



**Universidade Federal da Paraíba
Centro de Ciências Exatas e da Natureza
Departamento de Química
Programa de Pós-Graduação em Química**

Tese de Doutorado

**Algoritmo das Projeções Sucessivas para Seleção de Variáveis
Espectrais em Problemas de Classificação**

Márcio José Coelho de Pontes

Orientador: Prof. Dr. Mário César Ugulino de Araújo
2º Orientador: Prof. Dr. Roberto Kawakami Harrop Galvão

João Pessoa – Fevereiro de 2009



**Universidade Federal da Paraíba
Centro de Ciências Exatas e da Natureza
Departamento de Química
Programa de Pós-Graduação em Química**

Tese de Doutorado

**Algoritmo das Projeções Sucessivas para Seleção de Variáveis
Espectrais em Problemas de Classificação**

Márcio José Coelho de Pontes

Tese de Doutorado submetida ao Programa de Pós-Graduação em Química, da Universidade Federal da Paraíba, como parte dos requisitos para obtenção do título de Doutor em Química, área de concentração em “Química Analítica”.

Orientador: Prof. Dr. Mário César Ugulino de Araújo
2º Orientador: Prof. Dr. Roberto Kawakami Harrop Galvão

Bolsista (CAPES)

João Pessoa – Fevereiro de 2009

P814a Pontes, Márcio José Coelho de.

Algoritmo das projeções sucessivas para a seleção de variáveis espectrais em problemas de classificação / Márcio José Coelho de Pontes.- João Pessoa, 2009.

123p. : il.

Orientadores: Mário César Ugulino de Araújo, Roberto Kawakami Harrop Galvão

Tese (Doutorado) – UFPB/CCEN

1. Química Analítica. 2. Algoritmo das Projeções Sucessivas (SPA). 3. Análise Discriminante Linear (LDA). 4. Espectrometria UV-VIS. 5. NIR. 6.LIBS.

UFPB/BC

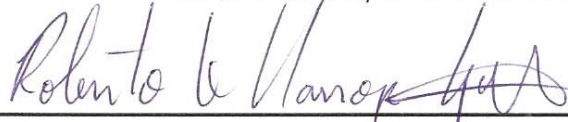
CDU: 543(043)

Algoritmo das Projeções Sucessivas para Seleção de Variáveis Espectrais em Problemas de Classificação

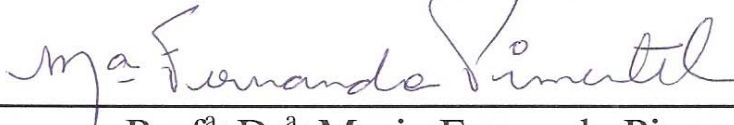
Aprovada pela banca examinadora:



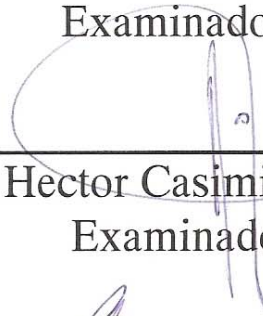
Prof. Dr. Mário César Ugulino de Araújo
Orientador/Presidente



Prof. Dr. Roberto Kawakami Harrop Galvão
2º. Orientador



Prof.ª Dr.ª Maria Fernanda Pimentel
Examinadora



Prof. Dr. Hector Casimiro Goicoechea
Examinador



Prof. Dr. Pedro Germano Antonino Nunes
Examinador



Prof. Dr. Wallace Duarte Fragoso
Examinador

**Em especial, a minha querida mãe, Miriam
Coelho, pela infinita paciência e por ser meu
exemplo de ser humano.**

**Ao meu pai José Pontes, pela amizade e
compreensão.**

Com carinho, dedico.

Agradecimentos

- A Deus;
- A toda minha família, pela educação, apoio e incentivo durante toda minha vida;
- À Universidade Federal da Paraíba, pelo apoio institucional;
- À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CAPES, pela bolsa concedida;
- Ao Prof. Dr. Mário César Ugulino de Araújo, pela orientação e confiança durante toda iniciação científica, mestrado e doutorado;
- Expresso minha profunda gratidão ao Prof. Dr. Roberto Kawakami Harrop Galvão, pela orientação, dedicação ao desenvolvimento desta tese e pelo atencioso acolhimento em todas as minhas visitas ao Instituto Tecnológico de Aeronáutica;
- Aos amigos Pablo Nogueira e Osmundo Neto pela ajuda na aquisição e preparação das amostras de óleos vegetais;
- Ao amigo Gledson Emídio, pelo companheirismo, boas discussões desse trabalho e pela obtenção dos espectros NIR de óleo diesel;
- Ao Cláudio Vicente da UFPE, pela realização da análise de referência das amostras de óleo diesel;
- A Urijatan Teixeira, Fátima Sanches e Francisco Antônio pelo registro dos espectros UV-VIS das amostras de café;
- Ao Programa Nacional de Cooperação Acadêmica (PROCAD) da CAPES, PROCAD 0081/05-1, pelo auxílio financeiro durante a missão de estudo;

- À Juliana Cortez, pela amizade e grande ajuda nas análises de solos;
- Ao Prof. Dr. Célio Pasquini, pelo acolhimento no Grupo de Instrumentação e Automação (GIA-UNICAMP) durante a missão de estudo;
- Ao Instituto Agrônomo de Campinas, IAC, pelo fornecimento das amostras de solos;
- Aos amigos Francisco Gambarra Neto e Sófacles Carreiro, pela amizade, companheirismo, boas conversas e valorosas discussões sobre Quimiometria.
- Aos Profs. Wallace Fragoso e Teresa Saldanha, pelas discussões, conselhos e ajuda durante o desenvolvimento da tese;
- A todos aqueles que fazem ou já fizeram parte da família LAQA, pela convivência agradável e a amizade cultivada nestes anos de trabalho;
- Aos amigos Sérgio, Alessandra, Edilene, Glauciene, Simone e Ricardo pelo companheirismo e apoio nas horas difíceis;
- À Liliana Lira, pela compreensão, paciência, carinho e ajuda no final da realização desse trabalho;
- Finalmente, a todos aqueles que de alguma forma contribuíram para a realização deste trabalho.

Sumário

Lista de Figuras.....	xi
Lista de Tabelas.....	xv
Lista de Abreviaturas e Siglas.....	xvii
Resumo.....	xviii
Abstract.....	xix
Publicações decorrentes do trabalho.....	xx

CAPÍTULO I. INTRODUÇÃO

1. INTRODUÇÃO	1
1.1. Aspectos gerais	1
1.2. Técnicas de reconhecimento de padrões	1
1.2.1. <i>Técnicas de reconhecimento de padrões não - supervisionadas</i>	4
1.2.2. <i>Técnicas de reconhecimento de padrões supervisionadas</i>	8
1.2.2.1. <i>SIMCA</i>	8
1.2.2.2. <i>LDA</i>	10
1.3. Seleção de variáveis	11
1.3.1. <i>Algoritmo das projeções sucessivas</i>	14
1.4. Objetivos	15

CAPÍTULO II. FUNDAMENTAÇÃO TEÓRICA

2. FUNDAMENTAÇÃO TEÓRICA	17
2.1. Pré-tratamento dos dados	17
2.2. PCA	18
2.3. SIMCA	20
2.4. LDA	21
2.5. Seleção de variáveis	22
2.5.1. <i>Algoritmo Genético</i>	24
2.5.2. <i>Stepwise</i>	25
2.5.3. <i>Algoritmo das Projeções Sucessivas para calibração multivariada</i>	27
2.5.4. <i>Algoritmo das Projeções Sucessivas para Classificação</i>	28

CAPÍTULO III. CLASSIFICAÇÃO DE ÓLEOS VEGETAIS

3. CLASSIFICAÇÃO DE ÓLEOS VEGETAIS	34
3.1. Introdução	34
3.1.1. <i>Óleos vegetais refinados</i>	34
3.2. Objetivos	37
3.3. Experimental	38
3.3.1. <i>Amostras</i>	38
3.3.2. <i>Equipamentos</i>	38
3.3.3. <i>Procedimento analítico</i>	39
3.3.4. <i>Tratamento dos dados e softwares</i>	39
3.5. Resultados e Discussões	41
3.4.1. <i>Espectros dos óleos vegetais</i>	41
3.4.2. <i>Análise exploratória dos dados</i>	42
3.4.3. <i>Classificação SIMCA</i>	43
3.4.4. <i>SPA-LDA</i>	44

3.4.5. GA-LDA.....	47
3.4.6. <i>Análise de sensibilidade ao ruído</i>	48
3.5. Considerações Finais	49
 CAPÍTULO IV. CLASSIFICAÇÃO DE ÓLEOS DIESEL	
4. CLASSIFICAÇÃO DE ÓLEOS DIESEL	51
4.1. Introdução	51
4.1.1. <i>Óleo diesel</i>	51
4.2. Espectrometria NIR	52
4.3. Objetivos	54
4.4. Experimental	54
4.4.1. <i>Amostras</i>	54
4.4.2. <i>Equipamentos</i>	54
4.4.3. <i>Procedimento analítico</i>	55
4.4.4. <i>Softwares</i>	55
4.5. Resultados e Discussões	56
4.5.1. <i>Espectros dos óleos diesel</i>	56
4.5.2. <i>Análise exploratória dos dados</i>	57
4.5.3. <i>SIMCA</i>	59
4.5.4. <i>SPA-LDA</i>	59
4.5.5. <i>GA-LDA</i>	61
4.5.6. <i>Análise de sensibilidade ao ruído</i>	62
4.6. Considerações Finais	63
 CAPÍTULO V. CLASSIFICAÇÃO DE CAFÉS	
5. CLASSIFICAÇÃO DE CAFÉS	65
5.1. Introdução	65
5.1.1. <i>Cafés</i>	65
5.2. Objetivos	66
5.3. Experimental	67
5.3.1. <i>Amostras</i>	67
5.3.2. <i>Equipamentos</i>	68
5.3.3. <i>Procedimento Analítico</i>	68
5.3.4. <i>Tratamento dos dados e softwares</i>	68
5.4. Resultados e Discussão	69
5.4.1. <i>Espectros das amostras de café</i>	69
5.4.2. <i>Análise exploratória dos dados</i>	70
5.4.3. <i>Classificação SIMCA</i>	71
5.4.4. <i>SPA-LDA</i>	72
5.4.5. <i>PCA e SIMCA com as variáveis selecionadas pelo SPA-LDA</i>	73
5.4.6. <i>Robustez dos modelos</i>	75
5.5. Considerações finais	76
 CAPÍTULO VI. CLASSIFICAÇÃO DE SOLOS BRASILEIROS	
6. Classificação de solos brasileiros	79
6.1. Introdução	79
6.1.1. <i>Solos brasileiros</i>	79
6.1.2. <i>Classificação de solos</i>	80
6.2. Espectroscopia de Emissão em Plasma Induzido por Laser	81
6.3. Compressão de dados (Transformada Wavelet)	83

6.4. Objetivos	84
6.5. Experimental	84
6.5.1. Amostras de solos Brasileiros	84
6.5.2. Instrumento LIBS.....	85
6.5.3. Aquisição dos espectros	86
6.5.4. Tratamento dos dados e softwares	86
6.6. Resultados e Discussões	87
6.6.1. Classificação no domínio espectral original	89
6.6.1.1. Classificação SIMCA	90
6.6.1.2. Modelos GA-LDA, SW-LDA e SPA-LDA.....	91
6.6.2. Classificação no domínio dos coeficientes wavelet.....	94
6.6.2.1. Classificação SIMCA no domínio wavelet	90
6.6.2.2. Modelos GA-LDA, SW-LDA e SPA-LDA no domínio wavelet.....	91
6.7. Considerações Finais	96
 CAPÍTULO VII. CONCLUSÃO	
7.0 CONCLUSÕES	98
7.1. Propostas futuras	99
Referências Bibliográficas	100
Anexos	121

Lista de Figuras

Figura 1.1. Disposição da matriz de dados espectrométricos.	4
Figura 2.1. Processo de seleção de variáveis com validação	23
Figura 2.2. Fluxograma do GA.	25
Figura 2.3. Ilustração da seqüência de projeções realizadas pelo SPA. (a): Primeira iteração. (b): Segunda iteração. Nesse exemplo, a cadeia de variáveis que inicia em x_3 deverá ser $\{x_3, x_1, x_5\}$	31
Figura 3.1. Sistema montado para o registro dos espectros de óleos vegetais. (A): amostra; (B): bomba peristáltica; (C): cubeta de fluxo; (D): descarte e (E): espectrofotômetro UV-VIS.	38
Figura 3.2. Representação gráfica do mecanismo de busca do algoritmo KS com três amostras selecionadas.	39
Figura 3.3. Espectros UV-VIS das amostras de óleos vegetais comestíveis analisados.	41
Figura 3.4. Gráfico dos escores obtidos pela PC2 <i>versus</i> PC1 para todas as 119 amostras de óleos vegetais. (●: milho, ●: girassol, ▲: canola e ■: soja).	42
Figura 3.5. Gráfico dos escores obtidos pela PC3 <i>versus</i> PC1 para todas as 119 amostras de óleos vegetais. (●: milho, ●: girassol, ▲: canola e ■: soja).	43
Figura 3.6. Custo da validação em função do número de variáveis selecionadas pelo SPA-LDA para o conjunto de dados de óleos vegetais. A seta indica o ponto de mínimo da curva do custo (0.1817), o qual ocorre em sete comprimentos de onda.	44
Figura 3.7. Espectro médio para cada tipo de óleo vegetal analisado. (a) milho, (b) soja, (c) canola e (d) girassol.	45
Figura 3.8. Gráfico dos escores da Função Discriminante 2 (FD2) <i>versus</i> Função Discriminante 1 (FD1) para todas as amostras de óleos vegetais (●: milho, ●: girassol, ▲: canola e ■: soja).	46
Figura 3.9. Gráfico dos escores da Função Discriminante 3 (FD3) <i>versus</i> Função Discriminante 1 (FD1) para todas as amostras de óleos vegetais (●: milho, ●: girassol, ▲: canola e ■: soja).	46

Figura 3.10. Espectro médio para cada tipo de óleo vegetal analisado. (a) milho, (b) soja, (c) canola e (d) girassol. Os dezesseis comprimentos de ondas selecionados pelo GA-LDA encontram-se indicados com círculos.	47
Figura 4.1. (a) Espectrofotômetro FT-IR utilizado para o registros dos espectros de óleos diesel. (b) cubeta de fluxo de quartzo de 1 cm de caminho óptico.	55
Figura 4.2. Espectro NIR originais das amostras de óleos diesel.....	56
Figura 4.3. Espectros NIR derivativos das amostras de óleos diesel analisadas.	57
Figura 4.4. Gráfico dos escores obtidos pela PC2 <i>versus</i> PC1 para todas as 128 amostras de óleos vegetais. (■: baixo teor de enxofre e ■: alto teor de enxofre).....	58
Figura 4.5. Gráfico dos escores obtidos pela PC3 <i>versus</i> PC1 para todas as 128 amostras de óleos diesel. (■: baixo teor de enxofre e ■: alto teor de enxofre).....	58
Figura 4.6. Custo da validação em função do número de variáveis selecionadas pelo SPA-LDA para o conjunto de dados de óleos diesel. A seta indica o ponto mínimo da curva do custo (0.5478), no qual ocorre em dois comprimentos de onda.	60
Figura 4.7. Espectro médio original (a) e derivativo (b) de óleo diesel com indicação dos comprimentos de onda selecionados pelo SPA.	60
Figura 4.8. Espectro médio original (a) e derivativo (b) de óleo diesel com indicação dos comprimentos de onda selecionados pelo GA.	61
Figura 5.1. Estruturas das principais moléculas presentes no café. (a) cafeína; (b) trigonelina e (c) ácido clorogênico.	65
Figura 5.2. Espectros UV-VIS das quatro classes de cafés analisadas.	69
Figura 5.3. Espectros médios das quatro classes de cafés (linhas sólidas) com limites de +/- um desvio padrão (linhas tracejadas).	70
Figura 5.4. Gráfico dos escores de PC2 \times PC1 para todas as 175 amostras de café. Descafeinado não vencido: ■; Descafeinado vencido: ■; Cafeinado não vencido: ●; Cafeinado vencido: ●.....	71
Figura 5.5. Gráfico de scree obtido pelo SPA-LDA para os espectros UV-VIS de cafés.	72

Figura 5.6. Espectros médios das quatro classes de cafés estudadas com os 15 comprimentos de onda selecionados pelo SPA-LDA.....	73
Figura 5.7. Gráfico dos escores de PC2 × PC1 resultante da PCA realizada em todas as 175 amostras com as 15 variáveis selecionadas pelo SPA-LDA. Descafeinado não vencido: ■; Descafeinado vencido: ■; Cafeinado não vencido: ●; Cafeinado vencido: ●.....	73
Figura 5.8. Escores de FD2 × FD1 obtidos pela LDA com as variáveis selecionadas pelo SPA. Descafeinado não vencido: ■; Descafeinado vencido: ■; Cafeinado não vencido: ●; Cafeinado vencido: ●.....	75
Figura 6.1. Implementação da transformada wavelet empregando um banco de filtros com dois níveis de decomposição. <i>H</i> e <i>G</i> representam os filtros digitais passa-baixas e passa-altas, respectivamente e $\downarrow 2$ denota a operação de sub-amostragem.....	83
Figura 6.2. Instrumento LIBS construído em laboratório. 1: laser; 2: espelho dicróico; 3: lente; 4: lente coletora de luz; 5: placa de posicionamento; 6: fibra ótica, 7: policromador echelle e 8: célula contendo a amostra. Os detalhes dos componentes 4 e 5 encontram-se ampliados no lado superior esquerdo da figura.....	85
Figura 6.3. (a): célula de latão contendo a amostra de solo. (b): suporte usado para deixar a superfície plana após algumas análises.....	85
Figura 6.4. Espectros LIBS originais de uma mesma amostra de Argissolo.....	87
Figura 6.5. Espectros LIBS pré-processados (SNV) da mesma amostra de Argissolo.....	88
Figura 6.6. Espectros LIBS das 149 amostras de solos brasileiros analisados.....	89
Figura 6.7. Gráficos dos escores obtidos por PC2 × PC1 para as 149 amostras de solos brasileiros. ●: Argissolo, ■: Latossolo e ▲: Nitossolo. O percentual de variância explicada para cada PC encontra-se indicado entre parênteses.....	89
Figura 6.8. Gráficos dos escores obtidos por PC3 × PC1 para as 149 amostras de solos brasileiros. ●: Argissolo, ■: Latossolo e ▲: Nitossolo. O percentual de variância explicada para cada PC encontra-se indicado entre parênteses.....	90
Figura 6.9. Gráfico de <i>scree</i> obtido pelo SPA-LDA para os espectros LIBS de solos brasileiros.....	91

- Figura 6.10.** Espectro médio da classe Nitossolo com as variáveis selecionadas pelo SPA-LDA. **(A)**, **(B)** e **(C)** indicam as regiões ampliadas na Figura 6.11.....92
- Figura 6.11.** Espectros LIBS médio ampliados de cada classe com as variáveis selecionadas pelo SPA-LDA. Argissolo: —; Latossolo: — e Nitossolo: —. 92

Lista de Tabelas

Tabela 1.1. Valores de acidez, índice de refração e viscosidade obtidos em 4 amostras de óleos vegetais analisadas	2
Tabela 1.2. Valores de acidez, índice de refração e viscosidade obtidos em 45 amostras de óleos vegetais analisadas	3
Tabela 3.1. Características físico-químicas e composição em ácidos graxos dos óleos refinados de canola, girassol, milho e soja.....	36
Tabela 3.2. Classes e quantidade de amostras de óleos vegetais analisadas.	38
Tabela 3.3. Número de amostras de treinamento, validação e teste selecionadas pelo KS para as quatro classes de óleos vegetais.....	40
Tabela 3.4. Resultados da classificação SIMCA para o conjunto de Teste de óleos vegetais em diferentes níveis de significância do Teste- <i>F</i> (1%, 5%, 10% e 25%).	43
Tabela 3.5. Resumo dos resultados (erros de classificação no conjunto de teste) para o SPA-LDA, GA-LDA e SIMCA (4 níveis de significância do teste- <i>F</i>) para o conjunto de dados de óleos vegetais.....	48
Tabela 3.6. Resumo dos resultados de classificação (Erros do Tipo I e Tipo II) obtidos pelos modelos SPA-LDA, GA-LDA e SIMCA no conjunto de teste de óleos vegetais contaminado pelo ruído.....	49
Tabela 4.1. Classes e quantidade de amostras de óleo diesel analisadas.....	54
Tabela 4.2. Número de amostras de treinamento, validação e teste selecionadas pelo KS para as duas classes de óleos diesel.	55
Tabela 4.3. Número de erros de classificação dos modelos SIMCA para o conjunto de teste de óleo diesel em diferentes níveis de significância do teste- <i>F</i> (1%, 5%, 10% e 25%)......	59
Tabela 4.4. Resumo dos resultados (erros de classificação para o conjunto de teste) para o SPA-LDA, GA-LDA e SIMCA (4 níveis de significância do teste- <i>F</i>) aplicados ao conjunto de dados de óleos diesel.....	62
Tabela 4.5. Número total de erros (Tipo I e Tipo II) obtidos pelos modelos SPA-LDA, GA-LDA e SIMCA no conjunto de teste de óleos diesel contaminado pelo ruído.....	62
Tabela 5.1. Número e tipo de amostras de cafés analisadas.	67

Tabela 5.2. Número de amostras de treinamento, validação e teste selecionadas pelo KS para as quatro classes de cafés.....	68
Tabela 5.3. Número de erros de classificação obtido pelos modelos SIMCA usando toda a faixa espectral para o conjunto de amostras de teste de cafés. O número de PCs é indicado entre parênteses.....	71
Tabela 5.4. Número de erros de classificação obtido pelos modelos SIMCA (em quatro níveis de significância do Teste- <i>F</i> : 1%, 5%, 10% e 25%) usando as 15 variáveis selecionadas pelo SPA para o conjunto de amostras de teste de café. O Número de PCs é indicado entre parêntese.....	74
Tabela 5.5. Resumo dos resultados (erros de classificação para o conjunto de teste) para o SPA-LDA e SIMCA (4 níveis de significância do teste- <i>F</i>) aplicados ao conjunto de dados café Os valores entre parênteses indicam o número de erros obtidos pelos modelos SIMCA construídos com as variáveis selecionadas pelo SPA-LDA.....	74
Tabela 6.1. Número de amostras de cada tipo de solo brasileiro analisado.....	84
Tabela 6.2. Número de amostras de treinamento, validação e teste para cada classe de solo.....	86
Tabela 6.3. Número de erros de classificação obtido pelos modelos SIMCA (nos níveis de significância do teste- <i>F</i> de 1%, 5%, 10% e 25%) para um conjunto de amostras de teste de solos.....	90
Tabela 6.4. Número de erros obtidos pelos modelos GA-LDA, SW-LDA e SPA-LDA em amostras de solos brasileiros do conjunto de teste.....	93
Tabela 6.5. Número de coeficientes wavelet necessários para explicar 95% da variância dos dados.....	94
Tabela 6.6. Número de erros de classificação SIMCA (no domínio wavelet) obtidos para o conjunto de teste.....	95
Tabela 6.7. Número de erros de classificação dos modelos GA-LDA, SW-LDA e SPA-LDA construídos no domínio wavelet (conjunto de teste). Os valores entre parênteses indicam o número de coeficientes wavelet selecionados pelos modelos.....	95
Tabela 6.8. Resumo final dos erros de classificação para os modelos SIMCA, GA-LDA, SW-LDA e SPA-LDA obtidos no domínio das variáveis originais e dos coeficientes wavelet.....	95

Lista de Abreviaturas e Siglas

CA	Análise canônica
Coif	Coiflet
DA	Análise discriminante
Db	Daubechies
F_{cal}	Valor calculado para o teste F
F_{crit}	Valor crítico adotado para o teste F
FD	Função discriminante
FT	Transformada de Fourier
G	Risco médio de uma classificação incorreta pela LDA
GA	Algoritmo genético
g_k	Risco de uma classificação incorreta do objeto \mathbf{x}_k da <i>k</i> -ésima amostra de validação
ICP-OES	Espectroscopia de emissão ótica com plasma indutivamente acoplado
LDA	Análise discriminante linear
LIBS	Espectrometria de emissão em plasma induzido por laser
MLR	Regressão linear múltipla
NIPALS	Mínimos Quadrados Parciais Iterativos não-lineares
NIR	Infravermelho próximo
NIRR	Refletância Difusa no Infravermelho Próximo
PCA	Análise de Componentes Principais
PCs	Componentes principais
PLS	Regressão por mínimos quadrados parciais
RMSEV	Raiz quadrada do erro médio quadrático de previsão para o conjunto de validação
SIMCA	Modelagem independente e flexível por analogia de classe
SPA	Algoritmo das projeções sucessivas
SW	Stepwise
Sym	Symlet
UV-VIS	Ultravioleta e visível
WC	Compressão wavelet
WT	Transformada wavelet

Resumo

Neste trabalho, o Algoritmo das Projeções Sucessivas (SPA: *Successive Projections Algorithm*), originalmente proposto para seleção de variáveis espectrais em modelos de Regressão Linear Múltipla (MLR: *Multiple Linear Regression*), é adaptado para o contexto de classificação. Para este propósito, uma nova função de custo associada ao risco médio de classificação incorreta pela Análise Discriminante Linear (LDA: *Linear Discriminant Analysis*) é concebida para guiar a seleção do SPA. O método proposto é ilustrado em quatro problemas de classificação. No primeiro exemplo, a espectrometria UV-VIS é adotada para classificar quatro tipos de óleos vegetais comestíveis (milho, soja, canola e girassol). No segundo caso, a espectrometria NIR é usada para discriminar amostras de diesel com respeito ao teor de enxofre (baixo ou alto). Nessas duas primeiras aplicações, o SPA é comparado com a Modelagem Independente e Flexível por Analogia de Classe (SIMCA: *Soft Independent Modeling of Class Analogy*) e com o Algoritmo Genético (GA: *Genetic Algorithm*) em termos do número de erros de classificação para o conjunto de amostras que não é usada no processo de modelagem (amostras de teste). No terceiro problema, a espectrometria UV-VIS é novamente usada para classificar extratos aquosos de cafés brasileiros torrados e moídos com respeito ao tipo (cafeinado/descafeinado) e ao estado de conservação (vencido e não vencido). Nos três primeiros estudos de caso, os modelos são também comparados em termos de sensibilidade ao ruído instrumental. Os espectros do conjunto de teste são contaminados com ruído extra e os modelos previamente obtidos (sem a adição do ruído) são aplicados para a classificação do novo conjunto de teste. Na última aplicação, o uso da Espectroscopia de Emissão em Plasma Induzido por Laser (LIBS: *Laser-Induced Breakdown Spectroscopy*) é investigado para classificação de solos em três diferentes ordens (Argissolo, Latossolo e Nitossolo). Para este caso, o SPA é comparado com um método de seleção de variáveis do tipo Stepwise (SW), bem como GA e SIMCA, em termos do número de erros para o conjunto de teste. Os resultados mostram que o SPA-LDA é superior ao SIMCA e comparável ao GA-LDA e SW-LDA com respeito à exatidão na classificação. Adicionalmente, o SPA-LDA é menos sensível ao ruído instrumental e mais parcimonioso do que as demais estratégias de classificação avaliadas.

Palavras-chave: SPA, Classificação, LDA, Espectrometria UV-VIS, NIR, LIBS, Óleos vegetais comestíveis, Diesel, Cafés e Solos Brasileiros.

Abstract

In this work, the Successive Projections Algorithm (SPA), originally proposed for spectral variable selection in Multiple Linear Regression (MLR) models, is adapted to the context of classification. For this purpose, a new cost function associated to the average risk of misclassification by Linear Discriminant Analysis (LDA) is used to guide SPA selection. The proposed approach is illustrated in four classification problems. In the first example, UV-VIS spectrometry is adopted to classify four types of edible vegetable oils (corn, soya, canola and sunflower). In the second case, NIR spectrometry is used to discriminate diesel samples with respect to sulphur content (low and high level). In the first two applications, SPA is compared with Soft Independent Modelling of Class Analogy (SIMCA) and a Genetic Algorithm (GA) in terms of classification errors in a set of samples not used in the model-building process (test samples). In the third problem, UV-VIS spectrometry is again used to classify aqueous extracts of Brazilian ground roast coffee with respect to type (caffeinated/decaffeinated) and conservation state (expired and non-expired). In the first three case studies, the models are also compared in terms of sensitivity to instrumental noise. The spectra of the test set are contaminated with extra noise and the models previously obtained (without the additional noise) are applied to the classification of the new test set. In the last application, the use of Laser-Induced Breakdown Spectroscopy (LIBS) is investigated for classification of Brazilian soils into three different orders (Argissolo, Latossolo and Nitossolo). In this case, SPA is compared with a Stepwise (SW) variable selection method, as well as GA and SIMCA, in terms of test set errors. The results show that SPA-LDA is superior to SIMCA and comparable to GA-LDA and SW-LDA with respect to classification accuracy. Moreover, SPA-LDA is less sensitive to instrumental noise and more parsimonious than the other classification strategies evaluated.

Keywords: SPA, Classification, LDA, UV-VIS spectrometry, NIR, LIBS, Edible vegetable oils, Diesel, Coffees and Brazilian soils.

Publicações decorrentes do trabalho

- [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*, **78:11, 2005**.
- [2] SOUTO, U. T. C. P.; PONTES, M. J. C.; SILVA, E. C.; GALVÃO, R. K. H.; ARAÚJO, M. C. U.; SANCHES, F. A. C.; CUNHA, F. A. S.; OLIVEIRA, M. S. R., UV-Vis spectrometric classification of coffees by SPA-LDA, *Food Chemistry*, **Artigo submetido e revisado, 2008**.
- [3] PONTES, M. J. C.; CORTEZ, J.; GALVÃO, R. K. H.; PASQUINI, C.; ARAÚJO, M. C. U.; COELHO, R. M.; CHIBA, M. K.; MADARI, B. E., Classification of brazilian soils by using LIBS and variable selection in the wavelet domain, *Analytica Chimica Acta*, **Artigo aceito, 2009**.

CAPÍTULO I
INTRODUÇÃO

1. INTRODUÇÃO

1.1. Aspectos gerais

No desenvolvimento da Química Analítica Clássica, técnicas baseadas em precipitação, extração ou destilação eram muito utilizadas para separação dos principais componentes de interesse. Nesse contexto, as análises qualitativas eram realizadas com o uso de reagentes que, em contato com o analito, produziam compostos que eram identificados pela sua cor, ponto de fusão, ponto de ebulição, solubilidade, etc^[1].

Devido à relativa simplicidade, bem como à confiabilidade dos resultados obtidos, muitos desses métodos clássicos são ainda utilizados. Contudo, com o surgimento de novos problemas analíticos, os pesquisadores passaram a investigar outros fenômenos completamente diferentes daqueles inicialmente observados, tais como: condutividade elétrica, absorção e/ou emissão de luz. Com base nessas propriedades físicas, foram desenvolvidos métodos instrumentais de análise (eletroanalíticos, espectrométricos, entre outros) que hoje abrangem um número elevado de aplicações^[1,2].

Os instrumentos analíticos modernos permitem a produção de uma grande quantidade de informação (variáveis) em um número elevado de amostras (objetos) que podem ser analisadas em curtos intervalos de tempo. Conseqüentemente, a aquisição e a manipulação de dados multivariados, muitas vezes complexos e de difícil interpretação, tornaram-se uma prática comum nos laboratórios de pesquisa, principalmente os de Química Analítica.

Para extrair o máximo de informação útil desses dados, recorre-se geralmente ao uso de procedimentos matemáticos e estatísticos. Nesse contexto, técnicas Quimiométricas como as de reconhecimento de padrões podem ser empregadas como uma alternativa vantajosa^[3-5].

1.2. Técnicas de reconhecimento de padrões

Os seres humanos sempre foram eficientes em tarefas que requerem a percepção de diferenças e semelhanças entre objetos, tais como distinguir um círculo de um quadrado ou triângulo. Em Química Analítica, as técnicas de reconhecimento de padrões (RP) envolvem um conceito semelhante. Essas técnicas têm por finalidade encontrar similaridades e diferenças entre grupos de amostras

Capítulo I. Introdução

que foram submetidas a algum tipo de análise, seja por técnicas instrumentais (espectrométricas, cromatográficas, entre outras), seja pela determinação de alguns parâmetros de interesse da amostra (por ex.: pH, densidade, concentração de algumas espécies, etc)^[4]. Após a realização dessas análises, o conjunto de dados é normalmente disponibilizado em uma tabela onde amostras (objetos) são apresentadas nas linhas e as propriedades medidas (variável) destes objetos, nas colunas.

A **Tabela 1.1** mostra um exemplo com 4 amostras de óleos vegetais refinados analisadas e três variáveis (acidez, índice de refração e viscosidade) medidas.

Tabela 1.1. Valores de acidez, índice de refração e viscosidade obtidos em 4 amostras de óleos vegetais analisadas^[6].

Amostra	Acidez (% m/m ácido oléico)	Índice de Refração	Viscosidade (mPa s)
1	0,028	1,4695	69,0
2	0,041	1,4725	65,2
3	0,029	1,4692	69,4
4	0,041	1,4725	65,2

Com base nos valores mostrados na **Tabela 1.1**, o ser humano é capaz de afirmar que as amostras 1 e 3 são semelhantes, enquanto que a amostra 2 é semelhante à 4. Fazer tal distinção é simples porque o conjunto de dados apresentado é bem reduzido (baixa dimensionalidade). Nos mais freqüentes problemas analíticos, um número maior de amostras e/ou variáveis é necessário para que se possa, de fato, construir padrões e modelos matemáticos seguros e confiáveis.

A **Tabela 1.2** mostra novamente uma matriz de óleos vegetais com as mesmas variáveis utilizadas na **Tabela 1.1**. Contudo, um número maior de amostras é apresentado.

Capítulo I. Introdução

Tabela 1.2. Valores de acidez, índice de refração e viscosidade obtidos em 45 amostras de óleos vegetais analisadas^[6].

Amostra	Acidez (% m/m ácido oléico)	Índice de Refração	Viscosidade (mPa s)
1	0,026	1,4695	69,2
2	0,029	1,4692	69,4
3	0,032	1,4695	69,2
4	0,028	1,4695	69,0
5	0,025	1,4690	68,4
6	0,038	1,4695	68,6
7	0,041	1,4695	69,2
8	0,039	1,4695	69,0
9	0,036	1,4690	70,0
10	0,028	1,4695	69,0
11	0,031	1,4690	70,0
12	0,075	1,4692	70,4
13	0,063	1,4690	71,4
14	0,076	1,4730	66,2
15	0,042	1,4727	66,0
16	0,037	1,4725	64,2
17	0,117	1,4730	64,2
18	0,105	1,4730	62,2
19	0,099	1,4730	64,2
20	0,070	1,4730	65,0
21	0,084	1,4730	65,4
22	0,041	1,4725	65,2
23	0,042	1,4730	65,0
24	0,062	1,4730	66,2
25	0,114	1,4725	66,2
26	0,025	1,4730	64,6
27	0,060	1,4730	64,4
28	0,078	1,4700	68,2
29	0,073	1,4700	67,0
30	0,042	1,4700	66,6
31	0,062	1,4700	68,2
32	0,085	1,4703	66,4
33	0,070	1,4703	66,8
34	0,064	1,4700	67,2
35	0,035	1,4700	65,4
36	0,050	1,4703	66,4
37	0,041	1,4725	65,2
38	0,065	1,4703	66,2
39	0,106	1,4700	66,2
40	0,080	1,4705	67,2
41	0,070	1,4703	67,0
42	0,127	1,4703	68,4
43	0,294	1,4720	63,0
44	0,208	1,4720	62,8
45	0,198	1,4718	63,4

Diferentemente da **Tabela 1.1**, torna-se difícil, a olho nu, identificar padrões com base nos valores disponíveis na **Tabela 1.2**. Problemas como estes se tornam ainda maiores quando estão envolvidas variáveis contínuas provenientes de diferentes métodos espectrométricos. Por exemplo, na espectroscopia de emissão em plasma induzido por laser (LIBS: *laser-induced breakdown spectroscopy*) é possível registrar valores de intensidade de emissão em mais de vinte mil

Capítulo I. Introdução

comprimentos de onda, com apenas um único espectro. Conseqüentemente, o uso de técnicas de RP tornou-se uma ferramenta indispensável na identificação, caracterização e avaliação de diferenças e similaridades entre grupos de amostras e/ou variáveis em diferentes conjuntos de dados, sobretudo os espectrométricos^[3,4].

A **Figura 1.1** é apresentada com o propósito de facilitar o entendimento da manipulação dos dados provenientes de medidas espectrométricas.

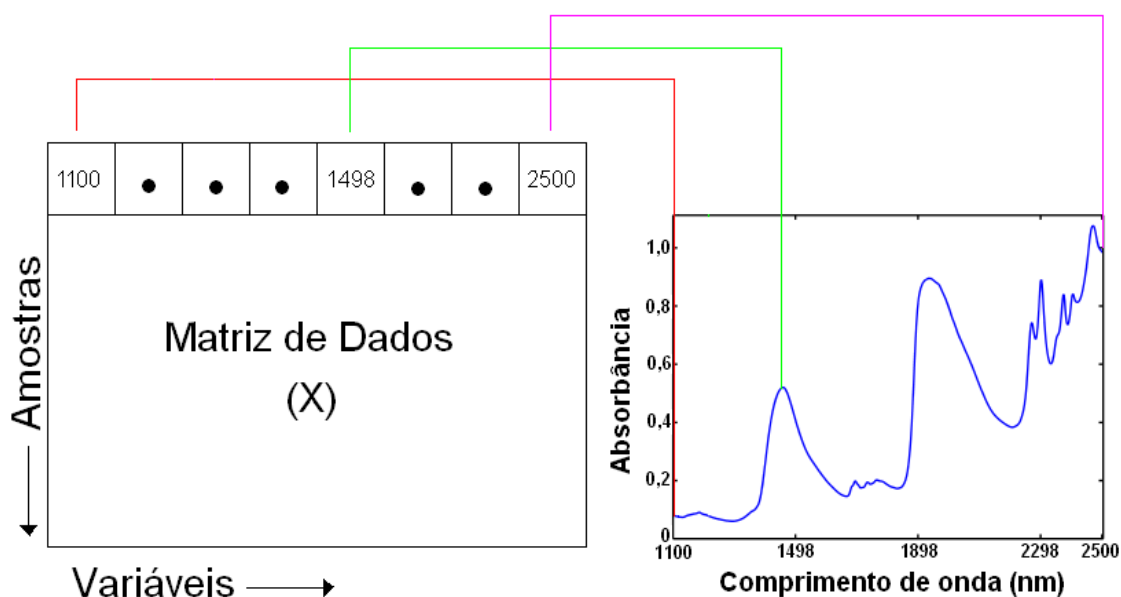


Figura 1.1. Disposição da matriz de dados espectrométricos.

A matriz de dados é organizada colocando-se, nas colunas, as variáveis que correspondem aos valores de absorvância de p comprimentos de onda. Neste exemplo, o espectro de absorção na região do infravermelho próximo (NIR: *Near Infrared*) foi registrado na faixa de 1100 a 2500 nm, com resolução de 2 nm. Então, para a aplicação das técnicas de reconhecimento de padrões, a matriz X terá 701 colunas correspondentes aos 701 comprimentos de onda (λ) e o número de linhas equivalente ao número de amostras (espectros) analisadas.

As técnicas de RP podem ser divididas em não-supervisionadas e supervisionadas^[3-5]. As características de ambas, bem como suas aplicações frente aos diferentes métodos serão mostradas nas próximas seções.

1.2.1. Técnicas de reconhecimento de padrões não-supervisionadas

As técnicas de RP não-supervisionadas avaliam a existência de agrupamentos sem utilizar o conhecimento prévio dos membros das classes, ou

Capítulo I. Introdução

seja, as amostras são examinadas utilizando apenas medidas de algumas propriedades com intuito de se observar agrupamentos naturais^[5].

As principais técnicas de RP não-supervisionadas são: análise de agrupamentos hierárquicos (*Hierarchical Cluster Analysis* – HCA) e análise de componentes principais (*Principal Component Analysis* – PCA)^[3,4,7]. Esses métodos são complementares, com muita aceitação por parte dos pesquisadores na análise de dados químicos.

A PCA é uma das técnicas de RP não-supervisionadas mais utilizadas. Nela, uma visão estatisticamente privilegiada e simples do conjunto de dados é fornecida através da criação de um novo conjunto de variáveis (novos eixos no espaço multidimensional), denominados Componentes Principais (PCs), que são ortogonais entre si e construídos da ordem da maior para a menor variância explicada dos dados. Teoricamente, o número de PCs é sempre igual ao número de variáveis. Entretanto, poucas componentes são responsáveis pela grande parte da variabilidade total dos dados. Em outras palavras, a PCA agrupa variáveis que estão altamente correlacionadas em novas variáveis, criando um conjunto que contém apenas as informações importantes e descartando as redundantes. Com isso, diminui-se o número de dimensões do sistema e cada amostra acaba sendo representada por um ponto em um espaço multidimensional menor, no qual é mais fácil a extração de informações e a observação de agrupamentos de amostras que apresentam características semelhantes^[4-5,7].

Aplicações envolvendo a PCA junto aos métodos espectrométricos têm sido freqüentemente apresentadas na literatura^[8-15]. Uma breve revisão de alguns trabalhos será mostrada abaixo.

As espectroscopias de emissão ótica e de massa com plasma indutivamente acoplado (ICP OES/ICP-MS) foram utilizadas para determinar traços de metais em chás de vários países africanos e asiáticos. A origem geográfica foi caracterizada pelas técnicas PCA e HCA (utilizando distância euclidiana e o método de *Ward*)^[8]. Além disso, um estudo supervisionado foi também realizado para procedimentos de classificação. Em ambas as técnicas não - supervisionadas, foi possível observar agrupamentos naturais referentes às amostras das diferentes origens.

A PCA foi utilizada com intuito de elucidar o potencial do uso da espectroscopia Raman (excitação em 785 nm) para medidas quantitativas de carotenóide, colágeno e gordura em músculo de peixe^[9]. Para o registro dos

Capítulo I. Introdução

espectros Raman, amostras do filé de peixes de espécies conhecidas foram utilizadas. Segundo os autores, foi possível associar os valores obtidos pelos escores com as espécies de peixes que apresentavam níveis elevados dos constituintes em estudo. A PC1 e PC2 descreveram, respectivamente, 65% e 13% da variância explicada dos dados. Os valores de pesos para a PC1 expressaram a variação do teor de gordura no peixe, visto que os picos associados com a mesma apresentaram valores positivos altos em 1301 cm^{-1} , 1265 cm^{-1} , 1076 cm^{-1} e 1064 cm^{-1} .

Uma avaliação de modificadores químicos na determinação direta e simultânea de Al, As, Cu, Fe, Mn e Ni em álcool etílico combustível por espectrometria de absorção atômica em forno de grafite (GFAAS) foi realizada através do uso da PCA^[10]. Os modificadores empregados nesse estudo foram: $\text{Pd}(\text{NO}_3)_2 + \text{Mg}(\text{NO}_3)_2$; W/Rh; W + co-injeção de $\text{Pd}(\text{NO}_3)_2 + \text{Mg}(\text{NO}_3)_2$ e, para cada um desses, foram utilizadas trinta amostras. Os resultados dos testes de adição e recuperação dos analitos frente aos diferentes modificadores foram utilizados como dados experimentais. A PCA possibilitou uma separação dos modificadores em função do intervalo de recuperação. Entre os modificadores estudados, aquele que apresentou os maiores teores de recuperação foi o W + co-injeção de $\text{Pd}(\text{NO}_3)_2 + \text{Mg}(\text{NO}_3)_2$, uma vez que o mesmo apresentou-se como a espécie de maior correlação positiva. Este então foi escolhido para a determinação direta e simultânea de Al, As, Cu, Fe, Mn e Ni.

A espectrometria RMN junto com a PCA foi utilizada para a diferenciação de vários tipos de amostras de chifres de cervo^[11]. Nesse trabalho, 3 PCs (93,5% da variância explicada acumulada) foram suficientes para uma boa discriminação entre as diferentes origens dessas amostras.

A PCA foi também utilizada para a discriminação de amostras de solos de diferentes origens geográficas. O estudo foi baseado nas atividades dos radionuclídeos (^{226}Ra , ^{238}U , ^{235}U , ^{40}K , ^{134}Cs , ^{137}Cs , ^{232}Th e ^7Be) detectadas pela espectrometria de raios gama^[12]. As três primeiras PCs totalizaram 81,7% da variância explicada dos dados (PC1, PC2 e PC3 explicaram 51,5%, 16% e 14,2%, respectivamente). Os radionuclídeos ^{226}Ra , ^{238}U e ^{232}Th apresentaram altos valores de pesos para a PC1 e foram os mais importantes para a discriminação das amostras de solos em diferentes localidades.

Capítulo I. Introdução

Técnicas quimiométricas de análise junto com espectroscopia de Fluorescência foram utilizadas com intuito de discriminar filés de peixes frescos dos descongelados^[13]. Inicialmente, a PCA foi aplicada aos espectros normalizados com intuito de se avaliar similaridades e diferenças nas duas classes de amostras. Duas PCs foram suficientes para uma boa discriminação, tendo a PC1 descrito 84,9% da variância explicada dos dados, enquanto que a PC2, 12,1%. A segunda etapa desse trabalho consistiu em utilizar os valores de escores obtidos pelas 5 primeiras PCs em um estudo supervisionado que utilizava a Análise Discriminante Fatorial (FDA). Nesse caso, também foram alcançados bons resultados.

Um conjunto de dados provenientes de espectros de absorvância na região do NIR foi utilizado por Pontes *et al.*^[14] em um estudo de classificação e verificação de adulteração de bebidas alcoólicas destiladas. O gráfico dos escores apresentado pelas duas primeiras PCs relevou uma boa discriminação entre as 4 classes de bebidas, assim como uma distinção entre as amostras adulteradas com etanol, metanol ou água em diferentes níveis de concentração.

Recentemente, a espectroscopia fotoacústica no infravermelho próximo com Transformada de Fourier (FTIR-PAS) foi avaliada quanto ao seu potencial para identificar amostras de solos^[15]. Os espectros PAS de 166 amostras pertencentes a cinco tipos de solos mediterrâneos foram registrados na região de 4000 a 400 cm^{-1} . Dois pré-processamentos foram realizados no domínio dos espectros: suavização pelo método de Savitzky-Golay^[16] (com janela de 25 pontos) e uma normalização pela área. A PCA foi, então, aplicada a esses dados com intuito de reduzir a dimensionalidade. Os valores de escores obtidos serviram como entrada para o emprego de outras técnicas de cunho supervisionado.

Em se tratando dos métodos de RP não-supervisionados, é importante ressaltar que a presença ou ausência de agrupamentos depende quase que exclusivamente dos valores das medidas que foram realizadas nas amostras. Em alguns sistemas de investigação, espera-se a presença de diversos agrupamentos, mas nem sempre isto acontece. Possivelmente, os motivos pelos quais isto não ocorra são: uso de alguma propriedade (por ex.: região do espectro, concentração de alguma espécie, etc) que não seja suficientemente discriminante, ou ainda, pela falta ou uso inadequado de algum pré-processamento.

Vale salientar que a PCA permite a realização de uma análise exploratória dos dados. Para classificar uma amostra futura como pertencente a um ou mais

Capítulo I. Introdução

agrupamentos inicialmente caracterizados, recorre-se às técnicas de RP supervisionadas que serão apresentadas na próxima seção.

1.2.2. Técnicas de reconhecimento de padrões supervisionadas

Diferentemente das técnicas de RP não-supervisionadas, as supervisionadas utilizam uma informação adicional sobre os membros das classes em estudo, ou seja, é necessário um conjunto de treinamento com objetos de categorias conhecidas para a elaboração de modelos que sejam capazes de identificar amostras desconhecidas^[3-5]. Antes da elaboração desses modelos, torna-se indispensável estabelecer quais medidas químicas são realmente adequadas para o processo de classificação, pois um mau planejamento experimental ou dados experimentais inadequados influenciam de forma negativa o desempenho dos métodos empregados.

Diversos Métodos de RP supervisionados têm sido aplicados em Química Analítica, mas eles diferem essencialmente na forma como a classificação é de fato alcançada. Dessa forma, a sua escolha dependerá da natureza dos dados, sobretudo com atenção ao número e tipo de variáveis empregadas no estudo. Contudo, a modelagem independente e flexível por analogia de classes (SIMCA: *Soft Independent Modeling of Class Analogy*) e a análise discriminante linear (LDA: *Linear Discriminant Analysis*) vêm se destacando nos últimos anos com um número elevado de aplicações^[3-4].

1.2.2.1. SIMCA

Proposto por Svante Wold^[17], o SIMCA é um método bem estabelecido na literatura e é largamente utilizado para classificação de amostras em conjuntos de dados com alta dimensionalidade. Nele, a localização dos objetos é modelada através do uso de componentes principais, ou seja, uma região no espaço multidimensional é delimitada através da construção de um modelo PCA para cada categoria de amostras. Um novo objeto será classificado como pertencente a uma ou mais classes previamente modeladas se possuir características que o permitam ser inserido no espaço multidimensional de algum (ns) dos modelos^[5,17].

O SIMCA tem sido aplicado com sucesso em diferentes matrizes, incluindo: medicamentos^[18] e ervas medicinais^[19], alimentos^[20], bebidas alcoólicas^[14], cosméticos^[21], entre outras^[22-25].

Capítulo I. Introdução

Candolfi *et al.*^[18] desenvolveram uma metodologia para identificar 10 excipientes utilizados em indústrias farmacêuticas. Nesse trabalho, o método SIMCA foi aplicado a espectros NIR na região de 1100 nm a 2468 nm (32 varreduras, com intervalos de 2 nm). Dois intervalos de confiança (95% e 99%) foram utilizados para a avaliação do desempenho dos modelos SIMCA em dados brutos, normalizados ou com a segunda derivada. Todas as amostras foram corretamente classificadas e o efeito dos diferentes pré-processamentos não influenciou nos resultados de classificação.

Espectros de reflectância difusa na região do visível-NIR (400 a 2500 nm) foram utilizados para classificar diferentes ervas medicinais (Giseng Radix, Austragali Radix e Smilacis Rhizoma). O SIMCA, entre os outros métodos de RP supervisionados avaliados, foi o que apresentou o melhor desempenho de classificação^[19].

A espectroscopia de Refletância Total Atenuada (ATR: *Attenuated Total Reflectance*) na região do infravermelho médio (MIR-IR) foi utilizada em combinação com a HCA e SIMCA para a autenticação de sucos de diferentes origens^[20]. Segundo os autores, 100% de acerto no conjunto de previsão foram alcançados pelos modelos SIMCA.

O potencial do método SIMCA foi explorado em um estudo de verificação de adulteração de bebidas alcoólicas destiladas^[14]. A espectroscopia NIR foi utilizada como técnica analítica desse trabalho e dois grupos de bebidas adulteradas foram avaliados: o primeiro consistia de amostras adulteradas no laboratório com 5% e 10% de água, etanol ou metanol. No segundo grupo, amostras que passaram por uma análise de referência em um órgão de fiscalização foram também estudadas. Segundo Pontes *et al.*^[14], um bom desempenho dos modelos SIMCA construídos com os dados NIR foi alcançado.

O SIMCA foi aplicado com sucesso em dados provenientes das Espectroscopias MIR – ATR ($4000-650\text{ cm}^{-1}$) e NIR (10000-4000 cm^{-1}) com fibra ótica para classificação de óleos de camélia autênticos e adulterados (5-25% m/m com óleo de soja)^[21]. Todas as amostras foram corretamente classificadas nesse estudo. Esse trabalho também explorou o método dos mínimos quadrados parciais (PLS: *Partial Least Squares*) para a predição da concentração do adulterante. Bons valores de correlação, RMSEC, RMSECV e RMPSEP foram obtidos.

Capítulo I. Introdução

Aplicações do SIMCA em outras técnicas como na espectroscopia de fluorescência^[22], Raman^[23], RMN^[24-25], entre outras^[3-4], são também citadas na literatura.

Nos últimos anos, um assunto de grande importância tem sido levantado em problemas de classificação. Há realmente a necessidade de utilizar, sobretudo com dados espectrométricos, todas as variáveis nas quais foram efetuadas as análises? Embora a literatura mostre muitas metodologias que utilizam toda a região do espectro^[3-4, 18-25], vários trabalhos têm mostrado que a capacidade preditiva dos modelos pode ser melhorada mediante uma seleção adequada de variáveis^[26-31], que idealmente deve eliminar as variáveis não informativas e reter aquelas que resultem em um número menor de erros de classificação, principalmente quando existir elevada sobreposição espectral. É nesse contexto que a LDA vem se destacando em aplicações envolvendo métodos espectrométricos em Química Analítica.

1.2.2.2. LDA

A LDA é um dos Métodos de RP supervisionados mais utilizados. Foi originalmente proposta por Fisher^[32] e é utilizada por muitos autores em diversas aplicações em Química Analítica, incluindo alimentos^[29, 31, 33], bebidas alcoólicas^[34], gasolina^[35], entre outras^[26-28, 30]. Este método é baseado na determinação de funções discriminantes lineares as quais maximizam a variância entre as classes e minimizam a variância dentro de cada classe. LDA pode ser considerada, assim como a PCA, um método de redução de dimensionalidade. Contudo, enquanto a PCA seleciona uma direção que retém a máxima variância dos dados em uma menor dimensão, a LDA seleciona uma direção que alcança a separação máxima entre as classes avaliadas.

Uma desvantagem da LDA em relação aos outros métodos de RP supervisionados é que a mesma é apropriada apenas para conjuntos de dados de pequenas dimensões. Além disso, a capacidade de generalização de modelos LDA pode ser comprometida por problemas de colinearidade^[36]. Dessa forma, o uso da LDA para classificação, sobretudo com dados espectrométricos, necessita de procedimentos de redução de dimensionalidade e/ou seleção de variáveis.

Nesse contexto, aplicações envolvendo técnicas de redução de dimensionalidade têm crescido consideravelmente nos últimos anos^[37-41]. Essas

Capítulo I. Introdução

técnicas têm como objetivo principal capturar a grande parte da informação útil em um número menor de variáveis. A PCA é uma das técnicas mais conhecidas e empregadas para esse propósito. Conforme descrito na seção 1.2.2.1, o bem estabelecido método SIMCA baseia-se na PCA, transformando as variáveis originais em variáveis latentes (PCs). Contudo, alguns autores vêm utilizando os valores de escores obtidos por um número restrito de PCs para serem aplicados junto aos modelos LDA^[15, 39, 42].

Estas transformações utilizadas pela PCA são eficientes em diversos problemas de classificação devido à sua capacidade em explicar a grande parte da informação dos dados em um número reduzido de variáveis. Apesar disso, a interpretação de tais modelos torna-se difícil porque nem sempre as novas variáveis possuem um significado físico e/ou físico-químico apropriado. Além disso, outro problema envolvendo este tipo de método é que, após a transformação, não é possível excluir regiões do sinal com baixa relação sinal/ruído, pois todas as variáveis originais contribuem para cada variável latente.

Em face do exposto, aplicações e/ou desenvolvimento de algoritmos de seleção de variáveis apresentam-se como uma importante área de pesquisa no âmbito de classificação^[43].

1.3. Seleção de variáveis

Em diversas aplicações envolvendo análise multivariada, tornam-se freqüentes problemas associados ao tamanho da matriz. De fato, conjuntos de dados com poucas amostras e/ou um elevado número de variáveis limitam a escolha e o desempenho do método de RP utilizado. Além disso, muitas dessas variáveis são irrelevantes ou redundantes, e as que são importantes são freqüentemente desconhecidas *a priori*. Algoritmos de seleção de variáveis surgem, portanto, como uma alternativa valiosa para minimizar ou contornar problemas desse tipo.

A pontuação de variáveis utilizadas no contexto dos métodos de classificação é um procedimento que pode ser utilizado quando se pretende trabalhar com seleção de variáveis. Algumas abordagens como a pontuação do classificador para variável individual (SVCR: *Single Variable Classifier Ranking*)^[44], Critério de Fisher^[32, 45-46] e critérios de correlação^[26, 47-50] têm sido utilizadas para o desenvolvimento de algoritmos de busca exaustiva ou seqüencial que escolhem melhores subconjuntos de variáveis.

Capítulo I. Introdução

O *Simulated Annealing* (SA), algoritmo de seleção de variáveis, foi originalmente proposto em 1953 por Metropolis *et al.*^[51] e popularizado após o trabalho de Kirkpatrick *et al.*^[52] em 1983. Trata-se de um método estocástico de busca global, cujo princípio está associado à Termodinâmica em simulações de “cozimento” de sólidos. No contexto da Química Analítica, a literatura é escassa de trabalhos envolvendo o SA em problemas de classificação. Destaca-se o trabalho apresentado por Llobet *et al.*^[53]. Nesse estudo, os autores avaliaram a espectrometria de massa com nariz eletrônico quanto ao seu potencial para classificação de presuntos espanhóis. O SA foi aplicado em redes neurais artificiais e 97,24% das amostras foram corretamente classificadas. O número de variáveis foi reduzido de 209 para 14.

O método de eliminação de variáveis não informativas (UVE: *Uninformative Variable Elimination*), proposto inicialmente por Centner *et al.*^[54] para calibração multivariada, foi adaptado para o contexto dos métodos de classificação por Wu *et al.*^[46]. Para esse propósito, modelos de análise discriminante pelos mínimos quadrados parciais (PLS-DA: *Partial Least Squares-Discriminant Analysis*) foram empregados e a taxa de classificação correta (TCC) foi adotada como parâmetro de desempenho do método.

O algoritmo genético (GA: *Genetic Algorithm*), muito utilizado em calibração multivariada, é um método de seleção de variáveis de natureza estocástica, assim como o SA. Contudo, o procedimento é realizado através de uma simulação de processos naturais da evolução com aplicação da teoria da evolução das espécies proposta por Darwin: “Quanto melhor um indivíduo se adaptar ao seu meio ambiente, maior será sua chance de sobreviver e gerar descendentes”^[55-57]. No contexto dos métodos de classificação, alguns trabalhos podem ser encontrados^[58-60].

O GA foi incorporado ao método de análise de variável canônica discreta (DCVA: *Discret Canonical Variate Analysis*) e comparado, em seis conjuntos de dados, com as técnicas LDA e redes neurais artificiais^[58]. Na maioria dos casos, o desempenho do GA-DCVA foi superior às demais estratégias.

Dharmaraj *et al.*^[59] utilizaram a espectroscopia FTIR para a classificação de diferentes origens geográficas da *Phyllanthus niruri* Linn., espécie de planta largamente difundida na Amazônia e eficiente para o tratamento de cálculo renal. O GA foi utilizado para seleção de variáveis em modelos LDA. O PCA-LDA e SIMCA

Capítulo I. Introdução

foram utilizados como estratégias de comparação. Apesar dos dois últimos apresentarem também bons desempenhos de classificação, o GA-LDA com apenas seis variáveis (número de onda) selecionadas foi o método que apresentou o melhor resultado (índice de acerto de predição de 100%).

Recentemente, Avci *et al.*^[60] combinaram o GA, a transformada wavelet discreta (DWT: Discrete Wavelet Transform) e redes neurais artificiais para a classificação de diferentes representações de texturas de imagens. Percentuais de 30% e 0,3% para o cruzamento e mutação, respectivamente, foram implementados nesse método. As amostras foram classificadas com um índice de acerto de predição de 95%.

O Stepwise (SW) é um outro algoritmo de seleção de variáveis que pode ser utilizado em problemas de classificação. Nele, a importância apresentada por cada variável independente dentro de um dado modelo é investigada. Para isso, as variáveis são excluídas ou adicionadas ao modelo de acordo com algum critério pré-estabelecido.

Sola-Larrañaga e Navarro-Blasco^[61] classificaram amostras de leite de vaca de diferentes variações sazonais e origens geográficas com base na determinação das concentrações de proteína, gordura, cinco minerais e traços de nove elementos (Fe, Zn, Cu, Mn, Se, Al, Cd, Cr e Pb) empregando o IR ou as Espectroscopias de Absorção Atômica (AAS: *Atomic Absorption Spectroscopy*) e de Emissão Atômica com Plasma Indutivamente Acoplado (ICP-AES: *Inductively Coupled Plasma Atomic Emission Spectroscopy*). Para esse propósito, o SW-LDA, com o critério de Wilks^[7], foi utilizado após um estudo preliminar realizado com a PCA. Bons resultados foram alcançados por essa metodologia.

Oliveros *et al.*^[62] testaram o SW-LDA com os critérios de Wilks^[7] e da distância de Mahalanobis para discriminação de diferentes origens geográficas de óleos de oliva.

Caneca *et al.*^[27] utilizaram a espectrometria no infravermelho (NIR e MIR) com o SW-LDA para a classificação de óleos lubrificantes em três diferentes estágios de desgaste. Para propósito de comparação, o método do *k*-ésimo vizinho mais próximo (KNN: *K-Nearest Neighbors*) foi também utilizado. Segundo os autores, foi possível alcançar um índice de acerto de 93% para ambas as regiões espectrais.

Capítulo I. Introdução

Como alternativa para contornar o problema de multicolinearidade entre as variáveis selecionadas, uma recente modificação do Stepwise foi proposta por Forina *et al.*^[63] com o uso de um procedimento de ortogonalização.

1.3.1. Algoritmo das Projeções Sucessivas

Em 2001, Araújo *et al.*^[64] apresentaram o Algoritmo das Projeções Sucessivas (SPA: *Successive Projections Algorithm*), técnica de seleção de variáveis que utiliza operações simples num espaço vetorial para minimizar problemas de colinearidade. O SPA mostrou ser um método eficiente para seleção de variáveis espectrais no contexto da calibração multivariada, especificamente quando aplicado à regressão linear múltipla (MLR: *Multiple Linear Regression*).

O objetivo do SPA consiste em buscar um subconjunto representativo pequeno de variáveis espectrais com ênfase na minimização da colinearidade. Com isso, torna-se possível utilizar modelos MLR que, embora simples e de fácil interpretação, podem ser severamente afetados por problemas de colinearidade^[65].

O trabalho inicial do SPA^[64] realizou uma análise espectrofotométrica simultânea de complexos de Co^{2+} , Cu^{2+} , Mn^{2+} , Ni^{2+} e Zn^{2+} com 4-(2-piridilazo) resorcinol, em misturas que continham os analitos nas faixas de concentração de 0,05 a 1,5 mg l⁻¹ na região do ultravioleta e de 0,02- 0,5 mg l⁻¹ na região do visível. Adicionalmente, foi realizado um estudo comparativo envolvendo outros métodos, tais como: GA, PLS e regressão por componentes principais (PCR: *Principal Component Regression*). O SPA-MLR alcançou os melhores resultados.

Após a apresentação inicial do SPA, novos artigos foram publicados com diferentes modificações e aplicações, incluindo ICP-AES^[66], UV-VIS^[67], FTIR^[68], espectrometria NIR^[68-70], entre outras^[71-72].

É importante ressaltar que, no que diz respeito aos métodos de classificação, a habilidade de generalização dos modelos LDA pode ser comprometida pela presença de colinearidade entre as variáveis^[7,36]. Por essa razão, a LDA se restringe normalmente a problemas de pequenas dimensões. Dessa forma, a minimização de colinearidade proporcionada pelo SPA deve ser útil também para modelos LDA.

1.4. Objetivos

Adaptar o SPA, originalmente proposto para seleção de variáveis espectrais em modelos MLR, para modelos baseados em LDA.

Demonstrar a eficiência do SPA-LDA em quatro estudos de caso com três diferentes técnicas espectroscópicas:

1. *Classificação de quatro tipos de óleos vegetais comestíveis (soja, milho, canola e girassol) utilizando a espectrometria de absorção molecular UV-Visível;*
2. *Classificação de amostras de óleos diesel (baixo e alto teor de enxofre) utilizando a espectrometria NIR;*
3. *Classificação de cafés com respeito ao tipo (decafeinado/cafeinado) e ao estado de conservação (vencido/não vencido);*
4. *Classificação de solos brasileiros em três diferentes ordens (Argissolo, Latossolo e Nitossolo).*

Realizar um estudo comparativo entre SPA-LDA, SIMCA e outros algoritmos de seleção de variáveis em função do número de erros para um conjunto externo de amostras.

CAPÍTULO II

FUNDAMENTAÇÃO TEÓRICA

2. FUNDAMENTAÇÃO TEÓRICA

2.1. Pré-tratamento dos dados

Em muitos casos, o sinal analítico proveniente de técnicas espectrométricas poderá apresentar intensidade com magnitudes diferentes, variação sistemática da linha de base e/ou ruído instrumental. Então, antes mesmo de se aplicar os métodos de RP, o emprego de técnicas de pré-processamento dos dados deve ser previamente avaliado. Possivelmente, o uso inapropriado ou a ausência dessa etapa inicial poderá prejudicar o desempenho dos métodos adotados.

Basicamente, três pré-tratamentos dos dados podem ser aplicados no domínio das variáveis: centralização dos dados na média, escalonamento e auto-escalonamento^[5, 7, 73]. A centralização dos dados na média consiste na subtração dos elementos de cada linha pela média da sua respectiva coluna. No escalonamento, cada elemento de uma linha é dividido pelo desvio padrão da sua respectiva variável. Com isso, todos os eixos da coordenada são conduzidos ao mesmo comprimento e, conseqüentemente, cada variável fica com a mesma influência na construção dos modelos. O auto-escalonamento consiste em centralizar os dados na média e, em seguida, efetuar o escalonamento. Com isso, as variáveis terão médias zero e desvios padrão igual a um. Tanto o escalonamento quanto o auto-escalonamento são utilizados quando se pretende atribuir os mesmos pesos às variáveis empregadas.

Segundo Martens e Naes^[65], o auto-escalonamento aplicado às variáveis é inapropriado para dados espectroscópicos, visto que esta transformação poderá, de certo modo, maximizar a presença de informações irrelevantes (ruído). Contudo, alguns autores vêm utilizando com sucesso a variação normal padrão (SNV: *Standard Normal Variate*)^[74-76] frente aos dados espectroscópicos. Em SNV, um auto-escalonamento no domínio das amostras é realizado, corrigindo os efeitos de espalhamento da radiação e tamanho das partículas. A **Equação 2.1** mostra a transformação utilizada pela SNV^[74].

$$\text{SNV}_A = \frac{\sum_{i=1}^p (x_i - \bar{x})}{\sqrt{\frac{\sum_{i=1}^p (x_i - \bar{x})^2}{p-1}}} \quad (2.1)$$

onde: SNV_A são as variações normais padrão de p comprimentos de onda para uma amostra \mathbf{A} ; x : valor do sinal analítico em i comprimento de onda da amostra \mathbf{A} e \bar{x} é a média dos valores de p comprimentos de onda da amostra \mathbf{A} , calculado conforme a **Equação 2.2**.

$$\bar{x} = \frac{\sum_{i=1}^p x_i}{p} \quad (2.2)$$

Outro pré-processamento muito utilizado no domínio das amostras é a 1ª derivada, cuja finalidade é corrigir problemas relacionados com a variação da linha de base, além de possibilitar uma melhor visualização de picos existentes nos sinais originais. Entretanto, aplicar tal pré-processamento em espectros com baixa relação sinal/ruído pode, em alguns casos, não ser uma boa alternativa, uma vez que os efeitos do ruído no conjunto de dados tendem a aumentar.

Várias técnicas podem ser utilizadas para a filtragem de ruído aleatório e aumento da relação sinal/ruído^[16, 77-78]. Entre elas, destaca-se pela sua simplicidade e eficiência, a suavização pelo método de Savitzky-Golay^[16] que ajusta um polinômio de baixa ordem aos pontos de uma janela pelos mínimos quadrados. A escolha do número de pontos utilizado na janela é de suma importância, pois um número elevado pode acarretar perda de informações e um número reduzido, a permanência de ruído. Uma vez estabelecidos os pré-tratamentos mais adequados para um determinado conjunto de dados, técnicas de RP poderão ser então aplicadas.

2.2. PCA

Os algoritmos dos mínimos quadrados parciais iterativos não-lineares (NIPALS: *Nonlinear Iterative Partial Least Squares*) e a decomposição por valores singulares (SVD: *Singular Value Decomposition*) têm sido freqüentemente utilizados para o cálculo da PCA^[79].

Capítulo II. Fundamentação Teórica

Em termos matemáticos, A PCA realiza uma decomposição de uma matriz de dados originais ou pré-processados, \mathbf{X} ($m \times n$), em dois conjuntos: escores (\mathbf{t}) e pesos (\mathbf{l}) que representam, respectivamente, as coordenadas das amostras e a contribuição de cada variável ao longo da PC. Os valores dos pesos, que podem variar entre -1 a 1, correspondem ao co-seno do ângulo entre a PC e os eixos das variáveis originais. Quanto maior for este valor em módulo, maior importância terá a variável na PC^[5].

O algoritmo NIPALS é o adotado pelo programa Unscrambler^[80] para realizar o cálculo da PCA. Para isso, considera-se uma matriz pré-processada \mathbf{X} de dimensões ($N \times J$), de modo que a j -ésima variável x_j esteja associada ao j -ésimo vetor coluna (\mathbf{x}_j , $j = 1, \dots, J$). O vetor \mathbf{x}_j que apresentar a maior norma é utilizado como uma estimativa inicial para \mathbf{t}_1 , que são os escores da PC1. Em seguida, projeta-se \mathbf{X} sobre \mathbf{t}_1 para estimar o vetor dos pesos (\mathbf{l}_1) para a PC1. Repete-se este procedimento até a convergência. De modo a facilitar a compreensão do algoritmo NIPALS^[81], uma seqüência utilizada para o cálculo das PCs é apresentada abaixo:

1. Escolhe-se um vetor \mathbf{x}_j de maior norma como estimativa inicial para \mathbf{t}_1 .

$$\mathbf{t}_1 = \mathbf{x}_j \quad (2.3)$$

2. Projeta-se \mathbf{X} sobre \mathbf{t}_1 para estimar os pesos (\mathbf{l}_1) para a PC1

$$\mathbf{l}_1 = \left[\frac{(\mathbf{t}_1^t \cdot \mathbf{X})}{(\mathbf{t}_1^t \cdot \mathbf{t}_1)} \right]^t \quad (2.4)$$

3. Normaliza-se o vetor \mathbf{l}_1 para comprimento 1.

$$\mathbf{l}_1 = \frac{\mathbf{l}_1}{\sqrt{(\mathbf{l}_1^t \cdot \mathbf{l}_1)}} \quad (2.5)$$

4. Projeta-se \mathbf{X} sobre \mathbf{l}_1 para obter uma nova estimativa de \mathbf{t}_1 (vetor de escores para PC1).

$$\mathbf{t}_1 = \mathbf{X} \cdot \mathbf{l}_1 \quad (2.6)$$

5. Estima-se o autovalor (a_1)

$$a_1 = \mathbf{t}_1^t \cdot \mathbf{t}_1 \quad (2.7)$$

6. Verifica-se a convergência

$$|a_0 - a_1| < \nu \quad (2.8)$$

Para verificar o cálculo da convergência nesta primeira etapa, adota-se a_0 , que é a variância explicada inicial do cálculo, como sendo igual a zero. Caso o módulo da diferença seja maior do que o valor adotado pela convergência (ν), que é normalmente na ordem de 10^{-4} ou menor, o cálculo retorna à etapa 2 e a_0 será igual a a_1 . Caso contrário, os valores de escores, pesos e variância explicada para PC1 serão \mathbf{t}_1 , \mathbf{l}_1 e a_1 , respectivamente. Nesse caso, o resíduo da PC1, denotado por \mathbf{X}_1 , é calculado por:

$$\mathbf{X}_1 = \mathbf{X} - \mathbf{t}_1 \cdot \mathbf{l}_1^t \quad (2.9)$$

Tal resíduo será utilizado para o cálculo da próxima PC.

O NIPALS tem a vantagem de não usar inversão de matrizes, o que torna o cálculo mais rápido para matrizes grandes^[80-81]. Já com o algoritmo SVD, os valores de escores e pesos são calculados simultaneamente e procedimentos de inversão de matriz são exigidos.

2.3. SIMCA

Para realizar o cálculo da distância da amostra ao modelo SIMCA, utilizam-se a variância residual para cada amostra da classe c , S_i (**Equação 2.10**), e a variância residual total, S_o (**Equação 2.11**).

$$S_i^c = \sqrt{\frac{\sum_{j=1}^p (res_j^c)^2}{p - A_c}} \quad (2.10)$$

$$S_o^c = \sqrt{\frac{\sum_{i=1}^{N_c} \sum_{j=1}^p (res_{ij}^c)^2}{(N_c - A_c - 1) \cdot (p - A_c)}} \quad (2.11)$$

onde N_c é o número de amostras pertencentes ao conjunto de treinamento da classe c ; A_c é o número de componentes principais utilizadas pela classe c ; p representa o

Capítulo II. Fundamentação Teórica

número de variáveis, i e j representam os índices das amostras e variáveis, respectivamente.

Um teste F é, então, utilizado para verificar a localização da amostra em relação ao(s) modelo(s). Compara-se o valor obtido pela **Equação 2.12** (F_{cal}) com um valor crítico (F_{crit}) que pode ser obtido empiricamente ou tabelado para um determinado nível de confiança e graus de liberdade. Se a amostra sob investigação apresentar um valor de F_{cal} menor do que o obtido pelo F_{crit} , a mesma pertencerá à classe em consideração.

$$F_{cal} = \frac{(S_i^c)^2}{(S_o^c)^2} \cdot \frac{N_c}{N_c - A_c - 1} \quad (2.12)$$

É importante ressaltar que dois tipos de erros podem ser apresentados em uma classificação SIMCA:

- **Tipo I:** a amostra não é classificada em sua classe verdadeira;
- **Tipo II:** a amostra é classificada em uma classe errada.

Portanto, uma mesma amostra poderá não ser classificada na sua classe verdadeira e ser ou não classificada em outra(s) classe(s).

2.4. LDA

Na LDA, a variável latente (Função Discriminante) é obtida através de uma combinação linear das variáveis originais. Quando um estudo de classificação apresentar c classes de amostras, $c - 1$ funções discriminantes poderão ser determinadas se o número de variáveis for maior do que c ^[7].

O processo de classificação da LDA está associado ao conceito da distância de Mahalanobis^[5,45], que pode ser definida da seguinte forma:

Seja $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_p]^T$ um objeto que deve pertencer a uma das c classes possíveis. Em caso de dados espectrométricos, as variáveis de classificação x_1, x_2, \dots, x_p podem corresponder, por exemplo, às medidas de absorvância realizadas em p comprimentos de ondas. O quadrado da distância de Mahalanobis $r^2(\mathbf{x}, \boldsymbol{\mu}_j)$ entre \mathbf{x} e o centro da j -ésima classe ($j = 1, 2, \dots, c$) é definido conforme a **Equação 2.13**.

$$r^2(\mathbf{x}, \boldsymbol{\mu}_j) = (\mathbf{x} - \boldsymbol{\mu}_j)^t \cdot \boldsymbol{\Sigma}_j^{-1} \cdot (\mathbf{x} - \boldsymbol{\mu}_j) \quad (2.13)$$

onde $\mu_j(p \times 1)$ e $\Sigma_j(p \times p)$ são, respectivamente, o vetor-média e a matriz de covariância para a classe $j^{[45]}$. Se os valores da média e covariância são desconhecidos (o que usualmente acontece), estimativas m_j e S_j podem ser empregadas no lugar de μ_j e Σ_j , respectivamente. Tais estimativas podem ser obtidas a partir de um conjunto de treinamento com objetos de classificação conhecida^[5]. É importante salientar que a LDA estima uma única matriz de covariância conjunta S , em vez de utilizar uma estimativa separada para cada classe. Este procedimento simplifica o modelo de classificação e resulta em superfícies de decisão lineares no \mathfrak{R}^p ^[5, 27, 45, 82]. Com esta modificação, o quadrado da distância de Mahalanobis entre o objeto x e o centro da j -ésima classe é calculado a partir da **Equação 2.14**.

$$r^2(x, m_j) = (x - m_j)^t \cdot S^{-1} \cdot (x - m_j) \quad (2.14)$$

O objeto x é, então, atribuído à classe j para a qual $r^2(x, m_j)$ tiver o menor valor.

Com intuito de se ter um problema bem condicionado, o número de amostras deverá ser maior do que o número p de variáveis a serem incluídas no modelo LDA. Caso contrário, a matriz de covariância estimada S será singular, o que inviabiliza o cálculo da matriz inversa na **Equação 2.14**. Portanto, o uso da LDA em dados espectrométricos depende, quase que totalmente, de procedimentos de seleção de variáveis.

2.5. Seleção de variáveis

Vários autores têm procurado definir seleção de variáveis baseado em diferentes critérios^[83]. Três definições são apresentadas abaixo:

1. *Clássica: Seleciona um subconjunto de M variáveis provenientes de um conjunto de N variáveis ($M < N$). Neste caso, uma função de custo é empregada para otimização.*
2. *Desempenho preditivo: seleciona subconjuntos de variáveis para melhorar ou não diminuir significativamente a habilidade preditiva dos modelos.*
3. *Aproximação da distribuição das classes originais: seleciona um subconjunto pequeno de variáveis de modo que a distribuição da classe resultante seja a mais próxima possível da distribuição da classe original que emprega todas as variáveis.*

A **Figura 2.1** mostra as quatro etapas comumente seguidas por um método de seleção de variáveis^[83]. Os detalhes desses passos serão mostrados posteriormente.

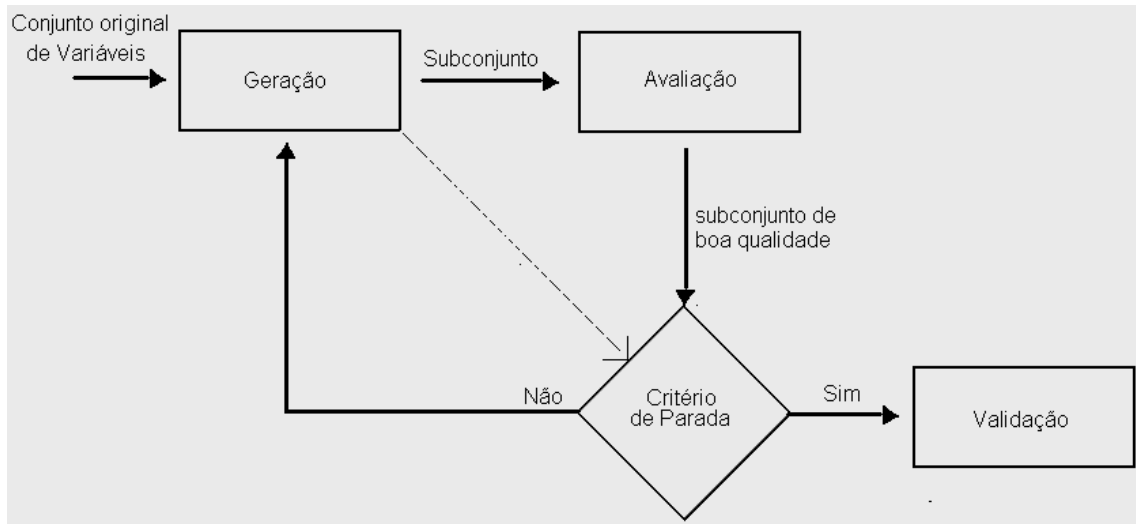


Figura 2.1. Processo de seleção de variáveis com validação^[83].

1. Geração

Esta etapa consiste essencialmente em gerar subconjuntos de variáveis para uma posterior avaliação. Pode-se começar: (i) com nenhuma variável, (ii) com todas ou (iii) com um subconjunto aleatório^[84]. Desta forma, as variáveis são iterativamente adicionadas ou removidas durante o processo.

2. Avaliação

Aqui, uma função de avaliação é adotada para medir a qualidade do subconjunto gerado pela etapa anterior. Obtém-se, então, um valor que é comparado com um outro previamente estabelecido. Se o subconjunto em investigação apresentar um valor que seja melhor, este será então substituído pelo anterior e assim por diante. Se não existir um critério apropriado para terminar o ciclo, a busca por uma cadeia de variáveis adequada se tornará um procedimento exaustivo e quase que impraticável^[84].

3. Critério de Parada

O critério de parada pode ser influenciado pelos procedimentos de geração e da função de avaliação^[84]. De fato, esta etapa quando baseada no processo de geração de subconjuntos de variáveis, inclui:

- (i) se um número de variáveis que serão selecionadas for pré-definido;
- (ii) se um número de iterações alcançadas for pré-estabelecido.

Já quando for baseada na função de avaliação, inclui:

- (i) se a adição ou remoção de qualquer variável não produzir um melhor subconjunto;
- (ii) se um subconjunto otimizado de acordo com alguma função de avaliação é obtido.

O ciclo continua até quando o critério de parada for satisfeito. Em seguida, é fornecido como saída um subconjunto de variáveis para o procedimento de validação.

4. Validação

O procedimento de validação é de extrema importância na aplicação de um método de seleção de variáveis. É adequado que se verifique a validade da cadeia de variável selecionada pelo processo utilizando diferentes testes. Recomenda-se, também, comparar os resultados da técnica com outros previamente estabelecidos ou ainda por outros métodos de seleção de variáveis que façam uso de conjuntos de dados simulados e/ou reais^[84].

2.5.1. Algoritmo Genético

O cromossomo biológico é composto de genes que são responsáveis por características específicas para cada indivíduo. Analogamente, torna-se possível construir um cromossomo artificial e simular um processo evolutivo natural.

Em se tratando de processos químicos, cada gene representa um parâmetro a ser otimizado. Dessa forma, o GA codifica subconjuntos de variáveis na forma de uma série de valores binários (0/1) denominados “cromossomos”. Cada posição (ou “gene”) no cromossomo está associada a uma das variáveis disponível para seleção. Uma população é, então, gerada a partir de um conjunto aleatório de indivíduos que podem ser vistos como possíveis soluções para o problema. Durante o processo de evolução, a população é avaliada e, para cada indivíduo, é dada uma nota ou índice que reflete a habilidade de adaptação para um determinado ambiente (aptidão). Dessa forma, os indivíduos mais aptos são preservados para o processo de seleção e os remanescentes (menos aptos), descartados^[55].

É importante destacar que estes indivíduos escolhidos para seleção podem ser modificados através de mutações ou cruzamentos. Com esta modificação, são

Capítulo II. Fundamentação Teórica

gerados descendentes para a próxima seleção com características genéticas de ambas as partes. Este procedimento é repetido até que um determinado critério de parada seja estabelecido (solução satisfatória alcançada, processo de busca estagnado ou número máximo de gerações atingido)^[26, 55].

O fluxograma do GA é apresentado na **Figura 2.2**.

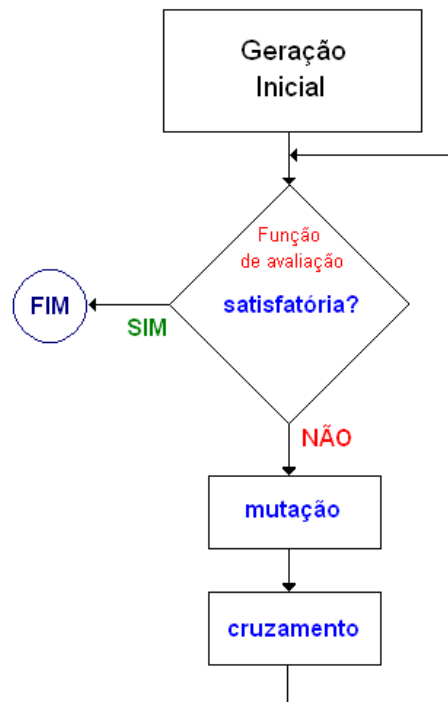


Figura 2.2. Fluxograma do GA.

2.5.2. Stepwise

O algoritmo inicia calculando o valor individual para cada variável espectral de acordo com seu poder de discriminação^[27]. O fator de discriminação (D_i)^[45] de uma variável x_i pode ser determinado a partir da **Equação 2.15**:

$$D_i = \frac{SB_i}{SW_i} \quad (2.15)$$

onde SW_i e SB_i são, respectivamente, as medidas de dispersão dentro da classe e entre as classes para a variável x_i . Calcula-se a dispersão SW_i ^[45] a partir da equação abaixo:

$$SW_i = \sum_{j=1}^C s_{ij} \quad (2.16)$$

onde s_{ij} é a dispersão de x_i na classe j , calculada conforme a **Equação 2.17**:

$$s_{ij} = \sum [x_i^k - m_{ij}]^2 \quad (2.17)$$

sendo x_i^k o valor de x_i na k -ésima amostra e m_{ij} é o valor médio de x_i na classe j , isto é:

$$m_{ij} = \frac{1}{n_j} \sum_{k \in I_j} x_i^k \quad (2.18)$$

Já a dispersão entre as classes, SB_i , é definida a partir da **Equação 2.19**:

$$SB_i = \sum_{j=1}^C n_j [m_{ij} - m_i]^2 \quad (2.19)$$

onde m_i é a média de x_i para todos os objetos do conjunto de treinamento.

Em cada passo, a variável x_i com o valor de D_i mais elevado é selecionada e o número de erros obtidos por validação cruzada *leave-one-out* é registrado. Antes do próximo passo, as variáveis que apresentarem uma alta correlação com a variável recém-selecionada são descartadas com intuito de evitar problemas de colinearidade. O algoritmo encerra o cálculo quando todas as variáveis forem avaliadas. O conjunto de variáveis que resultar em um menor número de erros de validação cruzada é então apresentado para o analista^[27]. Para uma melhor compreensão, os passos para esta estratégia de seleção de variável são mostrados abaixo:

Sejam v_{sel} e P os conjuntos contendo as variáveis já selecionadas e aquelas ainda disponíveis, respectivamente. Além disso, sejam γ um limiar de correlação ($0 < \gamma < 1$) adotado pelo usuário e N um contador que indica o número de variáveis já selecionadas. O algoritmo procede então da seguinte forma:

Passo 0 (inicializacao). $v_{sel} = \{ \}$, $P = \{1, \dots, p\}$, $N = 0$.

Passo 1. Calcular D_i para $1 \leq i \leq p$.

Passo 2. $i^* = \arg \max D_i$, $i \in P$.

Passo 3. Mover i^* de P para v_{sel} . Fazer $N = N + 1$.

Capítulo II. Fundamentação Teórica

Passo 4. Realizar um procedimento de validação cruzada leave-one-out usando as variáveis com índices em v_{sel} . Armazenar o número de erros resultantes em $E_{vc}(n)$.

Passo 5. Calcular o coeficiente de correlação múltipla (r_i) entre cada variável x_i com índice em P e as variáveis selecionadas com índices em V_{sel} . Tal coeficiente é obtido como:

$$r_i = \frac{\sigma(\hat{x}_i)}{\sigma(x_i)} \quad (2.20)$$

onde $\sigma(\cdot)$ denota o desvio padrão calculado no conjunto de treinamento e \hat{x}_i é uma estimativa de x_i obtida por regressão linear múltipla nas variáveis já selecionadas. Se o valor de r_i for próximo a um, então pode-se concluir que a inclusão da variável x_i não trará informação adicional ao modelo de classificação.

Passo 6. Excluir de P os índices das variáveis com coeficiente de correlação múltipla maior do que γ .

Passo 7. Se $P \neq \{\}$, retornar ao **Passo 2**.

Passo 8. Adotar o subconjunto de variáveis que apresentar o menor número de erros de validação cruzada, $E_{vc}(n)$, $n = 1, \dots, N$.

As variáveis selecionadas correspondem as primeiras n^* indexadas em V_{sel} . Caso diferentes subconjuntos de variáveis apresentem o mesmo número de erros de validação cruzada, o critério da parcimônia deverá ser obedecido escolhendo o subconjunto com menor número de variáveis.

Uma desvantagem desse algoritmo é a necessidade de se arbitrar um limiar para r_i . Tal escolha governa o descarte das variáveis, alterando o resultado final. Entretanto, é possível testar diferentes valores de limiar e, então, comparar os modelos LDA resultantes com base no número de erros obtidos em um conjunto separado de validação.

2.5.3. Algoritmo das Projeções Sucessivas para calibração multivariada

Em calibração multivariada, o SPA emprega conjuntos de calibração e validação, ambos com respostas instrumentais (\mathbf{X}) e valores medidos por um método de referência (y). A essência do SPA consiste em realizar operações de projeção na

Capítulo II. Fundamentação Teórica

matriz de calibração \mathbf{X}_{cal} ($Kc \times J$), cujas linhas e colunas correspondem a Kc amostras de calibração e J variáveis espectrais, respectivamente^[64].

A partir de cada uma das variáveis J disponível para o procedimento de seleção, o SPA constrói uma cadeia ordenada de Kc variáveis. Na construção dessa cadeia, cada elemento é selecionado de modo a exibir a mínima colinearidade com o anterior. Dessa forma, a colinearidade entre variáveis é avaliada pela correlação entre os vetores coluna da respectiva matriz de calibração \mathbf{X}_{cal} . Vale observar que, de acordo com este critério de seleção, não mais que Kc variáveis podem ser incluídas na cadeia^[64, 66].

Para cada uma das J cadeias de variáveis construídas como descrito acima, é possível extrair Kc subconjuntos de variáveis usando de um até Kc elementos na ordem em que eles foram selecionados. Assim, um total de $J \times Kc$ subconjuntos de variáveis podem ser formados. Para escolher o subconjunto mais apropriado, constroem-se modelos MLR, que são depois comparados em termos da raiz quadrada do erro médio quadrático para um conjunto de validação (RMSEV: *Root Mean Square Error of Validation*)^[64], calculado conforme a equação abaixo:

$$\text{RMSEV} = \sqrt{\frac{1}{Kv} \sum_{k=1}^{Kv} (y_v^k - \hat{y}_v^k)^2} \quad (2.21)$$

onde y_v^k e \hat{y}_v^k são, respectivamente, o valor de referência e o valor previsto para o parâmetro de interesse na k -ésima amostra de validação. Kv é número de amostras do conjunto de validação. Por fim, o algoritmo seleciona a cadeia de variáveis cujo modelo MLR levou ao menor RMSEV.

É importante ressaltar que, devido à simplicidade das operações algébricas envolvidas no SPA, o procedimento completo (incluindo a construção e validação dos modelos MLR) pode ser realizado em curtos intervalos de tempo para muitas aplicações.

2.5.4. Algoritmo das Projeções Sucessivas para Classificação^[43]

A informação da modelagem apresentada nos métodos de RP supervisionados está contida nos dados provenientes da resposta instrumental (matriz \mathbf{X}) e no índice de classes para cada amostra. Nesse caso, o RMSEV

Capítulo II. Fundamentação Teórica

(**Equação 2.21**) originalmente empregado no SPA não é uma métrica aplicável. Por esse motivo, uma nova função de custo foi concebida para guiar a seleção de variáveis.

A função de custo proposta refere-se ao risco médio G de uma classificação incorreta pela LDA. Assim como o RMSEV, esta função é calculada com base em um conjunto de validação, conforme descrito na **Equação 2.22**:

$$G = \frac{1}{K_v} \sum_{k=1}^{K_v} g_k \quad (2.22)$$

onde g_k (risco de uma classificação incorreta do objeto \mathbf{x}_k da k -ésima amostra de validação) é definido como:

$$g_k = \frac{r^2(\mathbf{x}_k, \mu_{lk})}{\min_{l_j \neq lk} r^2(\mathbf{x}_k, \mu_{lj})} \quad (2.23)$$

Na equação anterior, o numerador $r^2(\mathbf{x}_k, \mu_{lk})$ é o quadrado da distância de Mahalanobis entre o objeto \mathbf{x}_k (com índice de classe lk) e a média de sua classe (μ_{lk}). O denominador da **Equação 2.23** corresponde ao quadrado da distância de Mahalanobis entre o objeto \mathbf{x}_k e o centro da classe errada mais próxima. Idealmente, g_k deverá ser tão pequeno quanto possível, ou seja, o objeto \mathbf{x}_k deverá estar perto do centro da sua verdadeira classe e distante dos centros das demais classes.

Para iniciar o procedimento de seleção de variáveis no SPA-LDA, deve-se fornecer como entrada:

(i) Matrizes correspondentes às respostas instrumentais:

- *Conjunto de treinamento: Train* ($K_c \times J$);
- *Conjunto de validação: Val* ($K_v \times J$);
- *Conjunto externo para Teste: Test* ($K_t \times J$);

onde K_c , K_v e K_t representam o número de amostras para os conjuntos de treinamento, validação e teste, respectivamente. Esses conjuntos deverão ter o mesmo número de variáveis J .

(ii) Índices das classes:

- *Conjunto de treinamento: Group_Train* ($K_c \times 1$);
- *Conjunto de validação: Group_Val* ($K_v \times 1$);

Capítulo II. Fundamentação Teórica

- *Conjunto externo para Teste: Group_Test (Kt × 1);*

(iii) Número mínimo e máximo de variáveis a serem selecionadas

- *Número mínimo de variáveis: N1;*
- *Número máximo de variáveis: N2;*

É importante ressaltar que a construção de subconjuntos de variáveis com base no critério de minimização de colinearidade realizada pelo SPA-LDA resulta de uma seqüência de operações de projeções de vetores aplicadas às colunas da Matriz de treinamento (K_C, J). Contudo, antes mesmo de realizar tal procedimento, os objetos pertencentes a este conjunto são centralizados na média da sua própria classe. Então, torna-se necessário o uso dos índices de classes.

Considera-se que as respostas instrumentais (x) referentes ao conjunto de treinamento estejam em uma matriz \mathbf{X} de dimensões ($K_C \times J$), de forma que a j -ésima variável x_j esteja associada ao j -ésimo vetor coluna $\mathbf{x}_j \in \mathfrak{R}^{K_C}$. Sejam $\mathbf{M} = \min(K_C - C, J)$ o número máximo de variáveis que podem ser incluídas no modelo LDA e C é o número de classes envolvidas no problema.

Partindo de cada variável $\mathbf{x}_j, j = 1, \dots, J$, uma cadeia contendo \mathbf{M} variáveis é construída de acordo com as seguintes operações^[85]:

- **Passo 1:** *Início*

$\mathbf{z}^1 = \mathbf{x}_j$ (vetor que define as operações de projeção inicial)

$\mathbf{x}_k^1 = x_k, k = 1, \dots, J$

$L(1, j) = j$

- **Passo 2:** *Cálculo da matriz P^i de projeção no subespaço ortogonal a \mathbf{z}^i :*

$$\mathbf{P}^i = \mathbf{I} - \frac{\mathbf{z}^i (\mathbf{z}^i)^T}{(\mathbf{z}^i)^T \mathbf{z}^i} \quad (2.24)$$

onde \mathbf{I} é uma matriz identidade de dimensões apropriadas.

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

$$\mathbf{x}_k^{i+1} = \mathbf{P}^i \mathbf{x}_k^i \quad (2.25)$$

para $k = 1, \dots, J$.

- **Passo 4:** *Determinar o índice k^* do vetor de maior projeção e armazená-lo na matriz L .*

$$\mathbf{k}^* = \arg \max_{k=1, \dots, J} \|\mathbf{x}_k^{i+1}\| \quad (2.26)$$

Capítulo II. Fundamentação Teórica

$$L(i + 1, j) = k^* \tag{2.27}$$

- **Passo 5:** Fazer $z^{i+1} = x_k^{k+1}$ (Vetor que define as operações de projeção para a próxima iteração).
- **Passo 6:** Fazer $i = i + 1$. Se $i < M$, retornar ao passo 2.

Para cada iteração do procedimento apresentado acima, o vetor selecionado será aquele minimamente colinear em relação aos vetores selecionados nas iterações anteriores.

De modo a facilitar o entendimento das projeções, um exemplo simples (**Figura 2.3**)^[85] com $Kc = 3$ e $J = 5$ é apresentado para ilustrar os primeiros cinco passos do SPA.

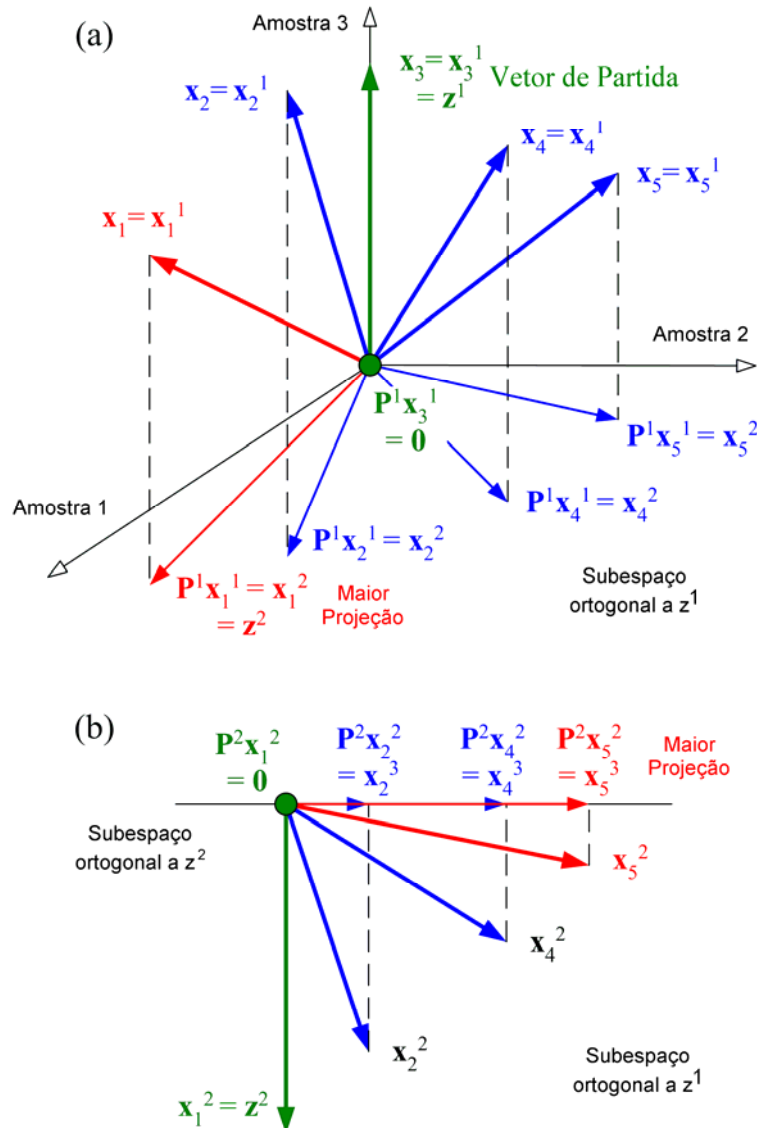


Figura 2.3. Ilustração da seqüência de projeções realizadas pelo SPA. (a): Primeira iteração. (b): Segunda iteração. Nesse exemplo, a cadeia de variáveis que inicia em x_3 deverá ser $\{x_3, x_1, x_5\}$.^[85]

Capítulo II. Fundamentação Teórica

O SPA-LDA foi adaptado utilizando a função $qr^{[71]}$ disponível no *software* Matlab. Com ela, torna-se possível gerar as cadeias de variáveis com maior eficiência computacional. Por *default*, a função qr adota a coluna com a maior norma como vetor de partida. Contudo, um procedimento foi realizado para forçar o algoritmo a iniciar pela *i-ésima* coluna da matriz de treinamento centrada na média ou auto-escalada.

Uma vez construídas as cadeias de variáveis candidatas, o SPA escolherá aquela que apresentar o menor risco médio de classificação incorreta pela LDA, como definido na **Equação 2.22**.

O algoritmo SPA-LDA tem como saída:

- **I**: Um vetor contendo a melhor cadeia de variáveis;
- **R**: Risco médio G de uma classificação incorreta pela LDA em função do número de variáveis usadas;
- **Lopt**: Matriz contendo as cadeias de variáveis associadas a R .

Além disso, o SPA-LDA fornece como resultado o número de erros (tanto para o conjunto de validação, como para teste), o valor ótimo de G e os índices das classes previstos pelo modelo LDA. Adicionalmente, o SPA-LDA apresenta um gráfico de *scree* que mostra a função de custo (Risco G , **Equação 2.22**) *versus* o número de variáveis. O número ótimo de variáveis corresponde ao mínimo dessa curva.

O SPA-LDA^[43] foi implementado com a sub-rotina *multilda.m* para o cálculo da LDA. Os códigos do programas encontram-se nos Anexos.

Com intuito de demonstrar a habilidade do SPA-LDA frente a diferentes matrizes e técnicas espectrométricas, quatro estudos de casos serão apresentados nos próximos capítulos:

- *Classificação de óleos vegetais utilizando a espectrometria de absorção molecular UV-VIS;*
- *Classificação de óleos diesel utilizando a espectrometria NIR;*
- *Classificação de cafés utilizando a espectrometria de absorção molecular UV-VIS;*
- *Classificação de solos brasileiros utilizando o LIBS.*

CAPÍTULO III
CLASSIFICAÇÃO DE ÓLEOS VEGETAIS

3. CLASSIFICAÇÃO DE ÓLEOS VEGETAIS

3.1. Introdução

3.1.1. Óleos vegetais refinados

Os óleos vegetais são gorduras cuja extração é realizada, em sua grande maioria, das sementes ou grão de espécies de plantas denominadas oleaginosas. São constituídos essencialmente por triglicerídios resultantes da esterificação de vários ácidos graxos (principais componentes) pelo glicerol. Os ácidos graxos poderão conter ou não, uma ou várias ligações duplas e o comprimento das respectivas cadeias pode variar de acordo com a origem botânica da planta de onde o óleo é extraído^[86].

Para que os óleos vegetais possam ser destinados ao consumo humano, um processo de refino deve ser empregado através da remoção de alguns componentes: ácidos graxos livres, proteínas, corantes naturais, umidade e compostos voláteis e inorgânicos. Este procedimento melhora a aparência, sabor, odor e garante uma maior estabilidade para o produto final^[86-87].

Por serem grandes fontes de nutrientes, os óleos vegetais previnem uma série de doenças, aceleram o metabolismo e auxiliam o bom funcionamento dos órgãos vitais^[86]. Em todo o mundo, existe uma grande variedade de oleaginosas utilizadas para a produção desses óleos. No Brasil, os principais óleos autorizados pela Agência Nacional de Vigilância Sanitária (ANVISA)^[88] para o comércio são:

- *algodão*
- *babaçu*;
- *oliva*;
- *uva*.
- *amendoim*;
- **canola**;
- *palma*;
- *arroz*;
- **girassol**;
- *palmiste*;
- *azeite saborizado*
- **milho**;
- **soja**;

Entre os óleos apresentados acima, canola, girassol, milho e soja foram utilizados nesse trabalho. O perfil de cada um será apresentado a seguir:

• Óleo de canola

Em comparação com a grande maioria das oleaginosas cultivadas para a produção de óleos vegetais refinados, o uso da canola para fins alimentícios só ocorreu por volta da década de 70, quando cientistas canadenses conseguiram,

Capítulo III. Classificação de óleos vegetais

através de aperfeiçoamento genético, desenvolver espécies das plantas *Brassica nabus* e *Brassica compestris* com teores de ácido erúico e glucosilونات aceitáveis. No Brasil, a cultura da canola só teve início a partir de 1974, especificamente nos estados do Rio Grande do Sul e Santa Catarina^[86-87].

O óleo de canola é um dos mais saudáveis. Quando comparado com os óleos de girassol, milho e soja, é o mais rico em gorduras insaturadas e apresenta o menor índice de gorduras saturadas. Conseqüentemente, o consumo desse óleo ajuda a controlar o nível de colesterol e combater os problemas do coração^[86,89].

• Óleo de girassol

Girassol, nome botânico *Helianthus annus* L., é matéria-prima para a fabricação de um óleo muito apreciado no mundo inteiro. Países como a Rússia, Estados Unidos, Argentina e China são os maiores produtores mundiais de óleo de girassol. No Brasil, o consumo ainda é pequeno, mas com intuito de atender a demanda comercial, o país vem importando cada vez mais grãos para produção e comercialização do óleo^[89].

O óleo de girassol apresenta alto índice de ácido linoléico (ômega 6) e Tocoferóis (Vitamina E), que auxiliam na redução dos níveis de colesterol do sangue. Devido à sua qualidade nutricional, o óleo de girassol pode ser considerado nobre. Além disso, existem outras aplicações com diferentes finalidades, tais como: indústria de cosméticos, farmacêutica, veterinária, alimentícia, fabricação de tintas, sabões, entre outras^[86].

• Óleo de milho

O milho (*Zea mays* L.) é um cereal rico em amido e proteínas. A sua maior utilização é destinada para a alimentação animal. Depois da soja, o milho é que apresenta o maior índice de produção no País. Em todo o mundo o Brasil ocupa um lugar de destaque no cultivo desse cereal, perdendo apenas para os Estados Unidos e China^[87,89].

Os principais ácidos graxos que compõem o óleo de milho são o linoléico e oléicos que podem chegar até 62% e 42%, respectivamente^[86]. Entre os quatro óleos analisados nesse trabalho, o óleo de milho é aquele que apresenta a menor proporção do ácido linolênico tri-insaturado, ácido graxo fortemente susceptível ao processo de oxidação^[86-87].

Capítulo III. Classificação de óleos vegetais

• Óleo de soja

Entre as várias oleaginosas cultivadas no Brasil, a soja (*Glycine max* L.) se destaca por apresentar a maior produção. De fato, esses óleos apresentam, em média, preços mais acessíveis e são freqüentemente consumidos pela população. Além de ser utilizado para fins alimentícios e em outros setores, o óleo de soja vem se destacando recentemente na produção do biodiesel^[90].

No que diz respeito à composição química do grão de soja, podem ser encontrados cerca de 40% de proteínas, 20% de lipídeos, 5% de minerais e 34% de carboidratos. Os principais ácidos graxos presentes nesse óleo são o linoléico, oléico, palmítico e linolênico^[86-87, 89].

As características físico-químicas, bem como os ácidos graxos presentes nesses quatro óleos vegetais refinados acima apresentados podem ser encontrados na **Tabela 3.1**.

Tabela 3.1. Características físico-químicas e composição em ácidos graxos dos óleos refinados de canola, girassol, milho e soja^[88].

	Canola	Girassol	Milho	Soja	
Características físico-químicas	Índice de Refração (40°C)	1,465-1,467	1,467-1,469	1,465-1,468	1,466-1,470
	Índice de Iodo (Wijs)	110-126	110-143	103-128	120-143
	Índice de Saponificação	182-193	188-194	187-195	189-195
	Matéria Insaponificável (g/100g)	< 2,0	< 1,5	< 2,8	< 1,5
	Acidez (g de ácido oléico/100g)	< 0,3	< 0,3	< 0,3	< 0,3
	Índice de Peróxido (meq/Kg)	< 10,0	< 10,0	< 10,0	< 10,0
	Brassicasterol (g/100g)	> 5,0	—	—	—
Ácidos graxos (g/100g)	C < 14	—	< 0,4	< 0,3	< 0,1
	Mirístico - C14:0	< 0,2	< 0,5	< 0,1	< 0,5
	Palmítico - C16:0	2,5-6,5	3,0-10,0	9,0-14,0	7,0-14,0
	Palmitoléico - C16:1	< 0,6	< 1,0	< 0,5	< 0,5
	Esteárico - C18:0	0,8-3,0	1,0-10,0	0,5-4,0	1,4-5,5
	Oléico (ω9) - C18:1	53,0-70,0	14,0-35,0	24,0-42,0	19,0-30,0
	Linoléico (ω6) - C18:2	15,0-30,0	55,0-75,0	34,0-62,0	44,0-62,0
	Linolênico (ω3) - C18:3	5,0-13,0	< 0,3	< 2,0	4,0-11,0
	Araquídico - C20:0	0,1 - 1,2	< 1,5	< 1,0	< 1,0
	Eicosenóico - C20:1	0,1 - 4,3	< 0,5	< 0,5	< 1,0
	Behênico - C22:0	< 0,6	< 1,0	< 0,5	< 0,5
	Lignocérico - C24:0	< 0,2	< 0,5	< 0,5	—
	Erúico- C22:1	< 2,0	< 0,5	—	—
	Tetracosenóico - C24:1	< 0,2	< 0,5	—	—

Capítulo III. Classificação de óleos vegetais

A autenticidade dos óleos vegetais comestíveis tem se tornado um assunto de grande importância, não apenas pelos fatores associados à saúde dos consumidores, mas também por razões econômicas. De fato, propriedades benéficas e adversas para a saúde humana dependem, entre outros fatores, do tipo e da qualidade de óleo consumido. Além disso, o valor agregado do produto varia de acordo com os custos associados à matéria-prima, processamento, refino, engarrafamento, transporte, estocagem, entre outros. Conseqüentemente, óleos de maior qualidade e mais caros podem, em muitos casos, ser alvos de falsificação ou adulteração com óleos de menor valor comercial.

Assim, um grande esforço tem sido empreendido por parte de vários pesquisadores no tocante ao desenvolvimento de novas metodologias analíticas que possam, de forma eficiente e segura, caracterizar e/ou autenticar óleos vegetais comestíveis. Nesse contexto, aplicações envolvendo a voltametria de onda quadrada^[29], assim como as espectrometrias NIR^[91], FTIR^[92] e Raman^[93] têm mostrado ser alternativas úteis frente aos métodos clássicos de análise (cromatográfica líquida e gasosa)^[91,94].

Neste capítulo, uma nova estratégia de classificação de quatro tipos de óleos vegetais comestíveis é proposta. Especificamente, a espectrometria UV-VIS é adotada para enfatizar a habilidade do SPA-LDA em lidar com sinais analíticos de baixa resolução, fortes sobreposições e baixa correlação entre espectro e estrutura molecular.

3.2. Objetivos

Avaliar o uso do SPA-LDA com a espectrometria UV-VIS para a classificação de óleos vegetais comestíveis (canola, girassol, milho e soja);

Comparar o SPA-LDA com o GA-LDA e com o SIMCA (em diferentes níveis de significância para o Teste- F : 1%, 5%, 10% e 25%) em função do número de erros obtidos para um conjunto externo de amostras (teste);

Avaliar os modelos SPA-LDA, GA-LDA e SIMCA quanto à sensibilidade ao ruído instrumental.

3.3. Experimental

3.3.1. Amostras

Cento e dezenove amostras de óleos vegetais comestíveis, de diferentes lotes e fabricantes foram adquiridas dos supermercados da cidade de João Pessoa, Paraíba. A **Tabela 3.2** mostra o número e as classes de amostras utilizadas nesse estudo.

Tabela 3.2. Classes e quantidade de amostras de óleos vegetais analisadas.

Classes	Número de amostras
Canola	29
Girassol	31
Milho	29
Soja	30
Total	119

3.3.2. Equipamentos

O sistema para registro dos espectros UV-VIS das amostras de óleos vegetais é apresentado na **Figura 3.1**. Ele é composto de um espectrofotômetro de absorção molecular UV-VIS HP 8453, modelo Hewlett Packard, de uma cubeta de fluxo de quartzo com 1 mm de caminho óptico e de uma bomba peristáltica Gilson (modelo Miniplus 3) equipada com tubos de PVC (1,85 mm de diâmetro).

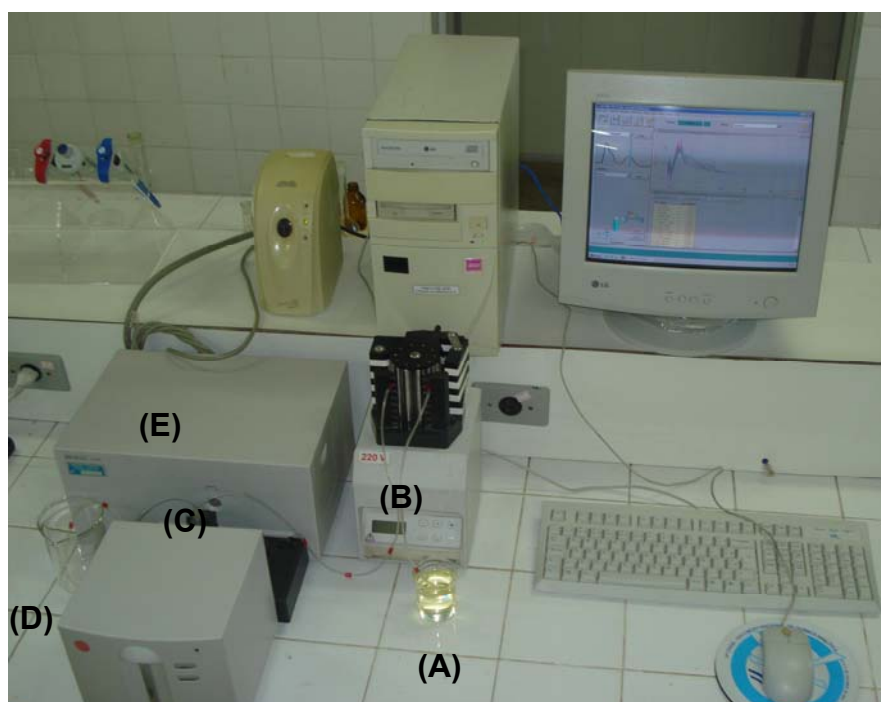


Figura 3.1. Sistema montado para o registro dos espectros de óleos vegetais. (A): amostra; (B): bomba peristáltica; (C): cubeta de fluxo; (D): descarte e (E): espectrofotômetro UV-VIS.

Capítulo III. Classificação de óleos vegetais

3.3.3. Procedimento analítico

Antes do registro dos espectros, as amostras foram diluídas com álcool n-butílico (99 % m/m) na proporção de 1:300 (v/v).

Inicialmente, o álcool n-butílico foi aspirado com o auxílio da bomba peristáltica e o espectro do branco foi obtido. Posteriormente, cada amostra de óleo diluída foi aspirada e o espectro foi registrado na região de 220 nm a 400 nm, com 1 nm de resolução. Cada espectro resultante apresentou 181 pontos (variáveis).

3.3.4. Tratamento dos dados e softwares

Antes da construção dos modelos, os dados foram divididos em três subconjuntos (treinamento, validação e teste) com o uso do algoritmo Kennard-Stone (KS)^[95]. Nesse algoritmo, as distâncias euclidianas entre os vetores das respostas instrumentais (\mathbf{x}) das amostras selecionadas são maximizadas.

Para uma melhor compreensão da forma como o algoritmo KS executa a busca por amostras mais representativas para o conjunto de treinamento, uma representação gráfica simplificada pode ser encontrada na figura abaixo. Os detalhes de cada passo são apresentados em seguida.

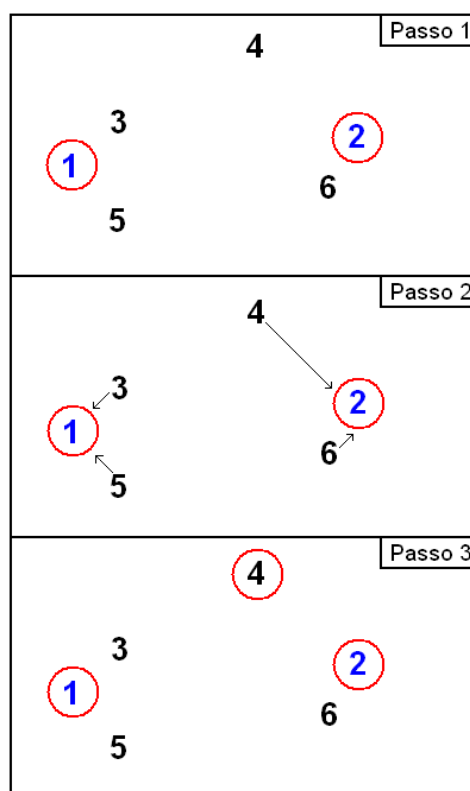


Figura 3.2. Representação gráfica do mecanismo de busca do algoritmo KS com três amostras selecionadas.

Capítulo III. Classificação de óleos vegetais

- **Passo 1:** Selecionam-se as duas amostras que apresentarem a maior distância euclidiana entre si. Nesse caso, as amostras 1 e 2.
- **Passo 2:** As menores distâncias entre as amostras remanescentes e as selecionadas são calculadas: D_{3-1} ; D_{5-1} ; D_{4-2} e D_{6-2} .
- **Passo 3:** Seleciona-se aquela amostra cuja distância obtida no passo anterior for maior. Nesse passo, a amostra 4 é, portanto, selecionada.

O procedimento descrito acima é repetido até que um número de amostras estipulado pelo analista seja alcançado.

O algoritmo foi aplicado nesse capítulo separadamente para cada classe. Os objetos do conjunto de treinamento foram inicialmente selecionados e os remanescentes foram divididos em validação e teste, de acordo com a ordenação do KS. As amostras de treinamento e validação foram utilizadas para o procedimento de modelagem (incluindo a seleção de variáveis para os modelos LDA e a determinação do número de PCs para os modelos SIMCA.) Já o conjunto externo de teste foi utilizado apenas para uma avaliação final dos modelos de classificação (SPA-LDA, GA-LDA e SIMCA).

O número de amostras para cada conjunto pode ser encontrado na **Tabela 3.3**.

Tabela 3.3. Número de amostras de treinamento, validação e teste selecionadas pelo KS para as quatro classes de óleos vegetais.

Classes	Conjuntos		
	Treinamento	Validação	Teste
Canola	12	6	11
Girassol	12	6	13
Milho	12	6	11
Soja	12	6	12
Total	48	24	47

O GA utilizado nesse estudo empregou cromossomos binários padrão com tamanho igual ao número de comprimentos de onda do espectro (o gene “1” indica o comprimento de onda selecionado)^[55]. A aptidão para cada variável foi considerada como o inverso do risco G (**Equação 2.22**) calculado usando os comprimentos de onda codificados no cromossomo. A probabilidade de um determinado indivíduo ser selecionado foi proporcional a sua aptidão (método da roleta). Operadores de cruzamento e mutação foram estabelecidos com probabilidades de 60% e 10%, respectivamente. O tamanho da população foi mantido constante e cada geração foi completamente substituída pelos seus descendentes. Entretanto, o melhor indivíduo foi automaticamente transferido para a próxima geração (elitismo) para evitar o

desperdício de boas soluções. O GA foi aplicado utilizando 100 gerações com 200 cromossomos cada. Além disso, o algoritmo foi repetido três vezes, a partir de populações diferentes. A melhor solução resultante dessas três realizações foi mantida e adotada como resultado do GA para este estudo de classificação.

O programa Unscrambler® 9.6 (CAMO S.A.) foi utilizado para a realização da PCA e construção dos modelos SIMCA. O número de PCs utilizado para cada modelo SIMCA foi encontrado estipulando um percentual de variância explicada maior ou igual a 95%.

As rotinas de classificação do SPA-LDA e GA-LDA e o algoritmo KS foram implementados no Matlab 6.5.

3.4. Resultados e Discussão

3.4.1. Espectros dos óleos vegetais

A **Figura 3.3** mostra os espectros de absorção molecular dos quatro tipos de óleos vegetais analisados na região de 220 nm a 400 nm.

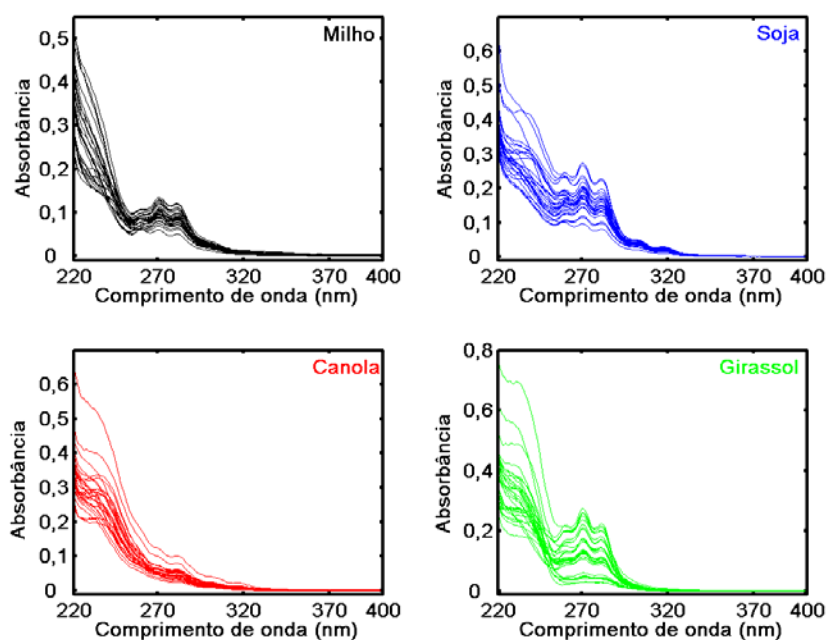


Figura 3.3. Espectros UV-VIS das amostras de óleos vegetais comestíveis analisados.

Como pode ser visto, as absorções ocorrem mais frequentemente na faixa do UV. Os óleos de girassol, soja e milho mostram bandas de absorção bem caracterizadas em torno de 260, 270 e 280 nm. Tal característica não é tão aparente nos espectros de canola. Quando comparados com os óleos de girassol e soja, os espectros de canola e milho apresentam claramente uma menor dispersão dentro da

Capítulo III. Classificação de óleos vegetais

classe. Especificamente, a maior variabilidade entre as amostras ao longo de toda faixa espectral pode ser encontrada para o óleo de girassol.

3.4.2. Análise exploratória dos dados

Com intuito de se realizar uma avaliação exploratória dos diferentes tipos de óleos estudados, uma PCA foi realizada para todo o conjunto de amostras. A **Figura 3.4** e **Figura 3.5** mostram os gráficos dos escores obtidos por PC2 *versus* PC1 e PC3 *versus* PC1, respectivamente.

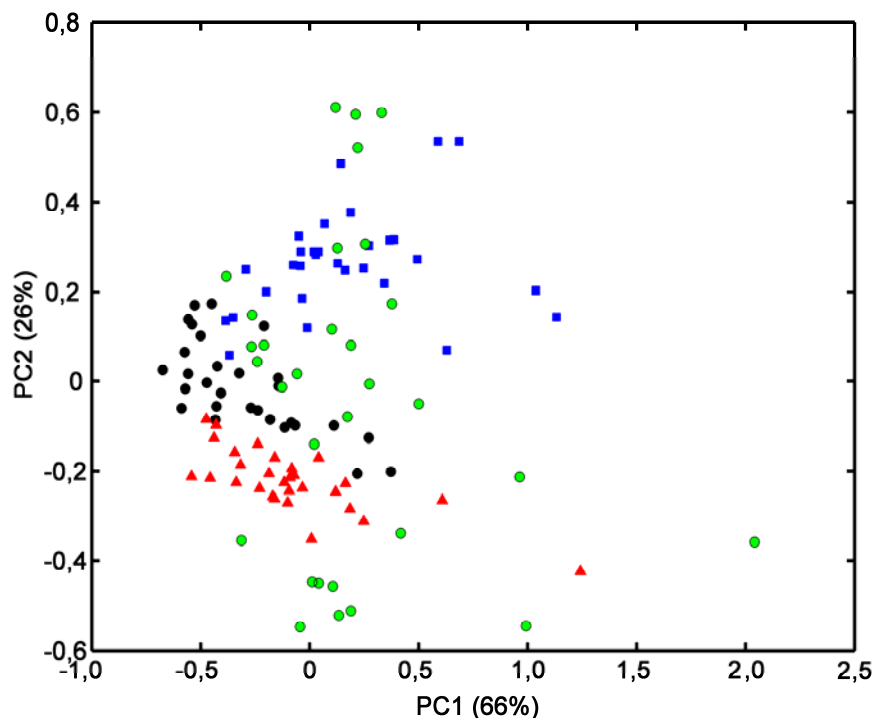


Figura 3.4. Gráfico dos escores obtidos pela PC2 *versus* PC1 para todas as 119 amostras de óleos vegetais. (●: milho, ●: girassol, ▲: canola e ■: soja).

O gráfico dos escores obtidos por PC2 *versus* PC1 mostra uma forte sobreposição entre as quatro classes de óleos (milho, girassol, canola e soja). Observa-se também que as classes girassol e soja apresentam-se com uma maior dispersão ao longo de PC1, variável latente com maior percentual de variância explicada (66%). O resultado obtido pela PCA está de acordo com o comportamento dos espectros apresentados na **Figura 3.3**, uma vez que as classes girassol e soja apresentam perfis espectrais bem parecidos e com uma maior variabilidade dentro da classe. Um total de 92% da variância dos dados é explicado por PC1 e PC2.

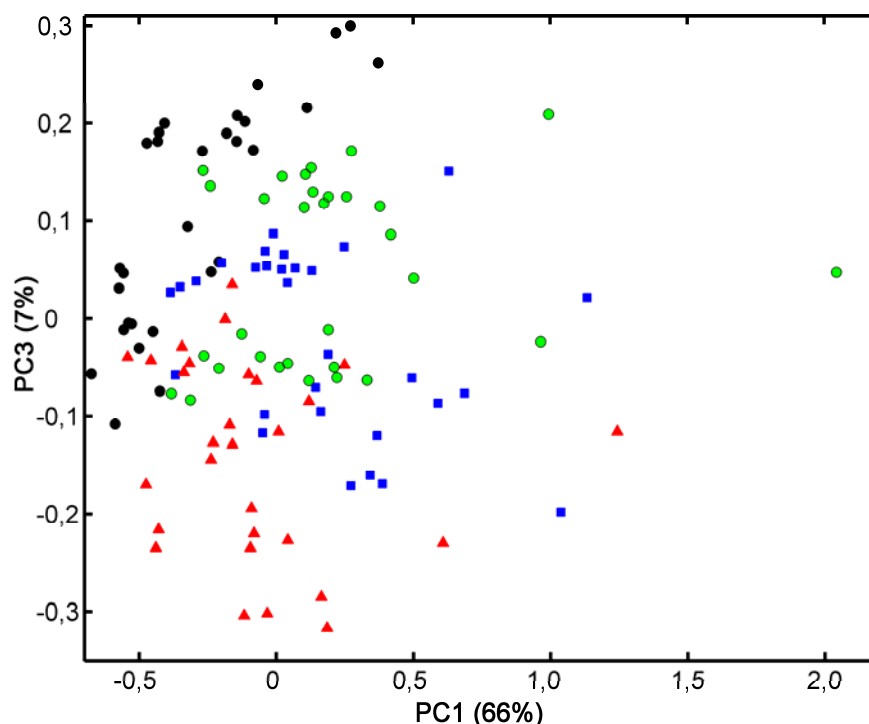


Figura 3.5. Gráfico dos escores obtidos pela PC3 *versus* PC1 para todas as 119 amostras de óleos vegetais. (●: milho, ●: girassol, ▲: canola e ■: soja).

No gráfico dos escores obtidos por PC3 *versus* PC1, uma grande sobreposição entre as classes é novamente observada. A PC3, apesar de explicar 7% dos dados, não revela tendências de separação entre quaisquer grupos. Desse modo, pode-se concluir que a PCA aplicada aos espectros UV-VIS não permitiu obter uma discriminação apropriada das classes para este estudo.

3.4.3. Classificação SIMCA

Modelos SIMCA foram construídos individualmente para classe de óleo utilizando a série de teste como técnica de validação. Quatro níveis de significância para o teste-*F* foram avaliados: 1%, 5%, 10% e 25%. Uma tabela resumida contendo os erros para o conjunto de teste é apresentada a seguir. O número de PCs adotado para cada modelo é indicado entre parênteses.

Tabela 3.4. Erros de classificação SIMCA para o conjunto de Teste de óleos vegetais em diferentes níveis de significância do Teste-*F* (1%, 5%, 10% e 25%).

Modelo	Milho (3 PCs)				Soja (3 PCs)				Canola (5 PCs)				Girassol (3 PCs)				
	Nível (%)	1	5	10	25	1	5	10	25	1	5	10	25	1	5	10	25
Milho	-	-	-	4	10	7	4	-	-	-	-	-	-	4	-	-	-
Soja	1	-	-	-	-	-	-	4	-	-	-	-	-	-	-	-	-
Canola	-	-	-	-	-	-	-	-	-	-	1	1	-	-	-	-	-
Girassol	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Na **Tabela 3.4**, os valores localizados nas células com tonalidade cinza indicam o número de erros do Tipo I. Podem-se citar, como exemplo, as quatro amostras da classe de soja que não foram corretamente classificadas pelo modelo soja construído com 3 PCs para um nível de significância de 25%.

Com base nos resultados apresentados pelos modelos SIMCA, pode-se perceber que o número de erros do Tipo I é baixo e só ocorre para os níveis de significância de 10% e 25%. Contudo, os erros do Tipo II são mais frequentes, sobretudo para 1% (19 erros). É importante observar que o número de erros do Tipo I aumenta e do Tipo II diminui, conforme o aumento do nível de significância do Teste-*F*. O resumo de ambos os erros obtidos pelos modelos SIMCA para o conjunto de Teste será apresentado mais adiante (**Tabela 3.5**) numa comparação com o SPA-LDA e GA-LDA.

3.4.4. SPA-LDA

O gráfico *scree* fornecido pelo SPA mostra o número de variáveis *versus* o custo da validação (**Figura 3.6**). Como resultado, apenas sete comprimentos de onda foram selecionados e todas as amostras do conjunto de validação foram corretamente classificadas.

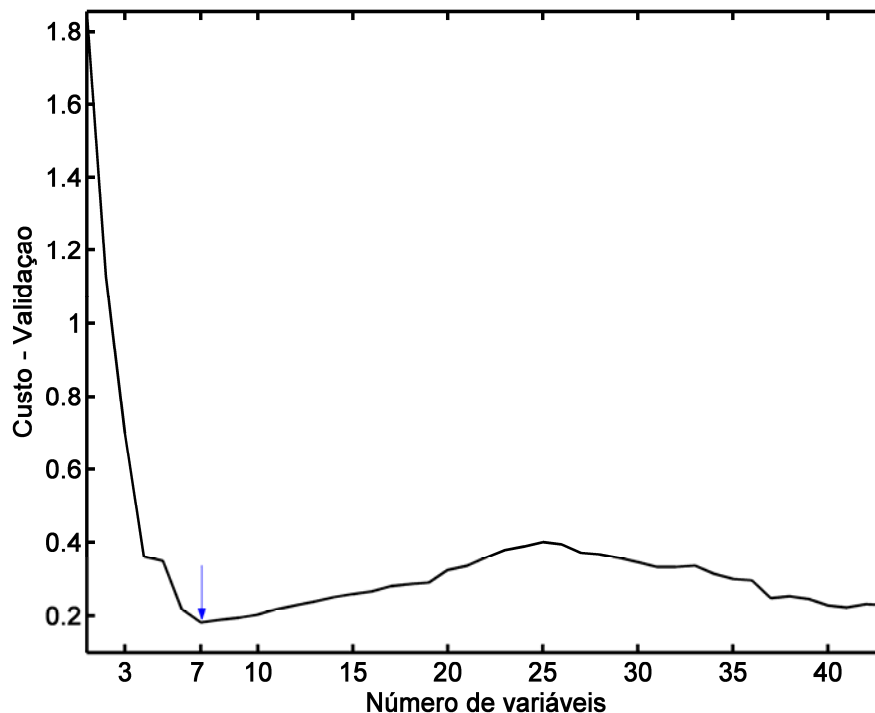


Figura 3.6. Custo da validação em função do número de variáveis selecionadas pelo SPA-LDA para o conjunto de dados de óleos vegetais. A seta indica o ponto de mínimo da curva do custo (0.1817), o qual ocorre em sete comprimentos de onda.

Capítulo III. Classificação de óleos vegetais

A **Figura 3.7** apresenta os espectros médios de cada classe de óleo vegetal com a indicação dos sete comprimentos de onda selecionados pelo SPA. Como se pode notar, os comprimentos de onda selecionados pelo SPA-LDA estão realmente associados a pontos característicos dos espectros (picos, vales, ombros e inflexões). Nenhuma variável foi selecionada em regiões de baixa intensidade de absorção. É interessante observar que os dois picos em torno de 260 nm e 280 nm não foram incluídos na seleção do SPA-LDA. Possivelmente, isto ocorreu devido à colinearidade entre estes picos com o pico correspondente à variável selecionada em 269 nm.

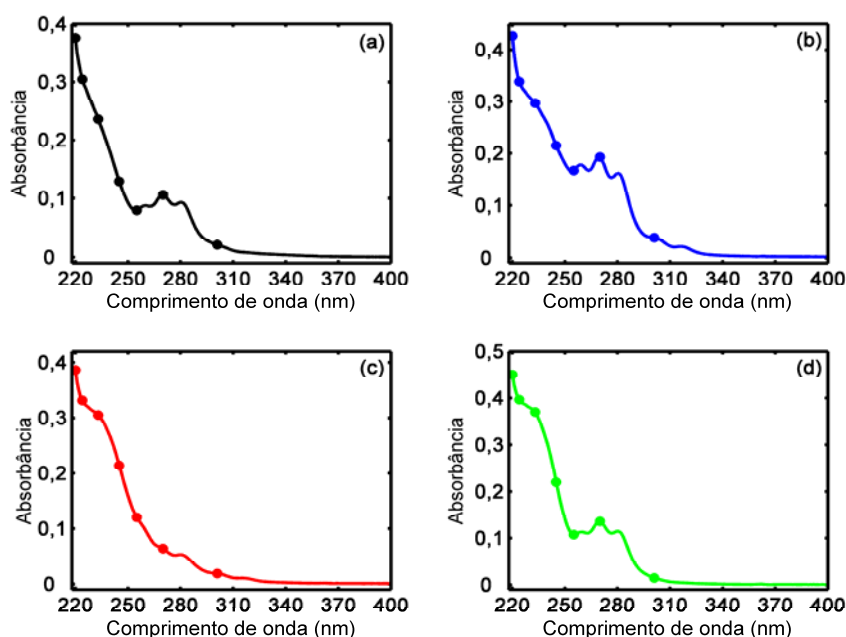


Figura 3.7. Espectro médio para cada tipo de óleo vegetal analisado. (a) milho, (b) soja, (c) canola e (d) girassol.

O modelo LDA obtido com as sete variáveis selecionadas pelo SPA foi, então, aplicado à classificação de amostras do conjunto de teste. Como resultado, apenas uma amostra foi classificada incorretamente (soja classificada como canola).

Os gráficos dos escores obtidos pela LDA utilizando os sete comprimentos de onda selecionados pelo SPA são apresentados nas **Figuras 3.8-3.9**.

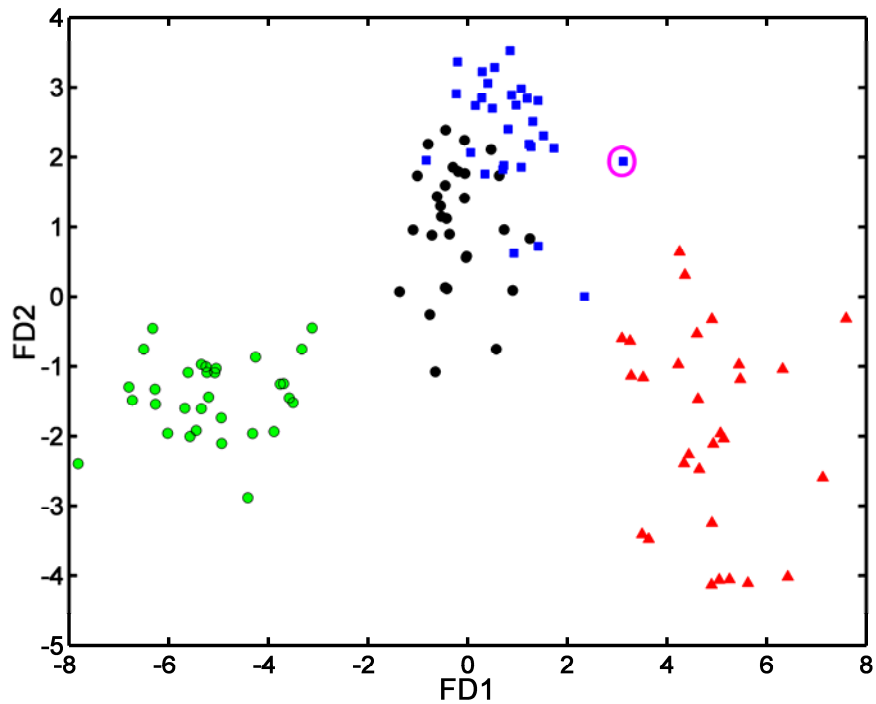


Figura 3.8. Gráfico dos escores da Função Discriminante 2 (FD2) versus Função Discriminante 1 (FD1) para todas as amostras de óleos vegetais (●: milho, ●: girassol, ▲: canola e ■: soja).

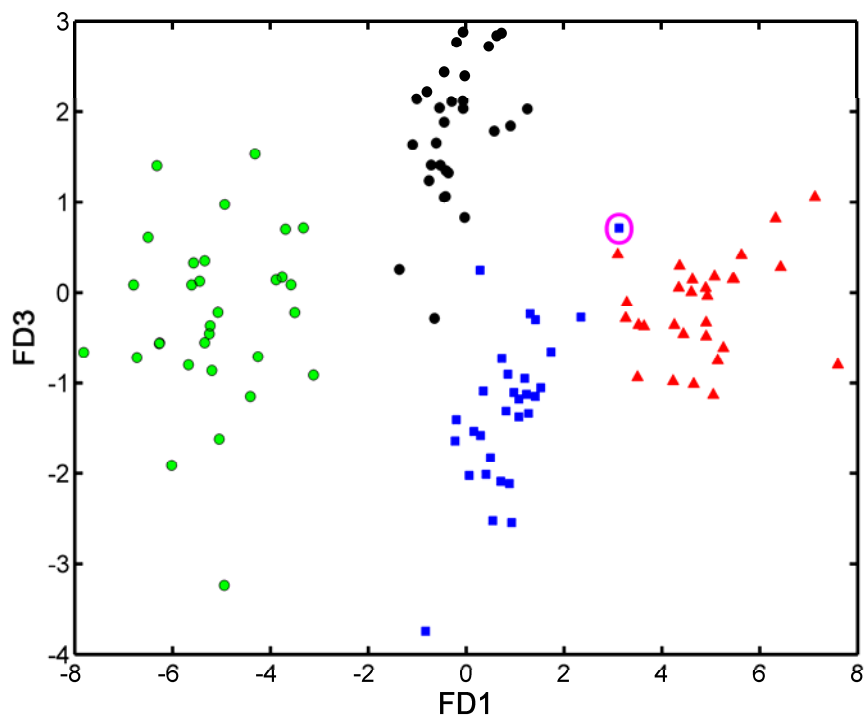


Figura 3.9. Gráfico dos escores da Função Discriminante 3 (FD3) versus Função Discriminante 1 (FD1) para todas as amostras de óleos vegetais (●: milho, ●: girassol, ▲: canola e ■: soja).

Diferentemente dos gráficos dos escores obtidos pela PCA (**Figura 3.4** e **Figura 3.5**), os resultados apresentados pela LDA revelam claramente uma discriminação entre as classes de óleos vegetais estudadas. De fato, é possível

observar na **Figura 3.8** que a FD1 é responsável por separar três grandes grupos: Girassol, Canola e um último formado pelas classes Milho e Soja. No gráfico dos escores obtidos pela FD3 × FD1 (**Figura 3.9**), uma maior separação das quatro classes pode ser observada. A amostra marcada com o círculo róseo foi aquela incorretamente classificada.

Além dos bons resultados obtidos em função do número de erros para o conjunto de teste, os dois gráficos apresentados pelas **Figuras 3.8-3.9** demonstram a eficiência do SPA-LDA em selecionar variáveis espectrais que sejam suficientemente discriminantes para o estudo de classificação de óleos vegetais comestíveis.

3.4.5. GA-LDA

Para propósito de comparação, o GA também foi empregado para seleção de variáveis em LDA. A **Figura 3.10** mostra os espectros médios de cada classe de óleo vegetal analisada com a indicação dos 16 comprimentos de onda selecionados pelo GA-LDA.

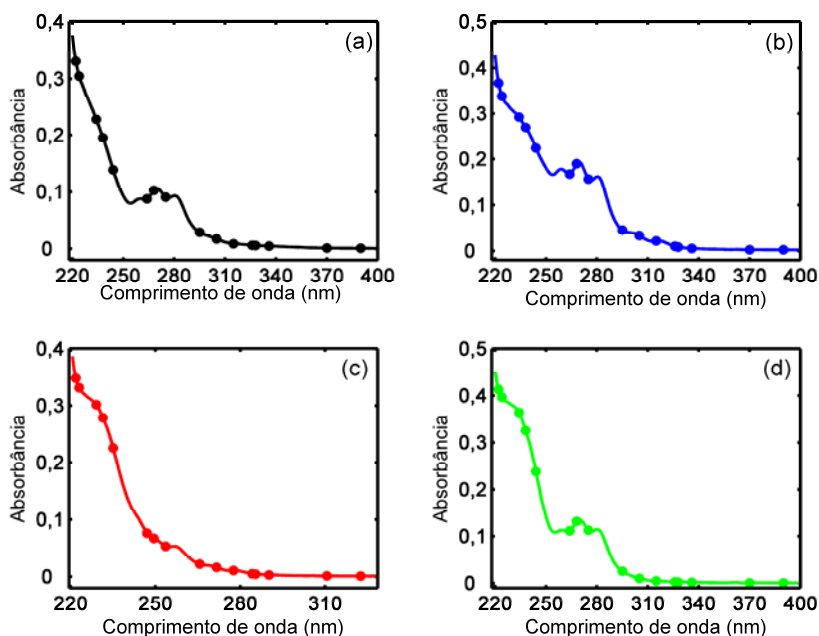


Figura 3.10. Espectro médio para cada tipo de óleo vegetal analisado. (a) milho, (b) soja, (c) canola e (d) girassol. Os dezesseis comprimentos de ondas selecionados pelo GA-LDA encontram-se indicados com círculos.

O modelo LDA resultante construído com as 16 variáveis selecionadas pelo GA classificaram corretamente todas as amostras para o conjunto de validação e teste. Contudo, é importante ressaltar que, diferentemente da solução obtida pelo

Capítulo III. Classificação de óleos vegetais

SPA, alguns dos comprimentos de onda selecionados pelo GA estão localizados em regiões onde nenhuma informação é evidenciada, incluindo uma parte do visível, onde a baixa relação sinal/ruído é pronunciada. A inclusão destas variáveis apresenta efeitos adversos, sobretudo no que diz respeito à sensibilidade do modelo LDA ao ruído instrumental. Este estudo será abordado na próxima seção.

A **Tabela 3.5** apresenta o resumo dos erros de classificação para o conjunto de teste de óleos vegetais.

Tabela 3.5. Resumo dos resultados (erros de classificação no conjunto de teste) para o SPA-LDA, GA-LDA e SIMCA (4 níveis de significância do teste-*F*) para o conjunto de dados de óleos vegetais.

	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

Fica claro a superioridade dos modelos LDA com seleção de variáveis (SPA e GA) frente aos modelos SIMCA (nos quatro níveis de significância do teste-*F*). Entre os resultados apresentados pelo SIMCA, a classificação realizada com 10% de nível de significância apresentou o melhor desempenho, em termos do número total de erros. É importante lembrar que quando uma classificação é realizada na LDA, a amostra que não for corretamente classificada na sua classe verdadeira, obrigatoriamente será incluída em uma errada. Portanto, o SPA-LDA apresentou um erro do Tipo I (amostra de soja não classificada como pertencente ao modelo Soja), que, conseqüentemente, também se qualifica como erro do tipo II (amostra de soja classificada como canola).

3.4.6. Análise de sensibilidade ao ruído

Nesta seção, um estudo de sensibilidade dos modelos LDA e SIMCA com respeito ao ruído instrumental foi realizado. Para isso, os espectros do conjunto de teste foram artificialmente contaminados com ruído gaussiano branco de média zero. O desvio padrão do ruído adicional foi da ordem de 10^{-3} . Então, os modelos previamente obtidos (sem o ruído adicional) foram aplicados para a classificação do novo conjunto externo de amostras.

A **Tabela 3.6** apresenta uma comparação entre os resultados de classificação dos modelos SPA-LDA, GA-LDA e SIMCA para o novo cenário. Os números entre parênteses indicam os resultados obtidos anteriormente (sem a adição do ruído). O nível de significância do teste-*F* empregado para o SIMCA foi

Capítulo III. Classificação de óleos vegetais

aquele que apresentou o melhor desempenho de classificação no estudo inicial (10%).

Tabela 3.6. Resumo dos resultados de classificação (Erros do Tipo I e Tipo II) obtidos pelos modelos SPA-LDA, GA-LDA e SIMCA no conjunto de teste de óleos vegetais contaminado pelo ruído.

	SPA-LDA	GA-LDA	SIMCA
Amostras	2 (2)	38 (0)	26 (5)

Como pode ser visto, o desempenho da previsão do modelo SPA-LDA não foi afetado pela introdução do ruído. Em contrapartida, o número de erros obtidos pelo GA-LDA após a inserção do ruído foi consideravelmente maior, passando de zero para 38 erros. Possivelmente, isto aconteceu porque algumas variáveis selecionadas pelo GA são pouco informativas, quando comparadas com as selecionadas pelo SPA. Conforme apresentado na **Figura 3.10**, algumas variáveis selecionadas pelo GA encontram-se em regiões com baixa relação sinal/ruído (próximo à faixa do visível). Então, a adição do ruído tornou-se muito mais prejudicial para o GA-LDA do que para o SPA-LDA. Finalmente, para o método SIMCA (para um nível de significância do teste- F de 10%), um comportamento similar ao GA foi obtido, tendo o número de erros aumentado de 5 para 26 após a adição do ruído.

3.5. Considerações Finais

Neste capítulo, foi apresentada a primeira aplicação do SPA-LDA em um problema de classificação envolvendo quatro tipos de óleos vegetais comestíveis. A espectrometria UV-VIS foi adotada para enfatizar a capacidade do SPA-LDA em lidar com sinais analíticos de baixa resolução, fortes sobreposições e baixa correlação entre espectro e estrutura molecular.

Quando comparado com o SIMCA, método de classificação freqüentemente utilizado, o SPA-LDA resultou em um número menor de erros para um conjunto de teste independente.

Uma comparação foi também realizada entre o SPA e o GA utilizando a mesma função de custo (Risco médio G de uma classificação incorreta pela LDA). O desempenho de classificação destes dois modelos foi semelhante. Contudo, o SPA-LDA foi consideravelmente mais robusto à introdução de um ruído extra nos espectros. Assim, pode-se concluir que o SPA-LDA deverá ser mais apropriado do que o GA-LDA se o modelo de classificação for usado em conjunção com espectrofotômetros de menor relação sinal/ruído em análises de rotina.

CAPÍTULO IV
CLASSIFICAÇÃO DE ÓLEOS DIESEL

4. CLASSIFICAÇÃO DE ÓLEOS DIESEL

4.1. Introdução

4.1.1. Óleo diesel

O óleo diesel é um combustível derivado do petróleo, que se apresenta na forma de um líquido viscoso, límpido, pouco volátil, de cheiro forte, nível de toxicidade mediano e formado basicamente por carbono e hidrogênio, com baixas concentrações de enxofre, nitrogênio e oxigênio^[96].

No Brasil, o óleo diesel é um dos combustíveis mais utilizados e sua produção é realizada a partir do refino do petróleo, pelo processo inicial de destilação fracionada, a temperatura entre 260°C e 340°C. Seu uso é destinado, entre outras finalidades, para gerar energia e movimentar máquinas e motores de grande porte, em motores de combustão interna e ignição por compressão (motores do ciclo diesel), tais como: caminhões, automóveis de passeios, ônibus, pequenas embarcações marítimas, locomotivas, navios, tratores, etc^[96].

O controle de qualidade do óleo diesel é baseado em parâmetros como teor de enxofre, temperaturas de destilação, corrosividade ao cobre, água e sedimentos, índice de cetano, densidade, viscosidade, entre outros. Para a determinação desses parâmetros, métodos estabelecidos de acordo com as normas da ASTM (American Society for Testing and Materials) são empregados^[70].

Certamente, o teor de enxofre é um dos principais parâmetros de qualidade avaliados em amostras de óleo diesel, uma vez que o enxofre é um elemento terminantemente indesejável em qualquer combustível devido à ação corrosiva de seus compostos e à formação de gases tóxicos como o SO₂ e SO₃, que ocorre durante a combustão do produto. Na presença de água, o SO₃ leva à formação de ácido sulfúrico (H₂SO₄), que é, além de poluente, altamente corrosivo para as partes metálicas dos equipamentos automotivos^[70].

A Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP), segundo a Portaria Nº 310 de 27 de dezembro de 2001^[97], classificou o óleo diesel automotivo em dois tipos:

- **Óleo diesel automotivo metropolitano (D).**

O óleo diesel metropolitano apresenta um índice menor de enxofre para minimizar os poluentes resultantes da combustão. Como o próprio nome indica, este

Capítulo IV. Classificação de óleos diesel

óleo é utilizado em regiões que possuem maiores frotas de automóveis em circulação, tais como em algumas capitais e em grandes cidades que necessitam de um maior controle ambiental^[96].

• Óleo diesel automotivo de interior (B)

Esse combustível contém uma percentagem maior de enxofre e é, certamente, o mais utilizado no Brasil, exceto para as regiões metropolitanas. Ele é usado em regiões onde não se tem um fluxo muito grande de veículos, como nas cidades do interior^[96].

A quantidade de enxofre presente no óleo diesel é, de fato, um assunto de grande importância, não apenas por razões econômicas, mas também para o meio ambiente. Então, o desenvolvimento de novas metodologias analíticas que sejam capazes de assegurar a qualidade do óleo diesel com respeito ao teor de enxofre torna-se de extrema utilidade.

Em 2008, uma polêmica em torno da redução do teor de enxofre no óleo diesel comercializado no país foi destaque em conferências e em debates promovidos por diferentes órgãos brasileiros. O Ministério Público Federal homologou em outubro de 2008 o acordo firmado com órgãos federais e estaduais, e representantes das produtoras de combustíveis e de automóveis para a redução da emissão de poluentes resultantes da queima do diesel.

Entre as técnicas analíticas empregadas para determinação de compostos de enxofre em amostras de diesel, podem-se citar: coulometria^[98], fluorescência de raios X^[99] e cromatografia com detecção por quimiluminescência^[100].

A espectrometria NIR, aliada à Quimiometria, vem se destacando nos últimos anos como uma metodologia eficiente, rápida e não destrutiva em análises de combustíveis. Especificamente para a determinação de enxofre em amostras de diesel, alguns trabalhos podem também ser encontrados^[70, 101].

Neste capítulo, o uso do SPA-LDA é avaliado juntamente com a espectrometria NIR para a classificação de amostras de óleo diesel com respeito ao teor de enxofre (baixo ou alto).

4.2. Espectrometria NIR

Em 1800, o alemão Frederick William-Herschel identificou a radiação no infravermelho próximo decompondo a luz por meio de um prisma de vidro e

Capítulo IV. Classificação de óleos diesel

movendo um termômetro através das cores para saber qual delas, no espectro, era responsável pelo calor produzido. Herschel observou, então, que existia uma pequena variação de temperatura depois do vermelho. Então, foi possível provar que havia radiação além do visível, mas esta descoberta foi fortemente ignorada até o início do desenvolvimento dos instrumentos mais modernos capazes de registrar os espectros. O uso da espectrometria NIR como técnica analítica de caráter quantitativo só começou em 1960 com o trabalho de Karl Norris, no Departamento de agricultura dos EUA^[102].

Certamente, a consolidação da espectrometria NIR como alternativa viável e compatível com as demais técnicas analíticas só ocorreu depois de 1980 com o avanço da eletrônica analógica e digital, da ciência dos materiais e, sobretudo, da Quimiometria. Atualmente, é possível encontrar um grande número de aplicações envolvendo o uso do NIR para análises rápidas e não-invasivas de alimentos, produtos farmacêuticos e agrícolas, polímeros, combustíveis, entre outras^[102].

A radiação NIR compreende a região delimitada pelos comprimentos de onda de 780 a 2500 nm do espectro eletromagnético. As bandas de absorção observadas nesta região são provenientes, quase que totalmente, de sobretons de transições vibracionais e de combinações das transições fundamentais associadas aos níveis energéticos vibracionais de grupos de átomos que ocorrem na região MIR^[103].

Geralmente, os espectros NIR resultam de transições vibracionais quantizadas associadas a átomos de C, O, N ou S ligados ao hidrogênio. Isto torna a técnica útil para a determinação de compostos orgânicos que contenham ligações C-H, N-H, O-H e S-H. Estas ligações, por serem de alta energia e por possuírem átomos de massa relativamente baixa, têm transições fundamentais na região do infravermelho médio^[102-103].

A natureza das bandas que caracterizam essa região confere uma intensidade de absorção cerca de 10 a 100 vezes inferior às absorções que ocorrem na região do infravermelho médio (MIR)^[102]. Somando-se a isso, o NIR apresenta bandas com alta sobreposição e de difícil interpretação. Estes fatores fazem com que o uso da espectrometria NIR, para realizar análises precisas e confiáveis, dependa quase que totalmente da utilização de métodos baseados em análise multivariada.

Capítulo IV. Classificação de óleos diesel

Em face do exposto, espera-se que o uso do SPA-LDA aplicado aos espectros NIR possa ser uma alternativa vantajosa para a classificação de amostras de óleo diesel com respeito ao teor de enxofre.

4.3. Objetivos

Avaliar a combinação da espectrometria de absorção molecular na região do NIR com o SPA-LDA para a classificação de óleos diesel com respeito ao teor de enxofre (baixo e alto);

Comparar o desempenho dos modelos SPA-LDA com o GA-LDA e SIMCA (em diferentes níveis de significância para o Teste-*F*: 1%, 5%, 10% e 25%) em função do número de erros para o conjunto externo de amostras (teste);

Avaliar os modelos SPA-LDA, GA-LDA e SIMCA quanto à sensibilidade ao ruído instrumental.

4.4. Experimental

4.4.1. Amostras

Cento e vinte e oito amostras de óleos diesel, coletadas de diferentes postos de combustíveis da região metropolitana da cidade do Recife, Pernambuco, foram utilizadas para esse estudo. O teor de enxofre presente nessas amostras foi determinado de acordo com o método de referência D4294-90 da ASTM. A faixa de concentração para essas determinações foi de 0,05-0,31% m/m.

Segundo a ANP^[97], o teor máximo de enxofre permitido para diesel metropolitano é de 0,20% m/m. Então, este valor foi adotado como limiar na divisão das classes. A **Tabela 4.1** mostra o número e as classes de amostras utilizadas nesse estudo.

Tabela 4.1. Classes e quantidade de amostras de óleo diesel analisadas

Classes	Número de amostras
Baixo teor de enxofre	69
Alto teor de enxofre	59
Total	128

4.4.2. Equipamentos

O sistema para registro dos espectros NIR dessas amostras é composto por um espectrômetro de Infravermelho FTIR, da Perkin Elmer, modelo Spectrum, série

Capítulo IV. Classificação de óleos diesel

GX, uma bomba peristáltica Gilson, um microcomputador Pentium e uma cubeta de quartzo de fluxo com 1 cm de caminho óptico.



(a)



(b)

Figura 4.1. (a) Espectrofotômetro FT-IR utilizado para o registros dos espectros de óleos diesel. (b) cubeta de fluxo de quartzo de 1 cm de caminho óptico.

4.4.3. Procedimento analítico

Inicialmente, foi registrado o espectro do branco apenas com o ar. O registro dos espectros NIR das amostras de óleo diesel foi realizado através do sistema apresentado na **Figura 4.1**. A bomba peristáltica foi utilizada para aspirar cada amostra até o preenchimento da cubeta de fluxo. Os espectros foram registrados com a média de dezesseis varreduras na região de 880 nm a 1600 nm e com uma resolução de 2 cm^{-1} . Antes do registro de cada espectro, pequenas quantidades de uma nova amostra foram aspiradas, em triplicatas, entre bolhas de ar para a realização da limpeza da cubeta. Durante a realização dessas medidas, a temperatura do ambiente e a umidade relativa do ar estavam em torno de $25,5 \text{ }^\circ\text{C}$ e 42%, respectivamente.

4.4.4. Softwares

Assim como no Capítulo III, o algoritmo KS foi utilizado para dividir os dados em três subconjuntos: treinamento, validação e teste. A **Tabela 4.2** apresenta o número e o tipo de amostra de óleo diesel para cada subconjunto.

Tabela 4.2. Número de amostras de treinamento, validação e teste selecionadas pelo KS para as duas classes de óleos diesel.

Classes	Conjuntos		
	Treinamento	Validação	Teste
Baixo teor de enxofre	26	15	28
Alto teor de enxofre	22	13	24
Total	48	28	52

Capítulo IV. Classificação de óleos diesel

As amostras de treinamento e validação foram utilizadas para o procedimento de modelagem (incluindo a seleção de variáveis para os modelos LDA e a determinação do número de PCs para os modelos SIMCA). Já o conjunto externo de teste foi utilizado apenas para uma avaliação final dos modelos de classificação (SPA-LDA, GA-LDA e SIMCA).

As configurações empregadas pelo GA foram idênticas àquelas utilizadas no capítulo anterior.

4.5. Resultados e Discussão

4.5.1. Espectros dos óleos diesel

A **Figura 4.2** mostra os espectros NIR de absorção molecular na região de 880 – 1600 nm das duas classes de óleo diesel empregadas nesse estudo.

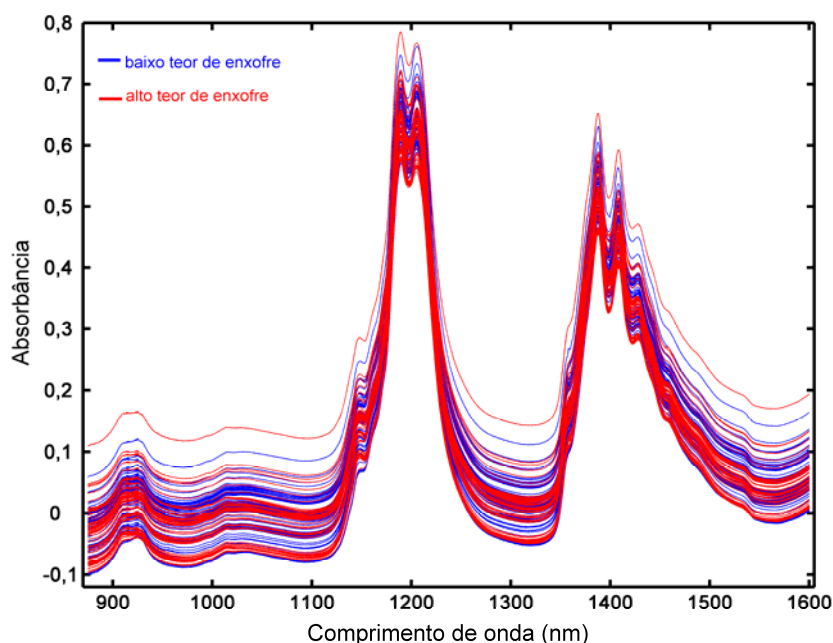


Figura 4.2. Espectro NIR originais das amostras de óleos diesel.

Como pode ser observado, existe uma grande variação sistemática da linha de base dos espectros originais ao longo de toda a região de trabalho. Com uma ampliação em cada espectro, pode-se também observar a presença de ruídos. Para corrigir esses problemas, a primeira derivada pelo método de suavização Savitzky-Golay^[16] foi empregada. Para esse procedimento, adotaram-se um polinômio de segunda ordem e uma janela de 11 pontos. A **Figura 4.3** mostra os espectros

Capítulo IV. Classificação de óleos diesel

derivativos resultantes das 128 amostras de óleos diesel. Após essa transformação, o número total de pontos resultantes foi de 1431 para cada espectro.

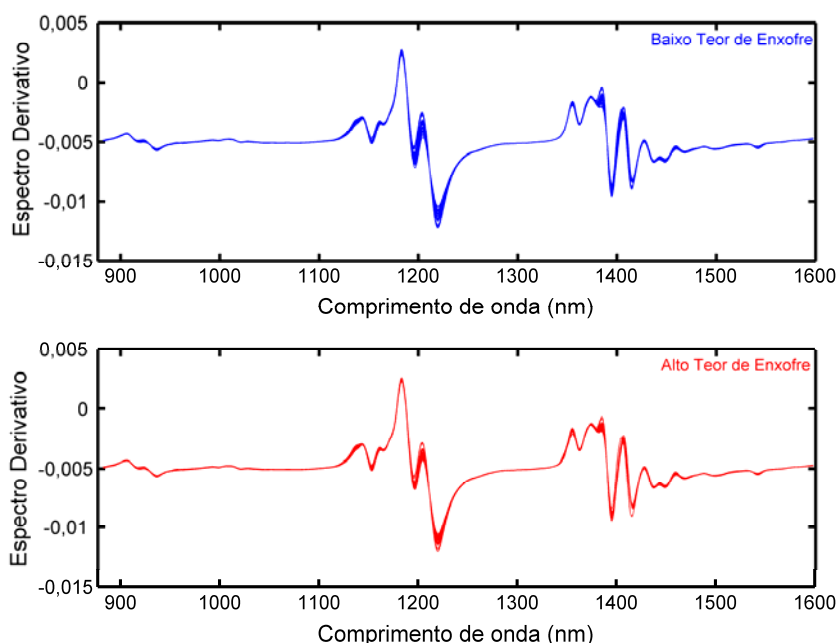


Figura 4.3. Espectros NIR derivativos das amostras de óleos diesel analisadas.

Diferentemente do estudo de caso apresentado no capítulo anterior, os espectros das duas classes envolvidas nesse problema de classificação são, de fato, muito parecidos. Esta sobreposição pode ser também evidente nos gráficos dos escores obtidos pela PCA que são apresentados na próxima seção.

4.5.2. Análise exploratória dos dados

Com intuito de realizar uma avaliação exploratória dos dois grupos de amostras, uma PCA foi aplicada aos 128 espectros derivativos. A **Figura 4.4** e **Figura 4.5** mostram os gráficos dos escores obtidos por PC2 *versus* PC1 e PC3 *versus* PC1, respectivamente.

O gráfico dos escores obtido por PC2 \times PC1 revela uma grande sobreposição entre as duas classes de óleos diesel envolvidas nesse problema. Este comportamento condiz com o perfil espectral dessas amostras, uma vez que poucas diferenças são observadas ao longo da faixa de trabalho. O gráfico dos escores obtidos por PC3 \times PC1 (**Figura 4.5**) mostra também esta tendência de sobreposição.

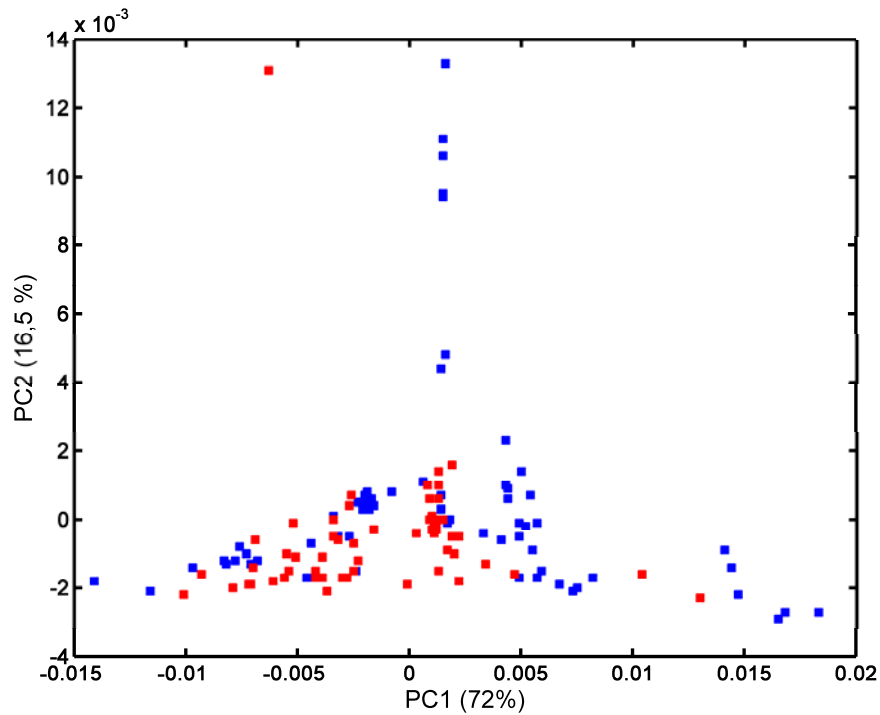


Figura 4.4. Gráfico dos escores obtidos pela PC2 versus PC1 para todas as 128 amostras de óleos vegetais. (■: baixo teor de enxofre e ■: alto teor de enxofre).

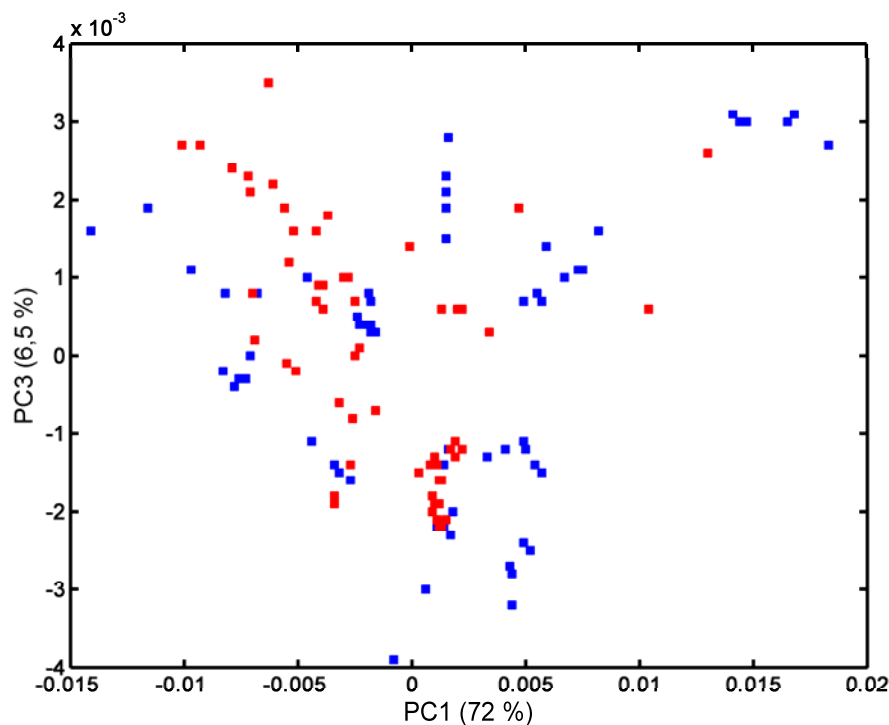


Figura 4.5. Gráfico dos escores obtidos pela PC3 versus PC1 para todas as 128 amostras de óleos diesel. (■: baixo teor de enxofre e ■: alto teor de enxofre).

Após a avaliação dos dois gráficos de escores apresentados pela **Figura 4.4 - 4.5**, pode-se concluir que a PCA, aplicada aos espectros NIR de óleos

Capítulo IV. Classificação de óleos diesel

diesel, mostrou-se pouco eficiente para obter uma discriminação entre as duas classes envolvidas nesse estudo.

4.5.3. SIMCA

Modelos SIMCA foram construídos individualmente para cada classe utilizando a série de teste como técnica de validação. Quatro níveis de significância para o teste- F foram avaliados: 1%, 5%, 10% e 25% e o conjunto de amostras de teste foi utilizado para comparação com as demais estratégias. A **Tabela 4.3** apresenta os resultados de classificação SIMCA para os quatro níveis de significância do teste- F .

Tabela 4.3. Número de erros de classificação dos modelos SIMCA para o conjunto de teste de óleo diesel em diferentes níveis de significância do teste- F (1%, 5%, 10% e 25%).

Modelo	baixo teor de enxofre (4 PCs)				alto teor de enxofre (8 PCs)			
	1	5	10	25	1	5	10	25
baixo teor de enxofre	-	-	1	4	28	25	23	15
alto teor de enxofre	23	23	20	5	-	-	1	4

A **Tabela 4.3** mostra ausência de erros do Tipo I para os níveis de significância de 1% e 5%. Em contrapartida, um número elevado de erros do Tipo II é apresentado para os dois tipos de amostras. Estes resultados estão de acordo com os obtidos pela PCA em todo o conjunto de dados. De fato, as duas classes estão se sobrepondo e, conseqüentemente, várias amostras acabam sendo classificadas em uma classe errada.

4.5.4. SPA-LDA

O gráfico de *scree*, apresentado na **Figura 4.6**, exhibe o mínimo em apenas duas variáveis, correspondentes aos comprimentos de onda 1200 nm e 1466 nm.

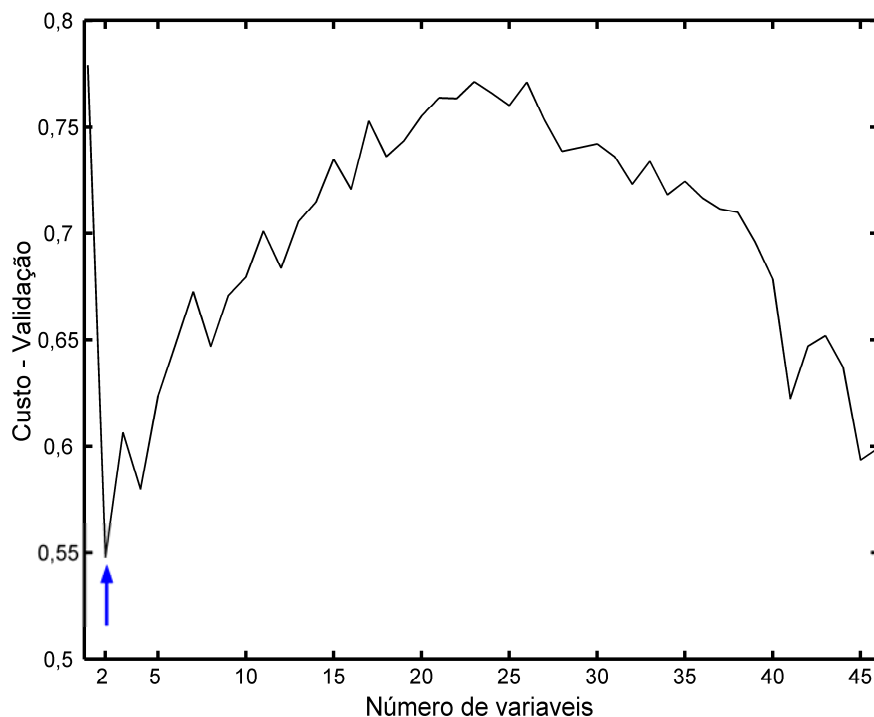


Figura 4.6. Custo da validação em função do número de variáveis selecionadas pelo SPA-LDA para o conjunto de dados de óleos diesel. A seta indica o ponto mínimo da curva do custo (0.5478), no qual ocorre em dois comprimentos de onda.

A **Figura 4.7** mostra a localização das duas variáveis selecionadas pelo SPA ao longo dos espectros médios (original e derivativo).

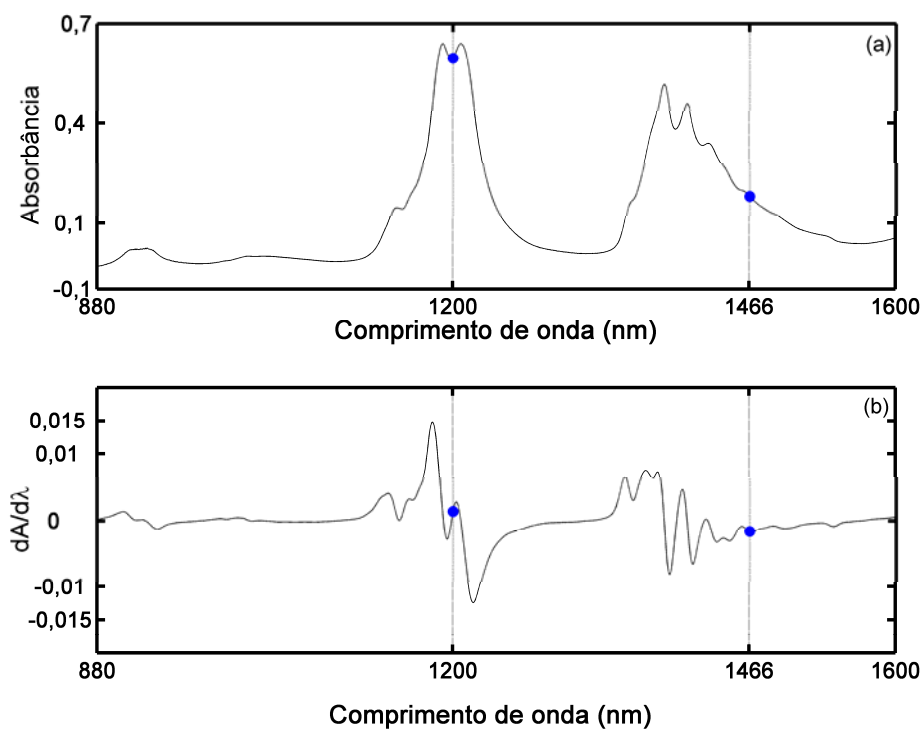


Figura 4.7. Espectro médio original (a) e derivativo (b) de óleo diesel com indicação dos comprimentos de onda selecionados pelo SPA.

Capítulo IV. Classificação de óleos diesel

Os modelos LDA resultantes construídos com essas duas variáveis foram aplicados ao conjunto de teste. Como resultado, onze amostras foram incorretamente classificadas (duas para a classe de baixo teor de enxofre e 9 para classe de alto teor de enxofre), totalizando um índice de acerto para o conjunto de Teste de 78,8%.

4.5.5. GA-LDA

Para efeito de comparação, modelos LDA foram construídos também com a seleção de variáveis pelo GA. O melhor resultado foi obtido com 25 comprimentos de onda, os quais são apresentados na **Figura 4.8**.

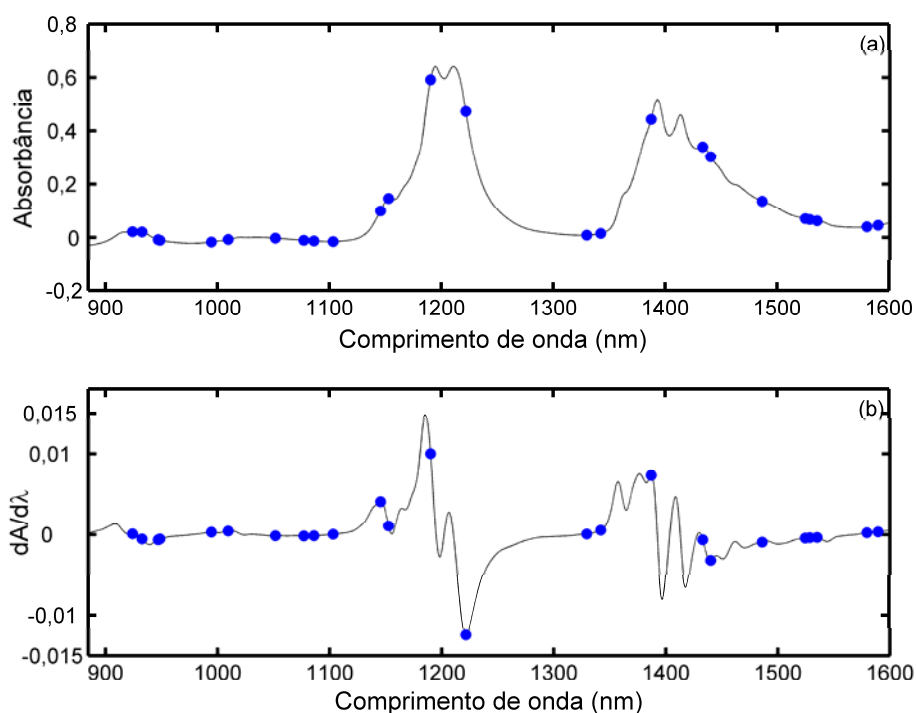


Figura 4.8. Espectro médio original (a) e derivativo (b) de óleo diesel com indicação dos comprimentos de onda selecionados pelo GA.

O modelo LDA resultante construído com as 25 variáveis selecionadas pelo GA, quando aplicado ao conjunto de Teste, apresentou um desempenho similar ao do SPA-LDA (onze erros de classificação). A **Tabela 4.4** apresenta o resumo dos erros de classificação para as três estratégias de modelagem empregadas nesse estudo.

Capítulo IV. Classificação de óleos diesel

Tabela 4.4. Resumo dos resultados (erros de classificação para o conjunto de teste) para o SPA-LDA, GA-LDA e SIMCA (4 níveis de significância do teste-*F*) aplicados ao conjunto de dados de óleos diesel.

	SPA-LDA	GA-LDA	SIMCA 1%	SIMCA 5%	SIMCA 10%	SIMCA 25%
Tipo I	11	11	-	-	2	8
Tipo II	11	11	51	48	43	20
Total	22	22	51	48	45	28

Como pode ser visto na **Tabela 4.4**, o desempenho dos modelos LDA com a seleção de variáveis pelo SPA ou GA é superior aos modelos SIMCA em todos os níveis de significância do teste-*F*. Os modelos SIMCA menos adequados foram aqueles avaliados para os níveis de significância de 1% e 5%, resultando em 51 e 48 erros, respectivamente. Os melhores resultados são alcançados com 25% (28 erros).

4.5.6. Análise de sensibilidade ao ruído

De forma semelhante ao capítulo III, os espectros do conjunto de teste foram contaminados com ruído. Nesse caso, o desvio padrão do ruído adicional foi da ordem de 10^{-5} . Novamente, os modelos previamente obtidos foram aplicados para a classificação do novo conjunto de teste de amostras de óleos diesel.

A **Tabela 4.5** apresenta uma comparação entre os resultados de classificação dos modelos SPA-LDA, GA-LDA e SIMCA para o novo cenário. Os números entre parênteses indicam os resultados obtidos anteriormente (sem a adição do ruído). O nível de significância do teste-*F* empregado para o SIMCA foi aquele que apresentou o melhor desempenho preditivo no estudo inicial (25%).

Tabela 4.5. Número total de erros (Tipo I e Tipo II) obtidos pelos modelos SPA-LDA, GA-LDA e SIMCA no conjunto de teste de óleos diesel contaminado pelo ruído.

	SPA-LDA	GA-LDA	SIMCA
Amostras	22 (22)	42 (22)	26 (28)

O desempenho do modelo GA-LDA aplicado ao conjunto de teste foi afetado pela adição do ruído, passando de 22 para 42 erros. Mais uma vez, a eficiência do SPA quanto à presença do ruído foi comprovada, mantendo-se constante o número de erros para o novo cenário. De fato, as variáveis selecionadas pelo GA (**Figura 4.8**), assim como aquelas selecionadas para o estudo de classificação de óleos vegetais (**Figura 3.10**), apresentaram uma tendência em cobrir toda a faixa espectral. Conseqüentemente, algumas regiões com pouca informação e/ou baixa

Capítulo IV. Classificação de óleos diesel

relação sinal/ruído foram utilizadas pelo modelo GA-LDA, resultando em um acréscimo no número de erros.

Para o SIMCA, ocorreu um decréscimo do número de erros após a inserção do ruído. Esse resultado aparentemente contraditório pode ser atribuído ao decréscimo do número de erros do Tipo II (de 20 para 13 erros), que não foi compensado com o aumento de erros do Tipo I (de 8 para 13 erros).

4.6. Considerações Finais

Neste capítulo, foi demonstrado o potencial do SPA-LDA para seleção de variáveis espectrais em mais um problema de classificação. Especificamente, a espectrometria de absorção molecular na região do NIR foi empregada em um estudo de classificação de amostras de óleo diesel com respeito ao teor de enxofre (baixo teor e alto teor).

O SPA-LDA mostrou um desempenho melhor que o SIMCA para todos os níveis de significância avaliados.

Em uma comparação com o GA-LDA, o SPA-LDA apresentou um desempenho similar para o conjunto de teste (22 erros). Contudo, o modelo LDA construído com a seleção de 25 comprimentos de onda pelo GA mostrou ser menos parcimonioso que o SPA-LDA, que utilizou apenas duas variáveis.

A superioridade do SPA-LDA frente às demais estratégias foi consolidada com o estudo de sensibilidade ao ruído. Assim como no capítulo III, a solução encontrada pelo GA foi desfavorecida com a seleção de algumas variáveis pouco informativas e/ou com baixa relação sinal/ruído. Conseqüentemente, o desempenho do modelo GA-LDA quando aplicado a um novo conjunto de teste contaminado com ruído, foi inferior ao SPA-LDA.

CAPÍTULO V
CLASSIFICAÇÃO DE CAFÉS

5. CLASSIFICAÇÃO DE CAFÉS

5.1. Introdução

5.1.1. Cafés

O café, tradicionalmente produzido a partir dos grãos torrados do fruto do cafeeiro, é uma das bebidas mais consumidas em todo o mundo. No Brasil, o café surgiu no estado do Pará por volta de 1727, mas devido às condições climáticas do país, o cultivo foi largamente difundido por outros estados, como Maranhão, Bahia, Rio de Janeiro, São Paulo, Paraná, Minas Gerais, entre outros. Inicialmente, a produção era voltada para o mercado interno, mas pouco tempo depois, o Brasil passou a ser um dos maiores produtores mundiais^[104].

A qualidade do café depende, entre outros fatores, do método de colheita, processamento, armazenamento e, sobretudo, da composição química dos grãos. Tanto a presença quanto a concentração de alguns compostos no café contribuem significativamente para o aroma e sabor característicos da bebida^[104]. Entre os compostos presentes no café, destacam-se a cafeína, trigonelina e os ácidos clorogênicos, apresentados na **Figura 5.1**.

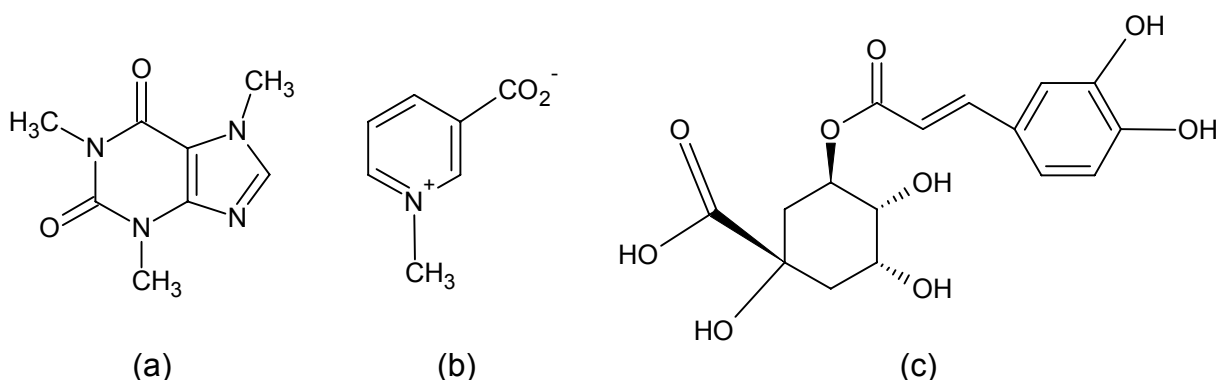


Figura 5.1. Estruturas das principais moléculas presentes no café. (a) cafeína; (b) trigonelina e (c) ácido clorogênico.

A cafeína (1,3,7-trimetilxantina), alcalóide principal encontrado no café, tem sido objeto de estudo de muitos pesquisadores, especialmente no que diz respeito a seus efeitos sobre a saúde humana. De fato, a literatura mostra que doses elevadas de cafeína podem provocar vários efeitos tóxicos, como arritmia, hipopotassemia, hiperglicemia, vômitos, convulsões, entre outros^[105-107]. Adicionalmente, tem-se registrado que a cafeína, consumida em pequenas quantidades, pode também provocar efeitos indesejáveis em pessoas que apresentam alta sensibilidade a esse

Capítulo V. Classificação de cafés

alcalóide. Tais efeitos incluem: redução da qualidade do sono^[108], aumento da pressão sanguínea^[109] e até problemas no desenvolvimento de fetos em mulheres grávidas^[110].

Tendo em vista os potenciais malefícios acima listados, cafés descafeinados podem ser uma opção conveniente, principalmente para consumidores que apresentem alta sensibilidade à cafeína. Então, uma avaliação da conformidade do café com respeito ao tipo (cafeinado e descafeinado) acaba sendo um assunto de grande relevância.

Um outro aspecto que merece destaque no que diz respeito à qualidade do café é o seu estado de conservação. Nesse contexto, o prazo de validade, definido como o tempo após o qual o produto se torna inaceitável para o consumo em uma determinada condição de estocagem, é normalmente indicado nos rótulos dos produtos. Entretanto, elevações na temperatura e na pressão parcial de oxigênio podem acelerar a degradação do café, causando uma redução substancial de sua qualidade. Assim, um procedimento de inspeção do estado de conservação torna-se necessário, não apenas para contornar os danos causados à saúde dos consumidores, mas também para desencorajar o comércio de produtos falsificados^[111].

O controle de qualidade, classificação ou autenticação de amostras de cafés têm sido realizados empregando técnicas instrumentais, tais como: HPLC^[112], GC-MS^[113], ICP OES^[114], RMN^[115] e espectrometria FTIR^[116]. Contudo, a grande maioria dessas técnicas necessita de reagentes perigosos e/ou equipamentos caros com elevado custo de operação e manutenção. Nesse contexto, a espectrometria UV-VIS poderá ser uma alternativa vantajosa para análises mais simples e de menor custo. Esta técnica tem sido empregada com sucesso para a determinação de cafeína^[117, 118], ácidos clorogênicos^[118] e teobromina^[117] em amostras de cafés.

5.2. Objetivos

Avaliar o SPA-LDA com a espectrometria UV-VIS para a classificação de cafés com respeito ao tipo (cafeinado/descafeinado) e prazo de validade (vencidos e não vencidos);

Capítulo V. Classificação de cafés

Comparar o desempenho dos modelos SPA-LDA com o SIMCA (em diferentes níveis de significância para o Teste- F : 1%, 5%, 10% e 25%) em função do número de erros para o conjunto de teste;

Avaliar a robustez dos modelos SPA-LDA e SIMCA com respeito à adição de ruído artificial aos espectros pertencentes ao conjunto de teste.

5.3. Experimental

5.3.1. Amostras

Cento e setenta e cinco amostras de grãos torrados e moídos de cafés descafeinados e cafeinados, de diferentes lotes e fabricantes, foram adquiridas de estabelecimentos comerciais e de indústrias de torrefação da cidade de João Pessoa, Paraíba. Essas amostras foram de misturas processadas a seco das variedades Robusta e Arábica, espécies mais empregadas pelas indústrias brasileiras. Em grãos verdes dessas variedades, o teor de cafeína pode variar entre 1 – 2% m/m, sendo mais alta na espécie robusta do que na arábica. É importante destacar que durante o processo de torrefação, este teor não é alterado significativamente^[119]. Segundo a ANVISA^[120], os cafés descafeinados comercializados no Brasil não devem apresentar teor de cafeína superior a 0,1% m/m.

Entre as cento e setenta e cinco amostras analisadas, noventa delas haviam sido estocadas em recipientes comerciais sem um rigoroso controle das condições ambientais por um período de 30 a 50 meses após do prazo de validade. Este elevado tempo de estocagem foi aqui adotado com intuito de garantir que tais amostras estivessem realmente inapropriadas para o consumo. A partir de agora, essas noventa amostras serão chamadas de “vencidas”. Para esse problema de classificação, foram consideradas quatro classes: descafeinado não vencido, cafeinado não vencido, descafeinado vencido e cafeinado vencido. A **Tabela 5.1** mostra o número de amostras em cada classe de cafés empregados nesse estudo.

Tabela 5.1. Número e tipo de amostras de cafés analisadas.

	Descafeinado não vencido	Cafeinado não vencido	Descafeinado vencido	Cafeinado vencido
Número de amostras	31	54	22	68

Capítulo V. Classificação de cafés

5.3.2. Equipamentos

O equipamento utilizado para o registro dos espectros UV-VIS das amostras de cafés (**Figura 3.1**) foi o mesmo empregado para o estudo de classificação de óleos vegetais (Capítulo III). Contudo, uma cubeta de quartzo de 10 mm de caminho óptico foi utilizada e não foi preciso elaborar um sistema em fluxo. Conseqüentemente, alguns componentes como bomba peristáltica, tubos de PVC e cela de fluxo foram dispensados.

5.3.3. Procedimento Analítico

O seguinte procedimento de extração aquosa, adaptado de Vitorino *et al*^[121], foi empregado nesse estudo:

- 1,0 g de cada mostra de café foi pesado e transferido para um papel de filtro posto sobre um funil de vidro;
- Três alíquotas de 50 mL de água destilada com temperatura entre 90 – 98 °C foram sequencialmente adicionadas sobre a amostra.
- Depois do resfriamento para a temperatura ambiente (cerca de 25 °C), os extratos foram diluídos na proporção de 1:20 (v/v) com água destilada.

Após o procedimento descrito acima, o espectro de cada extrato diluído foi imediatamente registrado na região de 225 – 353 nm, com 1 nm de resolução. O espectro do branco foi obtido apenas com água destilada.

5.3.4. Tratamento dos dados e softwares

O algoritmo KS foi novamente utilizado para dividir o conjunto de amostras em treinamento, validação e teste. O procedimento para dividir tais conjuntos foi realizado separadamente para cada classe, como no estudo de classificação de óleos vegetais e óleos diesel. O número de amostras para cada conjunto das quatro classes estudadas é apresentado na **Tabela 5.2**.

Tabela 5.2. Número de amostras de treinamento, validação e teste selecionadas pelo KS para as quatro classes de cafés.

Classes	Conjuntos		
	Treinamento	Validação	Teste
Descafeinado não vencido	17	7	7
Cafeinado não vencido	26	14	14
Descafeinado vencido	12	5	5
Cafeinado vencido	34	17	17
Total	89	43	43

Capítulo V. Classificação de cafés

As amostras de treinamento e validação foram usadas no procedimento de modelagem (incluindo seleção de variáveis pelo SPA para LDA, bem como determinação do número ótimo de PCs para o SIMCA). O terceiro conjunto (teste) foi, assim como nos outros estudos, usado para a avaliação final e comparação do desempenho dos modelos de classificação.

PCA e SIMCA foram realizados com a versão 9.6 do Unscrambler® (CAMO S.A). SPA-LDA e o KS foram executados em Matlab® 6.5.

5.4. Resultados e Discussão

5.4.1. Espectros das amostras de café

Os espectros de absorção molecular UV-VIS dos extratos aquosos das 175 amostras de cafés, obtidos na faixa de 225 nm a 353 nm, são apresentados na **Figura 5.2**.

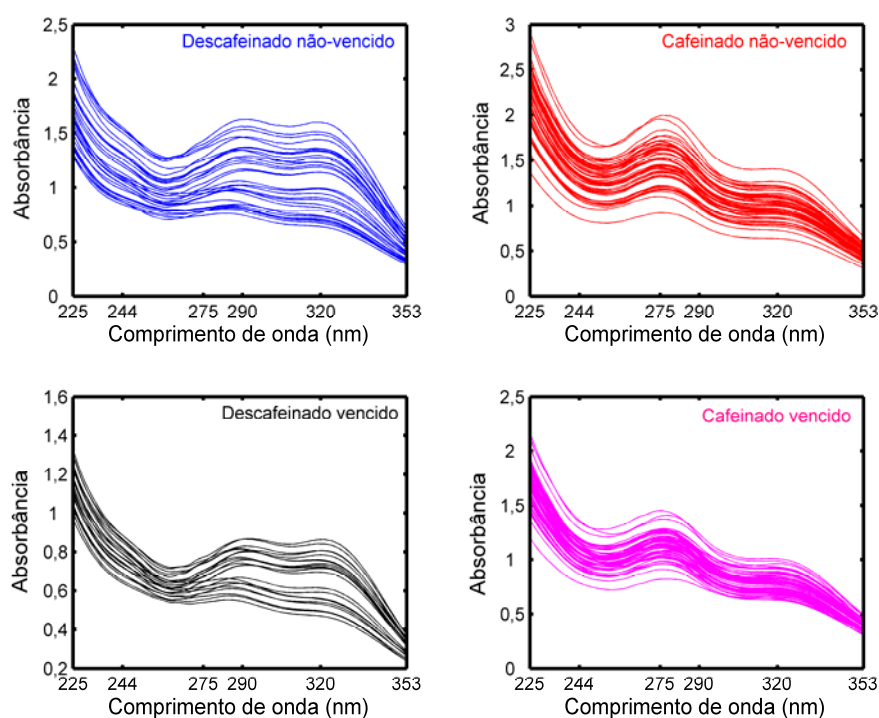


Figura 5.2. Espectros UV-VIS das quatro classes de cafés analisadas.

A região compreendida entre 225-353 nm está associada com as transições eletrônicas $n \rightarrow \pi^*$ das moléculas de cafeína, ácidos clorogênicos e trigonelina. A banda em torno de 275 nm está associada à absorção do cromóforo C=O da cafeína^[119], enquanto que as bandas em torno de 290 nm e 320 nm estão associadas às absorções de ácidos clorogênicos e trigonelina, respectivamente^[121].

Capítulo V. Classificação de cafés

Com intuito de investigar a repetitividade das medidas espectrais, quatro amostras (uma para cada classe) foram empregadas. Dez extrações foram, então, realizadas para cada amostra e três espectros foram registrados para cada extrato. Com estas medidas, o desvio padrão conjunto estimado para cada variável espectral variou na faixa de $5,5 \times 10^{-4}$ a $2,4 \times 10^{-3}$ em absorbância. Tais valores, correspondentes à variabilidade da resposta instrumental para uma determinada amostra, são muito menores que a variabilidade inter-amostras, conforme a **Figura 5.3**.

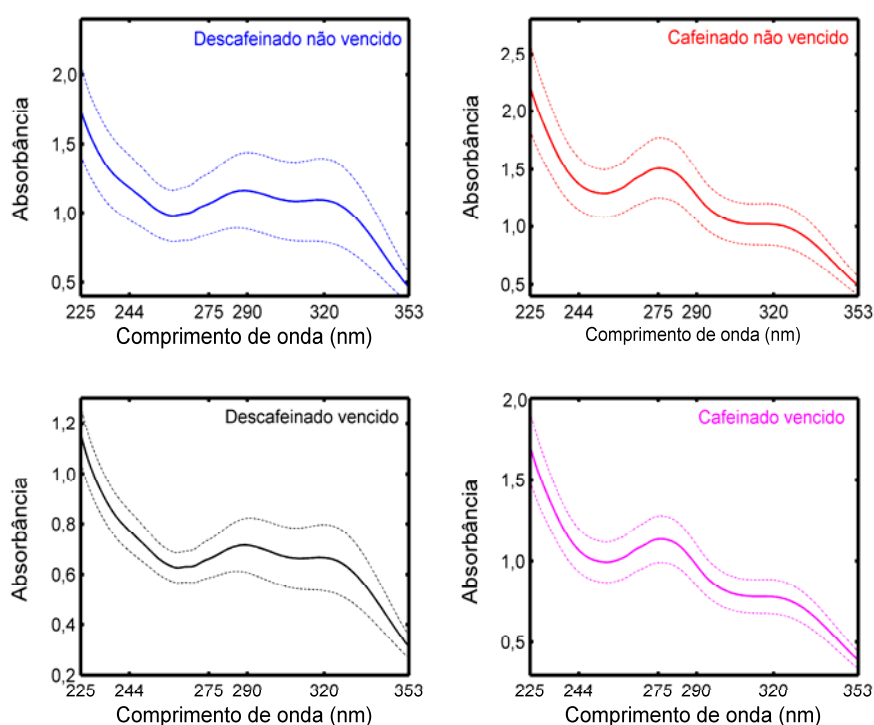


Figura 5.3. Espectros médios das quatro classes de cafés (linhas sólidas) com limites de +/- um desvio padrão (linhas tracejadas).

5.4.2. Análise exploratória dos dados

A **Figura 5.4** apresenta o gráfico dos escores de PC2 \times PC1 resultante da aplicação da PCA aos espectros UV-VIS dos cafés. Para esse caso, uma razoável discriminação entre as classes de cafés cafeinados e descafeinados é alcançada. Contudo, observa-se uma sobreposição entre as classes de amostras vencidas e não vencidas, principalmente em se tratando de cafés cafeinados.

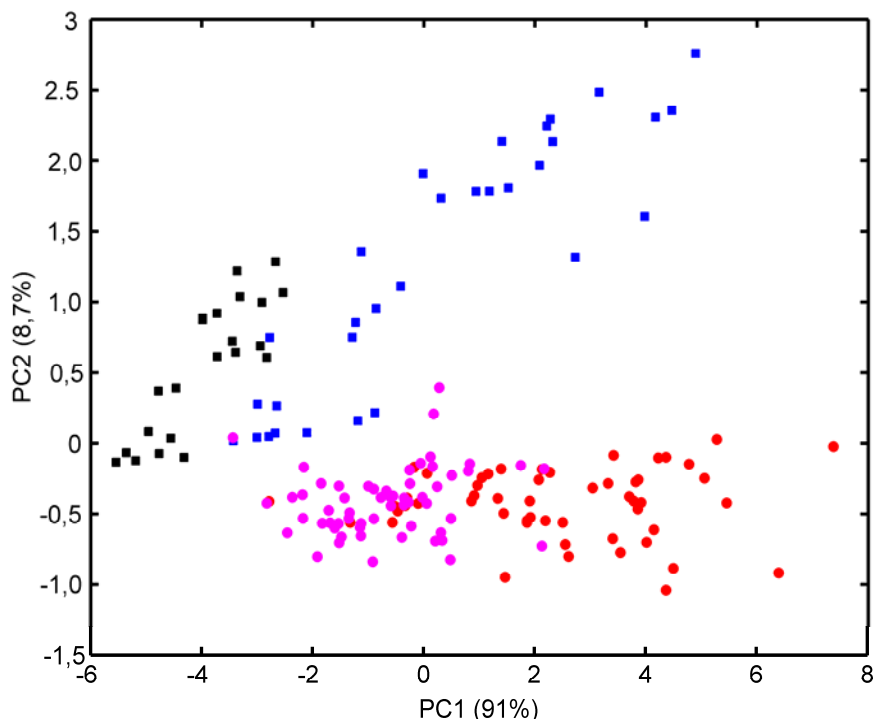


Figura 5.4. Gráfico dos escores de PC2 × PC1 para todas as 175 amostras de café. Descafeinado não vencido: ■; Descafeinado vencido: ■; Cafeinado não vencido: ●; Cafeinado vencido: ●

5.4.3. Classificação SIMCA

Modelos SIMCA foram construídos para cada uma das quatro classes de cafés empregando toda a região do espectro. A **Tabela 5.3** apresenta os resultados de classificação obtidos para o conjunto de teste em quatro níveis de significância do Teste-F (1%, 5%, 10% e 25%).

Tabela 5.3. Número de erros de classificação obtido pelos modelos SIMCA usando toda a faixa espectral para o conjunto de amostras de teste de cafés em quatro níveis de significância do Teste-F (1%, 5%, 10% e 25%). O número de PCs é indicado entre parênteses.

Nível (%)	Modelo															
	Desc. n/ vencido (2PCs)				Caf. n/ vencido (2PCs)				Desc. vencido (2PCs)				Caf. vencido (2PCs)			
	1	5	10	25	1	5	10	25	1	5	10	25	1	5	10	25
Desc. n/ vencido	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-
Caf. n/ vencido	-	-	-	-	-	-	1	6	-	-	-	-	9	8	6	4
Desc. vencido	3	1	-	-	-	-	-	-	-	1	1	1	-	-	-	-
Caf. vencido	-	-	-	-	17	17	8	4	-	-	-	-	-	-	-	-

Com base nos resultados apresentados na **Tabela 5.3**, pode-se concluir que os modelos SIMCA não foram adequados no que diz respeito ao estado de conservação das amostras de cafés cafeinados. Esses resultados condizem com os gráficos de escores da PCA (**Figura 5.4**), uma vez que uma elevada sobreposição entre as classes de cafés cafeinados não vencidos e vencidos foi observada. Entre

Capítulo V. Classificação de cafés

os quatro níveis de significância avaliados, 10% e 25% foram aqueles que apresentaram o melhor desempenho de classificação (um total de 16 erros).

5.4.4. SPA-LDA

O gráfico de *scree* apresentado na **Figura 5.5** mostra que o número ótimo de variáveis selecionadas pelo SPA-LDA foi 15, que correspondem aos comprimentos de onda: 225, 226, 227, 229, 231, 235, 244, 259, 271, 274, 280, 293, 324, 339 e 353 nm, indicados na **Figura 5.6**. O modelo LDA obtido com essas variáveis foi, então, aplicado ao conjunto de Teste. Como resultado, todas as amostras foram corretamente classificadas (índice de acerto de 100%).

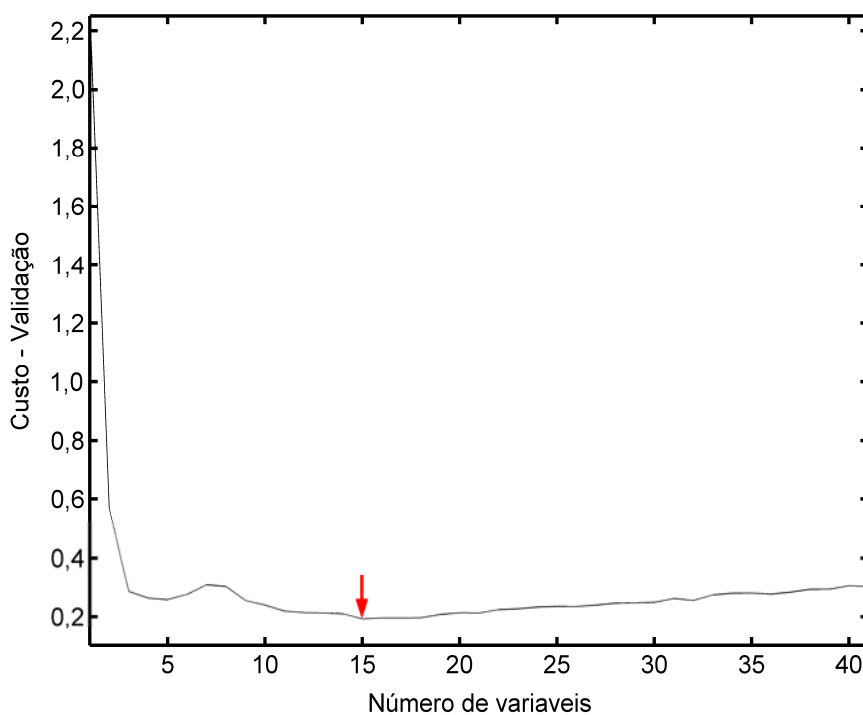


Figura 5.5. Gráfico de *scree* obtido pelo SPA-LDA para os espectros UV-VIS de cafés.

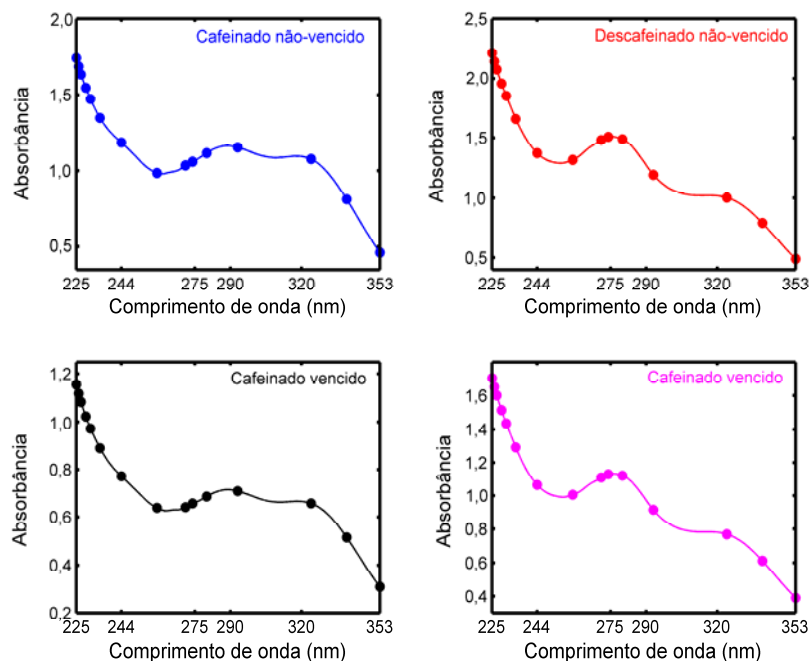


Figura 5.6. Espectros médios das quatro classes de cafés estudadas com os 15 comprimentos de onda selecionados pelo SPA-LDA.

5.4.5. PCA e SIMCA com as variáveis selecionadas pelo SPA-LDA

Com intuito de avaliar o poder discriminatório das variáveis selecionadas, uma PCA (**Figura 5.7**) para as 175 amostras e modelos SIMCA individuais para cada classe foram refeitos usando apenas aquelas 15 variáveis selecionadas pelo SPA-LDA.

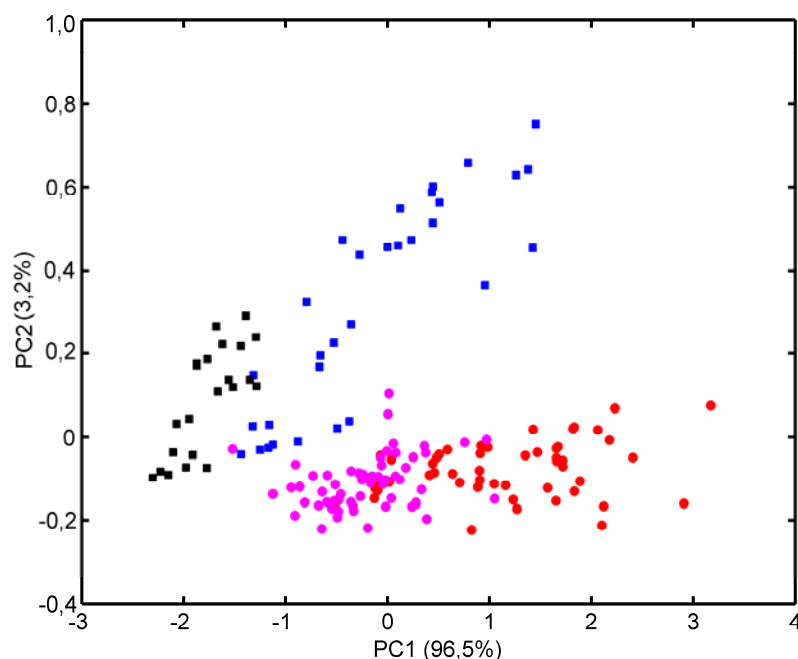


Figura 5.7. Gráfico dos escores de PC2 \times PC1 resultante da PCA realizada em todas as 175 amostras com as 15 variáveis selecionadas pelo SPA-LDA. Descafeinado não vencido: ■; Descafeinado vencido: ■; Cafeinado não vencido: ●; Cafeinado vencido: ●.

Capítulo V. Classificação de cafés

O perfil do gráfico dos escores apresentado na **Figura 5.7** é muito parecido com aquele obtido pela PCA empregando o espectro completo (**Figura 5.4**), ou seja, pouca discriminação entre as amostras vencidas das não vencidas é apresentada, principalmente para a classe dos cafeinados.

Os resultados de classificação SIMCA com os comprimentos de onda selecionados pelo SPA-LDA são apresentados na **Tabela 5.4**.

Tabela 5.4. Número de erros de classificação obtido pelos modelos SIMCA (em quatro níveis de significância do Teste-*F*: 1%, 5%, 10% e 25%) usando as 15 variáveis selecionadas pelo SPA para o conjunto de amostras de teste de café. O número de PCs é indicado entre parênteses

Nível (%)	Modelo															
	Desc. n/ vencido (2PCs)				Caf. n/ vencido (2PCs)				Desc. vencido (2PCs)				Caf. vencido (2PCs)			
	1	5	10	25	1	5	10	25	1	5	10	25	1	5	10	25
Desc. n/ vencido	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-
Caf. n/ vencido	-	-	-	-	-	1	1	5	-	-	-	-	8	7	5	4
Desc. vencido	3	1	-	-	-	-	-	-	-	1	-	-	-	-	-	-
Caf. vencido	-	-	-	-	16	16	16	7	-	-	-	-	-	-	-	-

Com essa nova abordagem, o modelo SIMCA para o nível de significância de 25% foi o que apresentou o melhor resultado (17 erros). De forma geral, os resultados de classificação SIMCA (para os quatro níveis de significância) apresentaram um desempenho semelhante ao SIMCA construído em toda faixa espectral. Para os níveis de significância 1% e 5%, houve um pequeno decréscimo no número total de erros para os modelos construídos com as variáveis selecionadas pelo SPA-LDA. Em contrapartida, para os níveis 10% e 25%, um aumento dos erros foi apresentado. Portanto, esses resultados confirmam que a informação discriminatória apresentada no espectro completo foi preservada nas variáveis selecionadas pelo SPA-LDA. A **Tabela 5.5** resume os resultados de classificação obtidos pelos modelos SIMCA e SPA-LDA.

Tabela 5.5. Resumo dos resultados (erros de classificação para o conjunto de teste) para o SPA-LDA e SIMCA (4 níveis de significância do teste-*F*) aplicados ao conjunto de dados café. Os valores entre parênteses indicam o número de erros obtidos pelos modelos SIMCA construídos com as variáveis selecionadas pelo SPA-LDA.

	SPA-LDA	SIMCA 1%	SIMCA 5%	SIMCA 10%	SIMCA 25%
Tipo I	0	0 (0)	1 (2)	2 (1)	8 (6)
Tipo II	0	29 (27)	26 (24)	14 (21)	8 (11)
Total	0	29 (27)	27 (26)	16 (22)	16 (17)

Fica claro, com base no número de erros apresentados pelas três estratégias, que o desempenho do SPA-LDA foi superior ao SIMCA, tanto com

modelos construídos com o espectro completo, como para aqueles que empregaram as variáveis selecionadas pelo SPA-LDA.

Os valores de escores das duas primeiras funções discriminantes (FD2 × FD1) obtidos pela LDA para o conjunto de dados completo (**Figura 5.8**) confirmam também o bom desempenho do SPA-LDA. A separação das quatro classes de cafés na **Figura 5.8** é muito mais evidente do que nos gráficos dos escores obtidos por PC2 × PC1 (**Figura 5.4 e Figura 5.7**).

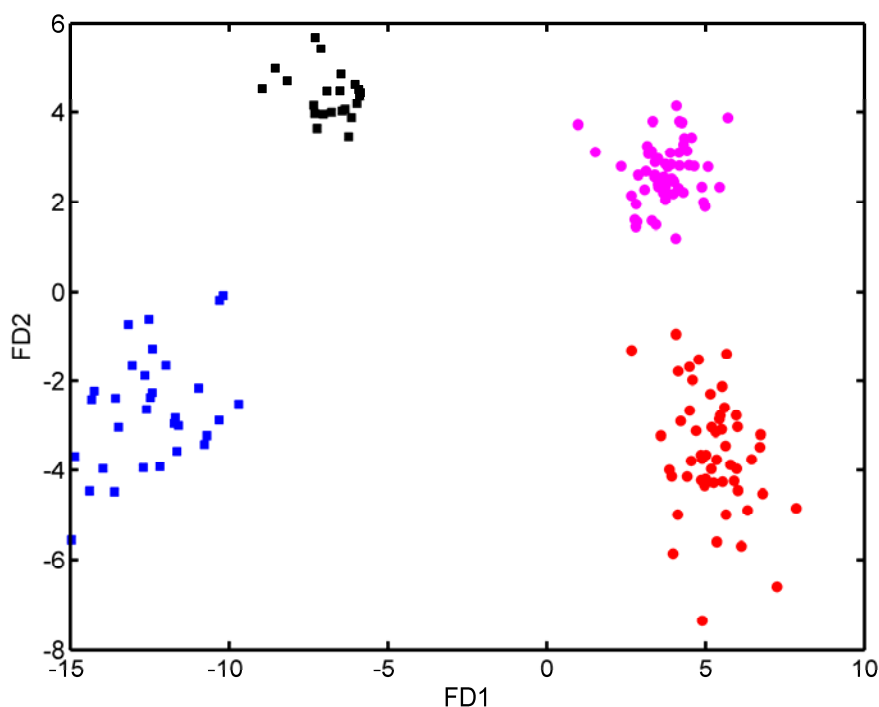


Figura 5.8. Escores de FD2 × FD1 obtidos pela LDA com as variáveis selecionadas pelo SPA. Descafeinado não vencido: ■; Descafeinado vencido: ■; Cafeinado não vencido: ●; Cafeinado vencido: ●.

Como pode ser visto na **Figura 5.8**, a FD1 está sendo responsável por indicar a separação das amostras de cafés em dois grandes grupos: cafeinados e descafeinados com valores de escores positivos ou negativos. Já a FD2 discrimina quanto ao estado de conservação (vencidos e não vencidos), ou seja, com valores de escores positivos, encontram-se amostras com prazo de validade expirado, enquanto que negativos, amostras dentro do prazo de validade.

5.4.6. Robustez dos modelos

Com intuito de avaliar a robustez dos modelos SPA-LDA e SIMCA com respeito ao ruído instrumental, uma simulação de Monte Carlo^[122] foi realizada. Para

este propósito, ruídos artificiais foram adicionados aos espectros do conjunto de teste. O nível do ruído adicionado foi de acordo com o desvio-padrão conjunto calculado no estudo de repetitividade das medidas espectrais (apresentado na seção 5.4.1). Para o espectro de cada uma das 43 amostras do conjunto de teste, dez realizações de ruídos diferentes foram adicionadas, resultando em 430 espectros ruidosos. Então, os modelos SPA-LDA e SIMCA previamente elaborados foram novamente avaliados quanto ao seu desempenho de classificação para o novo conjunto de teste contaminado pelo ruído. Com o SPA-LDA, foi possível alcançar um índice de acerto de 96% (de 430 amostras, 19 foram incorretamente classificadas). Com o SIMCA (com 25% de significância) construído com espectro completo, 80 erros do Tipo I e 120 erros do Tipo II foram obtidos. Finalmente, para o SIMCA construído com os 15 comprimentos de onda selecionados pelo SPA-LDA, o número de erros do Tipo I e Tipo II foi 80 e 124, respectivamente.

5.5. Considerações finais

Neste capítulo, foi demonstrado mais uma vez o bom desempenho do SPA-LDA em um novo problema de classificação. Especificamente, a espectrometria UV-VIS foi adotada como técnica analítica para o desenvolvimento de uma nova metodologia analítica capaz de classificar amostras de cafés com respeito ao tipo (cafeinado e descafeinado) e ao estado de conservação (vencidos e não vencidos).

A partir da realização de uma PCA em todo o conjunto de dados, foi possível observar uma elevada sobreposição entre as amostras de cafés pertencentes à classe vencidas e não vencidas, principalmente para o grupo dos cafeinados.

Modelos SIMCA construídos com base em toda a região espectral foram pouco eficientes na classificação de amostras com respeito ao estado de conservação. Em particular, um número elevado de erros do Tipo II foi encontrado para a classe dos cafeinados vencidos e não vencidos.

O SPA-LDA, quando aplicado aos espectros UV-VIS, selecionou quinze variáveis (comprimentos de onda) e classificou corretamente todas as amostras do conjunto de teste (índice de acerto de 100%).

Modelos SIMCA foram reconstruídos utilizando as mesmas quinze variáveis selecionadas pelo SPA-LDA. Os resultados de classificação foram semelhantes àqueles obtidos com toda região espectral. Isto indica que a informação

Capítulo V. Classificação de cafés

discriminatória apresentada no espectro completo foi preservada nas variáveis selecionadas pelo SPA-LDA.

Um estudo de robustez dos modelos SPA-LDA e SIMCA foi realizado contaminando as amostras do conjunto de Teste com ruído artificial. Os modelos anteriormente elaborados foram utilizados para a classificação desse novo conjunto contaminado com ruído. O desempenho do SPA-LDA (96% de acerto) foi superior ao SIMCA (toda região espectral) e ao SIMCA construído com as variáveis selecionadas pelo SPA-LDA.

CAPÍTULO VI
CLASSIFICAÇÃO DE SOLOS BRASILEIROS

6. Classificação de solos brasileiros

6.1. Introdução

6.1.1. Solos brasileiros

O solo pode ser definido como uma coleção de corpos naturais, constituídos por partes sólidas, líquidas e gasosas. São tridimensionais e formados por minerais e compostos orgânicos que ocupam a maior parte do manto superficial das extensões continentais do planeta^[123].

No Brasil, a classificação de solos é um assunto de grande interesse cuja motivação é essencialmente direcionada à necessidade de levantamentos e estudos pedológicos, bem como para a determinação da utilização e manejo dos solos.

A classificação de solos atualmente vigente no País é baseada em conceitos e na evolução do antigo sistema americano, proposto inicialmente em 1938 por Baldwin *et al.*^[124] e modificado posteriormente por Thorp e Smith em 1949^[125]. Porém, os critérios adotados para tal classificação vêm mudando constantemente de acordo com o avanço da atual ciência dos solos. Essas mudanças ocorrem desde a alteração de alguns conceitos pré-estabelecidos como na criação, desmembramento e formação de novas classes e subclasses.

A classificação de solos no território nacional é estruturada em seis diferentes níveis categóricos (Ordem, Subordem, Grande Grupo, Subgrupo, Família e Série). O nível mais genérico é a Ordem, no qual é possível fazer uma distinção de verdadeiras províncias de solos. O nível mais detalhado e preciso é a Série, que separa unidades bastante homogêneas. Entre a Ordem e a Série, ocorre uma diminuição do grau de generalização e aumento do grau de especificação^[123].

Recentemente, a última avaliação realizada pelo Sistema Brasileiro de Classificação de Solos (SIBCS) definiu 13 ordens para os solos: Argissolo, Latossolo, Nitossolo, Neossolo, Vertissolo, Cambissolo, Chernossolo, Luvisolo, Espodossolo, Planossolo, Plintossolo, Gleissolo e Organossolo^[123]. As três primeiras foram utilizadas nesse problema de classificação e, segundo o Instituto Agrônomo de Campinas (IAC), as ordens Argissolo e Latossolo são as mais abundantes na América do Sul e os nitossolos ocupam aproximadamente 1% do território brasileiro.

• Argissolos

Argissolos são solos minerais, não-hidromórficos, de textura média a arenosa, possuem argila de atividade baixa ou alta e apresentam uma clara diferenciação entre os horizontes (designação dada a várias camadas do solo) A, B e C. Quando bem manejados, são relativamente férteis e indicados às atividades agropastoris. Apresentam horizonte B de cor avermelhada até amarelada e teores de óxidos de ferro inferiores a 15%. Possuem profundidades variadas e uma variabilidade ampla de classes texturais^[123].

• Latossolo

Entre as 13 ordens de solos pré-estabelecidas no território brasileiro, os latossolos são os mais abundantes. Esses solos são muito intemperizados, com pouca reserva de nutrientes para as plantas. Cerca de 95% dos latossolos são ácidos (pH entre 4,0 e 5,5) e, geralmente, apresentam problemas quanto à fertilidade. Apresentam pouca distinção entre os horizontes A, B e C e possuem cores que variam de vermelhas muito escuras a amareladas (normalmente escuras no horizonte A, vivas no B e mais claras no C). Além disso, apresentam alta permeabilidade à água, o que os torna possível de serem trabalhados em grandes amplitudes de umidade^[123].

• Nitossolo

São solos minerais, não-hidromórficos, com cor vermelho-escura a arroxeada, eutróficos (grande maioria) e comumente derivados do intemperismo de rochas básicas. A textura dessa ordem de solo pode variar de argilosa a muito argilosa. Além disso, são solos bastante porosos (porosidade normalmente superior a 50%), com teores de ferro (Fe_2O_3) muito elevados (podendo chegar a mais de 15%) e tem como principal limitação, o grande risco de erosão^[123].

6.1.2. Classificação de solos

A classificação de solos brasileiros é realizada com o uso de parâmetros físicos, químicos e morfológicos. Contudo, os métodos de referência para a determinação desses parâmetros são laboriosos e consomem muito tempo, principalmente porque necessitam de procedimentos de tratamento das amostras. Além disso, alguns critérios de classificação são extremamente subjetivos e difíceis de serem quantificados.

Capítulo VI. Classificação de solos brasileiros

Alguns artigos têm sido publicados com o uso de parâmetros como fertilidade^[126] e características morfológicas^[127] do solo para a finalidade de classificação. Entretanto, poucos trabalhos têm abordado o desenvolvimento de técnicas analíticas e/ou procedimentos de tratamento de dados para simplificar o uso de sistemas de classificação já existentes^[128-130].

Zagatto^[128] determinou 20 elementos presentes em extratos de amostras de solos utilizando a AAS ou ICP OES. Os valores obtidos por essas determinações foram utilizados, junto com as técnicas KNN e SIMCA, para propósito de classificação de alguns tipos de solos brasileiros. Nessa época a Quimiometria era ainda muito rudimentar, mas mesmo assim, um índice de acerto de 80% foi alcançado.

Demattê *et al.*^[129] utilizaram a espectroscopia de reflectância na região do visível e NIR (450-2500 nm) em uma avaliação de vários tipos de solos. Diferentes profundidades foram utilizadas para determinar as classes de solos pela interpretação descritiva das curvas espectrais e da análise estatística. Os resultados foram favoravelmente comparados com os do método convencional em termos de demarcação e classes dos solos detectadas.

Mouazen *et al.*^[130] também realizaram um estudo de classificação de solos utilizando a espectroscopia de reflectância na região VIS-NIR. Diferentemente do trabalho apresentado por Demattê *et al.*^[129], os autores utilizaram a região compreendida entre 306,5 a 1710,9 nm para classificar texturas de solos. Uma PCA foi inicialmente realizada sobre os espectros e os valores de escores obtidos pelas cinco primeiras PCs foram empregados pela Análise Discriminante Fatorial (FDA: *factorial discriminant analysis*). Quatro diferentes classes de solos foram classificadas com índices de acerto de 85,7% e 81,8% para os conjuntos de treinamento e teste, respectivamente. Em seguida, duas classes semelhantes foram unidas formando um único grupo e o desempenho de classificação aumentou para 89,9% (treinamento) e 85,1% (teste).

6.2. Espectroscopia de Emissão em Plasma Induzido por Laser (LIBS)

A espectroscopia de emissão em plasma induzido por laser (LIBS: *Laser Induced Breakdown Spectroscopy*) é uma técnica analítica moderna, em fase de desenvolvimento e ainda pouco explorada no Brasil. Nela, um laser pulsado é focalizado em uma área restrita da superfície da amostra com irradiância (potência

Capítulo VI. Classificação de solos brasileiros

por unidade de área) da ordem de GW.cm^{-2} ^[131]. A amostra é, então, aquecida de modo a vaporizar uma pequena quantidade das espécies que a constituem, promovendo a formação de um micro-plasma de elevada temperatura. Conseqüentemente, processos como atomização e excitação são desencadeados e, durante a relaxação, átomos ou íons que foram excitados no plasma geram um espectro de emissão característico do material volatilizado, que é usado como resposta analítica^[132-133].

Com o LIBS, é possível determinar quase todos os elementos da tabela periódica em amostras no estado sólido, líquido ou gasoso^[133]. Além disso, a técnica tem como vantagens^[131-133]:

- *Possibilitar análises quase que não destrutivas, ou seja, pouca quantidade de amostra é perdida durante as medidas (cerca de nanograma a micrograma).*
- *Pouco ou nenhum tratamento na amostra é necessário. Assim, a análise se torna mais rápida, mais barata e livre de geração de resíduos.*
- *Possibilidade de realizar medidas a distância. Assim, é promovida uma maior segurança do analista em locais de alta periculosidade e inacessíveis às tradicionais técnicas analíticas.*

Apesar dessas vantagens, a técnica encontra-se em estado de desenvolvimento e, devido às dificuldades apresentadas na calibração, um grande esforço tem sido necessário para o desenvolvimento de metodologias de caráter quantitativo. No contexto das análises qualitativas, o LIBS tem sido empregado com sucesso para a classificação de diferentes amostras, tais como ligas, objetos arqueológicos, polímeros, explosivos, entre outras^[131-132]. Contudo, apenas um artigo já foi publicado explorando as medidas provenientes do LIBS para a classificação de solos^[134].

No trabalho apresentado por Bousquet *et al.*^[134], os autores utilizaram inicialmente espectros LIBS com 50000 variáveis. Estes foram reduzidos para 68 pontos correspondentes às linhas espectrais de oito elementos presentes no solo (alumínio, silício, ferro, cálcio, manganês, potássio, titânio e magnésio). A PCA foi então aplicada aos dados reduzidos e apenas duas classes de amostras foram discriminadas.

De fato, poucas metodologias têm sido desenvolvidas explorando os atrativos do LIBS com a análise multivariada, especificamente para aplicações de

métodos de reconhecimento de padrões e seleção de variáveis. Devido à grande complexidade e à alta dimensionalidade normalmente presentes nos espectros LIBS, espera-se que o uso do SPA-LDA em combinação com procedimentos de compressão de dados (Transformada Wavelet) possa ser uma alternativa vantajosa para a classificação de solos brasileiros.

6.3. Compressão de dados (Transformada Wavelet)

Devido aos espectros LIBS apresentarem um número elevado de variáveis, a aplicação de técnicas de seleção de variáveis em dados como esses pode envolver um elevado tempo computacional. Para contornar este inconveniente, técnicas de compressão baseadas na Transformada Wavelet podem ser empregadas para essa finalidade.

A Transformada Wavelet (WT: *Wavelet Transform*) é uma ferramenta de processamento multiresolucional de sinais que tem sido adotada para remoção de ruído, extração de características e compressão de sinais instrumentais [135-137].

A WT de um espectro $\mathbf{x} = [x(\lambda_1) \ x(\lambda_2) \ \dots \ x(\lambda_j)]$, onde λ_j é o j -ésimo comprimento de onda, pode ser obtida usando uma estrutura de banco de filtros digitais [136,138] apresentada na **Figura 6.1**.

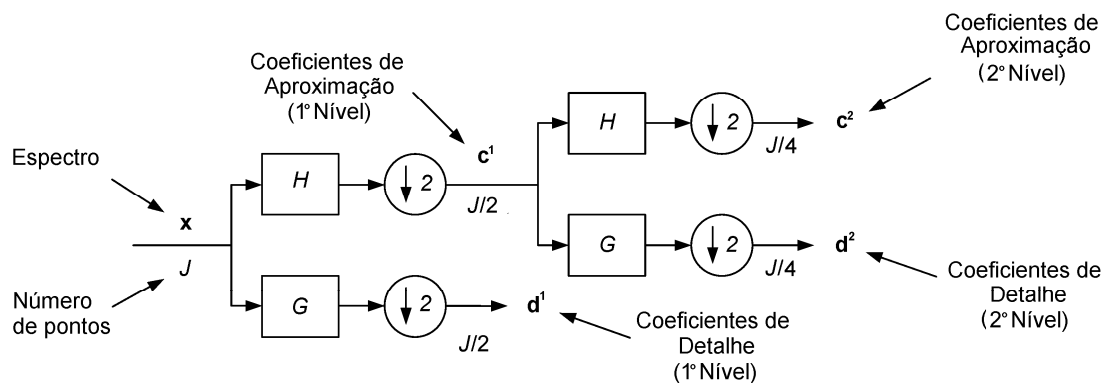


Figura 6.1. Implementação da transformada wavelet empregando um banco de filtros com dois níveis de decomposição. H e G representam os filtros digitais passa-baixas e passa-altas, respectivamente e $\downarrow 2$ denota a operação de sub-amostragem.

A estrutura básica do banco de filtros consiste em um par de filtros passa-baixas (H) e passa-altas (G), que são aplicados a um sinal de entrada \mathbf{x} de comprimento J . Em seguida, uma operação de sub-amostragem ($\downarrow 2$) é realizada, descartando um de cada dois coeficientes. Como saída, têm-se os coeficientes de aproximação (\mathbf{c}), versão suavizada do espectro e os coeficientes de detalhe (\mathbf{d}), tipicamente correspondentes a ruído de alta frequência. Esta operação pode ser

repetida sucessivas vezes para os coeficientes de aproximação até o número de níveis de decomposição especificados pelo analista. O resultado da transformação inclui os coeficientes de aproximação no último nível de decomposição, como também os coeficientes de detalhe obtidos nos vários níveis do banco de filtros^[140]. Tal resultado será chamado daqui por diante de “coeficientes wavelet”.

Os filtros H e G empregados no banco de filtro são tipicamente de comprimento finito. Isto implica que cada coeficiente de aproximação ou detalhe corresponde a uma faixa reduzida de comprimentos de onda dentro do espectro. Vale salientar que diferentes tipos de filtro podem ser empregados na transformação, não havendo regra simples e geral para guiar a escolha do tipo mais apropriado^[136,139]. Recomenda-se, portanto, testar diferentes filtros e adotar alguma métrica de comparação.

6.4. Objetivos

Avaliar o uso do LIBS quanto ao seu potencial para a classificação de solos brasileiros em três diferentes ordens: argissolo, latossolo e nitossolo;

Comparar o desempenho de classificação do SPA-LDA com o SIMCA (em quatro níveis de significância do Teste- F : 1%, 5%, 10% e 25%), GA-LDA e SW-LDA em função do número de erros para o conjunto de teste;

Avaliar o uso da compressão wavelet (WC : *Wavelet Compression*) para diminuir a dimensionalidade dos espectros LIBS e reduzir o esforço computacional envolvido na construção dos modelos de classificação.

6.5. Experimental

6.5.1. Amostras de solos brasileiros

Cento e quarenta e nove amostras de solos de três diferentes ordens (argissolo, latossolo e nitossolo), coletadas do horizonte B, foram fornecidas e previamente classificadas pelo Instituto Agrônomo de Campinas. O número de amostras para cada tipo de solo é apresentado na **Tabela 6.1**.

Tabela 6.1. Número de amostras de cada tipo de solo brasileiro analisado.

	Argissolo	Latossolo	Nitossolo
Número de Amostras	46	84	19

Antes do registro dos espectros LIBS, as amostras de solos foram secas a uma temperatura de 105°C por aproximadamente 2,5 h. Em seguida, essas

amostras foram trituradas e peneiradas para um tamanho de partícula maior do que 250 μm e menor do que 350 μm . Posteriormente, foram armazenadas em recipientes de vidro e estocadas em dessecadores até o início das análises.

6.5.2. Instrumento LIBS

A **Figura 6.2** mostra o Instrumento LIBS construído pelo Grupo de Instrumentação e Automação em Química Analítica (GIA) da UNICAMP que foi utilizado para o registro dos espectros.

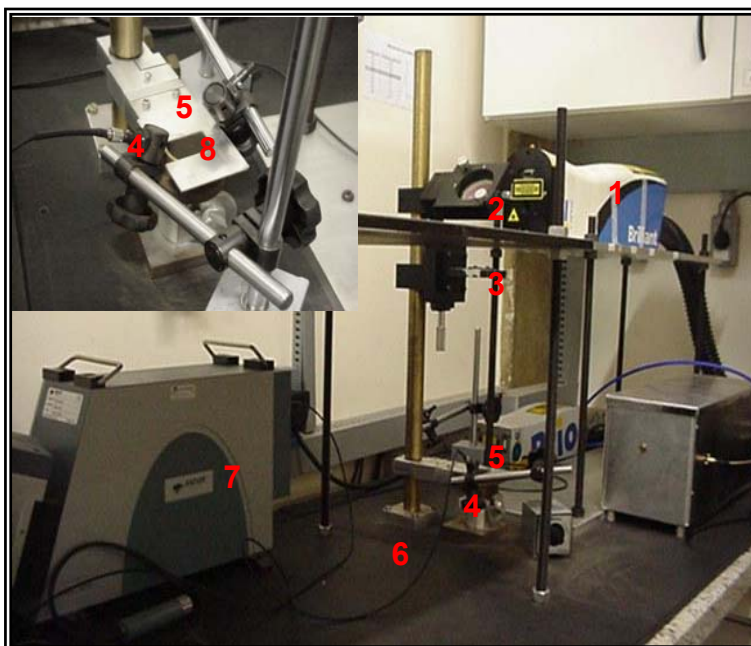


Figura 6.2. Instrumento LIBS construído em laboratório. 1: laser; 2: espelho dicróico; 3: lente; 4: lente coletora de luz; 5: placa de posicionamento; 6: fibra ótica, 7: policromador echelle e 8: cela contendo a amostra. Os detalhes dos componentes 4 e 5 encontram-se ampliados no lado superior esquerdo da figura.

Uma célula (**Figura 6.3a**) foi construída para compactar as amostras de solos e registrar os espectros LIBS. Na **Figura 6.2**, a mesma encontra-se ajustada próximo à placa de posicionamento, perto da lente coletora de luz.

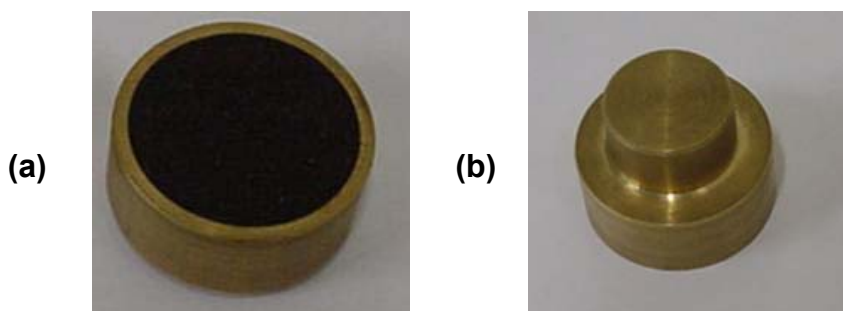


Figura 6.3. (a): célula de latão contendo a amostra de solo. (b): suporte usado para deixar a superfície plana após algumas análises.

6.5.3. Aquisição dos espectros

Para cada amostra de solo, aplicava-se um pulso de laser por locação em diferentes pontos da amostra compactada na célula (**Figura 6.3a**). Conseqüentemente, a superfície da amostra se tornava irregular e com algumas crateras. Para deixá-la novamente plana e uniforme, o suporte apresentado na **Figura 6.3.b** foi utilizado depois de cada cinco disparos do laser. Este procedimento foi repetido até completar 30 espectros por amostra. O espectro médio dessas medidas foi utilizado para o tratamento quimiométrico.

A energia do laser, o tempo de atraso e o tempo de integração foram 110 mJ/pulso, 500 ns e 10 μ s, respectivamente. A distância foco-amostra foi de 0,5 cm. Os espectros foram adquiridos na região de 203,13 nm a 987,64 nm. Cada espectro resultante apresentou 26624 variáveis.

6.5.4. Tratamento dos dados e softwares

A SNV foi utilizada como pré-tratamento inicial dos dados. Em seguida, os conjuntos foram divididos em treinamento, validação e teste utilizando, assim como nos capítulos anteriores, o algoritmo KS. O número de amostra para cada conjunto e classe é apresentado na **Tabela 6.2**.

Tabela 6.2. Número de amostras de treinamento, validação e teste para cada classe de solo.

Classe	Conjuntos		
	Treinamento	Validação	Teste
Argissolo	24	11	11
Latossolo	40	22	22
Nitossolo	11	4	4
Total	75	37	37

As amostras de treinamento e validação foram usadas no procedimento de modelagem (incluindo seleção de variáveis pelo SPA para LDA, escolha do limiar adotado para o SW e para a determinação do número ótimo de PCs para o SIMCA). O terceiro conjunto (teste) foi usado, assim como nos outros capítulos, para a avaliação final e comparação do desempenho dos modelos de classificação.

Para o procedimento de WC, 22 wavelets foram testadas (Symlet 4 – 10, Daubechies 1 – 10, Coiflet 1 – 5)^[140-141] utilizando, cada uma, o nível máximo possível de decomposição. Um percentual de 95% da variância explicada dos dados retidos no processo de compressão foi utilizado como limiar de corte.

SNV, PCA e SIMCA foram realizados com o programa Unscrambler®, versão 9.6 (CAMO S.A.) Quatro diferentes níveis de significância do teste-*F* para classificação SIMCA foram investigados. As rotinas para WC, GA-LDA, SW-LDA e SPA-LDA foram implementadas em Matlab® versão 6.5. Para o GA, foram utilizadas 200 gerações com 400 cromossomos cada. As probabilidades de cruzamento e mutação foram de 60% e 10%, respectivamente, como nos capítulos III e IV. O GA-LDA foi repetido três vezes, a partir de populações diferentes. A melhor solução (em termos do valor de aptidão) resultante dessas três realizações foi empregada. O SW-LDA utilizado nesse estudo foi o mesmo empregado por Caneca *et al.*^[27], descrito na seção 2.4.2. Contudo, sete valores de limiar para o coeficiente de correlação múltipla foram testados: 0,1; 0,2; 0,5; 0,7; 0,8; 0,9 e 0,95. O melhor limiar foi selecionado com base no número de erros de classificação para o conjunto de validação.

6.6. Resultados e Discussão

A **Figura 6.4** apresenta quatro espectros LIBS de uma mesma amostra de solo na faixa de 203,13 – 987,64 nm.

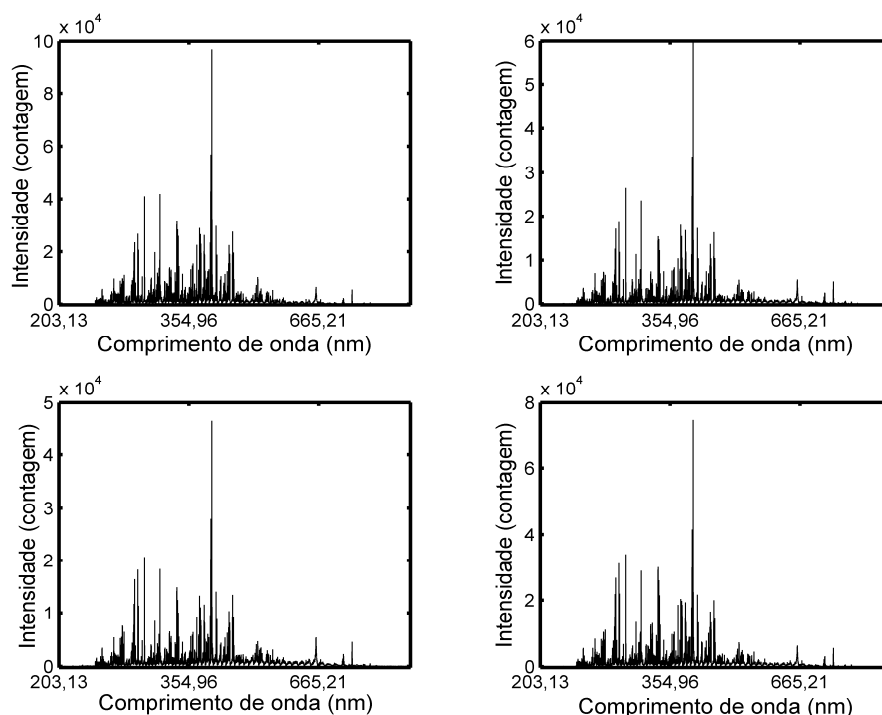


Figura 6.4. Espectros LIBS originais de uma mesma amostra de Argissolo.

Como pode ser visto, existe uma grande variabilidade na intensidade de emissão para cada um desses espectros. Como a técnica LIBS é baseada em uma

análise pontual, este problema pode ser atribuído à heterogeneidade das amostras de solos. Além disso, a formação do plasma é um processo complexo e difícil de ser reproduzido, o que contribui também para a variabilidade dos espectros obtidos em diferentes locações da mesma amostra. Este problema foi minimizado através do emprego da SNV (**Figura 6.5**), que foi aplicada aos espectros individuais para cada amostra.

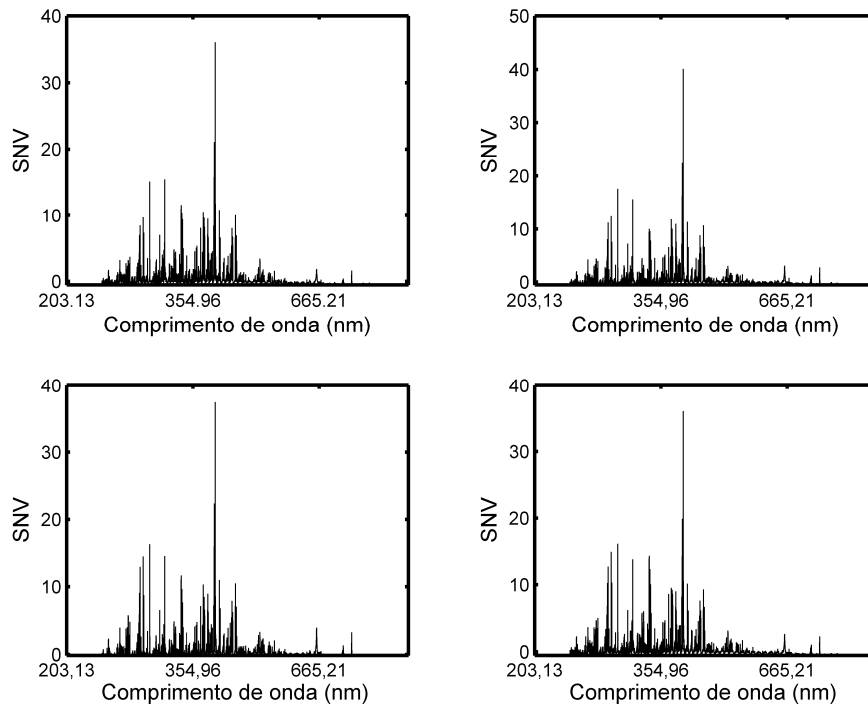


Figura 6.5. Espectros LIBS pré-processados (SNV) da mesma amostra de Argissolo.

Em seguida, a média dos 30 espectros pré-processados por amostra foi calculada. Os espectros resultantes para amostras de cada uma das três classes são apresentados na **Figura 6.6**. Nestes, torna-se difícil distinguir as diferentes ordens de solos com base no aspecto visual dos espectros. Constata-se, portanto, a importância do emprego de técnicas de reconhecimento de padrões e seleção de variáveis.

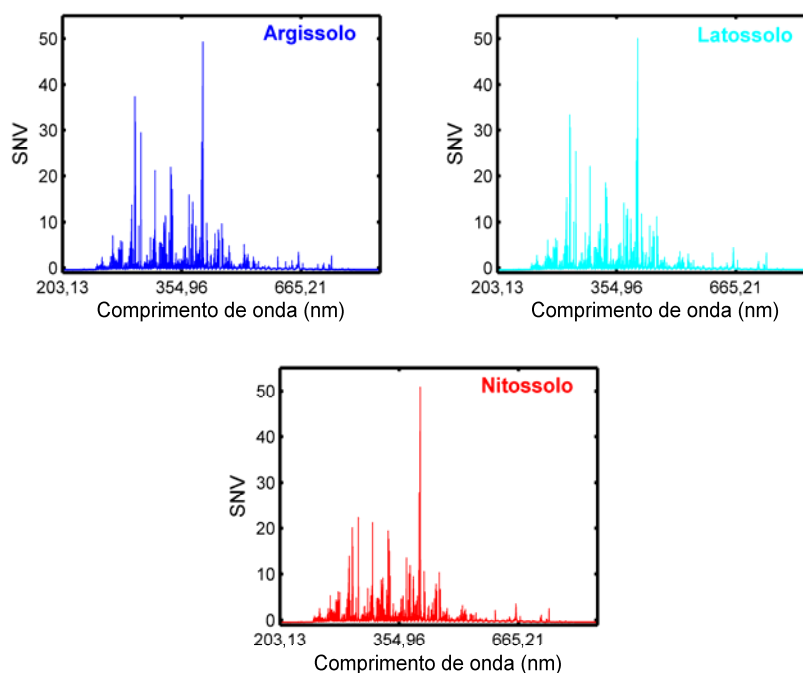


Figura 6.6. Espectros LIBS das 149 amostras de solos brasileiros analisados.

6.6.1. Classificação no domínio espectral original

Inicialmente, uma PCA foi aplicada nas 149 amostras de solos brasileiros estudadas. Os gráficos dos escores obtidos por PC2 × PC1 e PC3 × PC1 são apresentados na **Figura 6.7** e **Figura 6.8**, respectivamente.

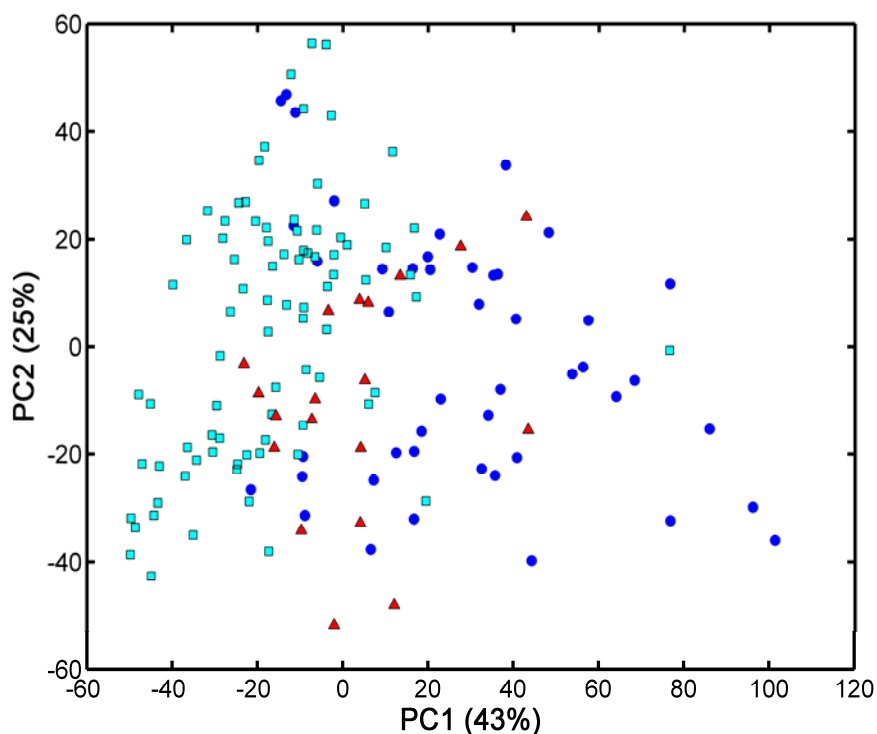


Figura 6.7. Gráficos dos escores obtidos por PC2 × PC1 para as 149 amostras de solos brasileiros. ●: Argissolo, ■: Latossolo e ▲: Nitossolo. O percentual de variância explicada para cada PC encontra-se indicado entre parênteses.

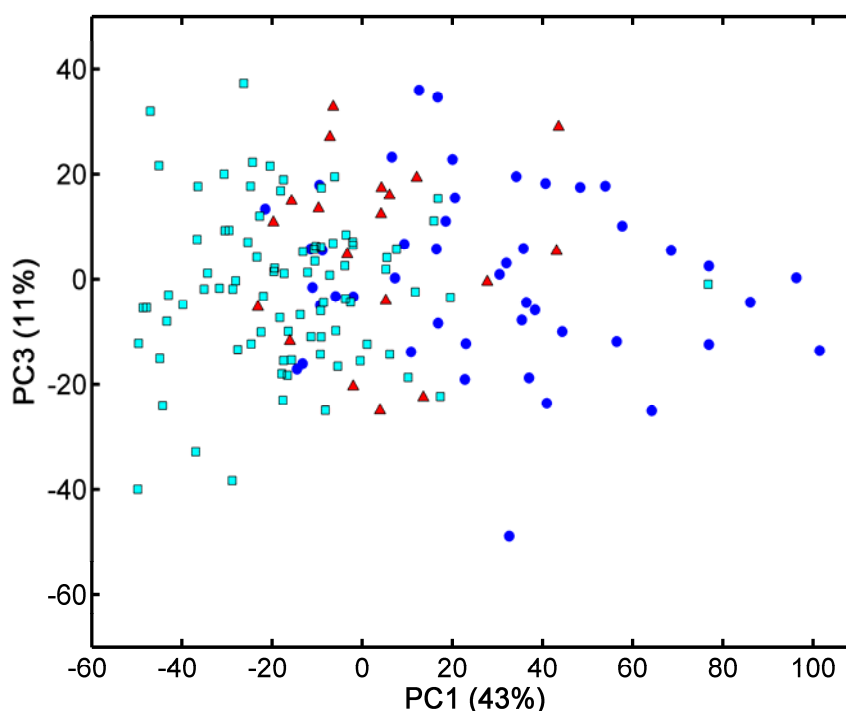


Figura 6.8. Gráficos dos escores obtidos por PC3 × PC1 para as 149 amostras de solos brasileiros. ●: Argissolo, ■: Latossolo e ▲: Nitossolo. O percentual de variância explicada para cada PC encontra-se indicado entre parênteses.

Em ambos os gráficos, pouca discriminação entre as três ordens de solo pode ser observada. Assim como nos problemas de classificação apresentados nos capítulos III e IV, pode-se concluir que a PCA foi pouco eficiente no contexto da caracterização dos diferentes grupos de amostras envolvidos.

6.6.1.1. Classificação SIMCA

Modelos SIMCA com diferentes níveis de significância do teste- F (1%, 5%, 10% e 25%) foram construídos para cada classe e o número de erros de classificação para o conjunto de teste encontra-se na **Tabela 6.3**.

Tabela 6.3. Número de erros de classificação obtido pelos modelos SIMCA (nos níveis de significância do teste- F de 1%, 5%, 10% e 25%) para um conjunto de amostras de teste de solos.

Modelo	Argissolo (7 PCs)				Latossolo (7 PCs)				Nitossolo (8 PCs)			
	1	5	10	25	1	5	10	25	1	5	10	25
Argissolo	0	0	0	1	10	8	7	5	7	7	7	6
Latossolo	21	21	20	15	0	0	0	3	14	14	14	14
Nitossolo	4	4	4	3	4	4	4	2	0	0	0	0

Os erros do Tipo I, destacados nas células de cor cinza, foram poucos, ocorrendo apenas em amostras pertencentes às classes Argissolo e Latossolos a um nível de significância de 25%. Em contrapartida, os erros do Tipo II foram mais

Capítulo VI. Classificação de solos brasileiros

freqüentes para todas as amostras e em todos os níveis de significância do teste-*F*. Os resultados apresentados pelos modelos SIMCA condizem com aqueles obtidos pela PCA. De fato, uma grande dispersão e sobreposição das classes de solos são observadas nas **Figuras 6.7 e 6.8**, o que leva a um número maior de erros do Tipo II.

6.6.1.2. Modelos GA-LDA, SW-LDA e SPA-LDA

Entre os três algoritmos de seleção de variáveis avaliados, o GA foi aquele que apresentou o número maior de variáveis (12) e erros de classificação para o conjunto Teste (11 de 37 amostras foram incorretamente classificadas).

Para o SW-LDA, sete valores de limiar foram testados, conforme descrito na seção 6.5.4. Cinco comprimentos de onda foram selecionados e o limiar escolhido, baseado no número menor de erros para o conjunto de validação, foi 0,1. O modelo LDA resultante com as cinco variáveis selecionadas apresentou seis erros para o conjunto de teste.

Com o SPA-LDA, cinco variáveis foram também selecionadas, conforme o gráfico de *scree* apresentado na **Figura 6.9**.

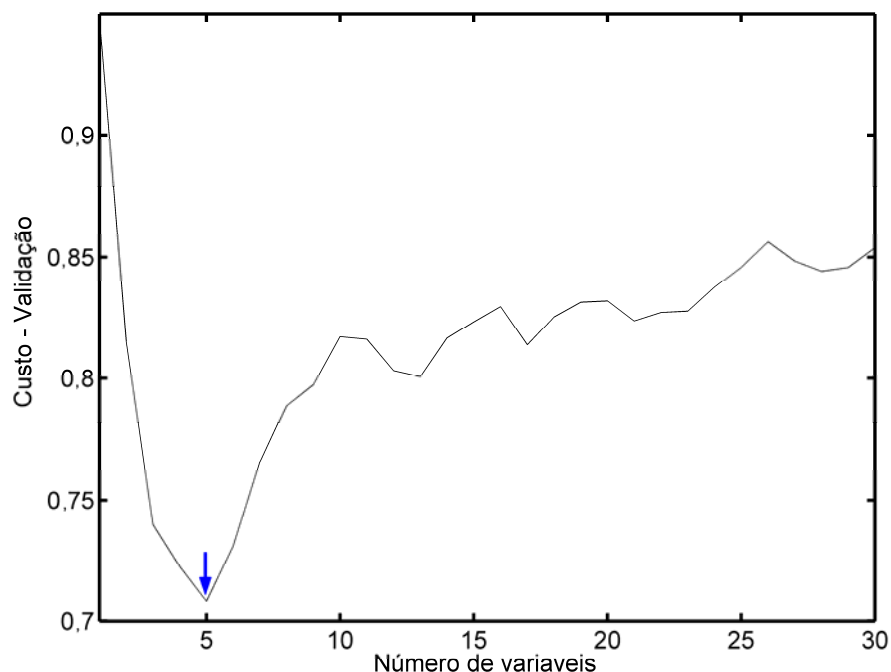


Figura 6.9. Gráfico de *scree* obtido pelo SPA-LDA para os espectros LIBS de solos brasileiros.

Os cinco comprimentos de onda selecionados pelo SPA-LDA encontram-se destacados na **Figura 6.10 e Figura 6.11**.

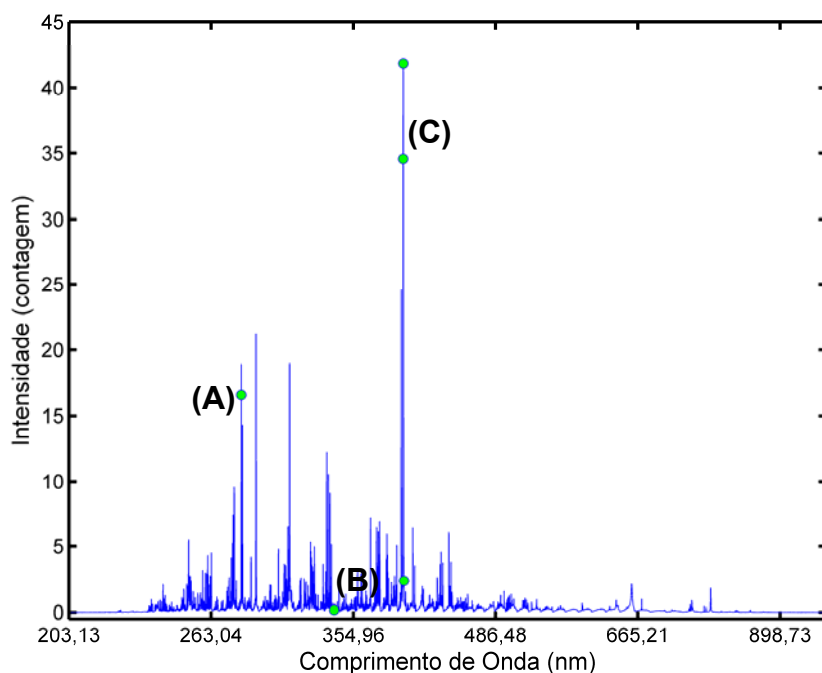


Figura 6.10. Espectro médio da classe Nitossolo com as variáveis selecionadas pelo SPA-LDA. (A), (B) e (C) indicam as regiões ampliadas na Figura 6.11.

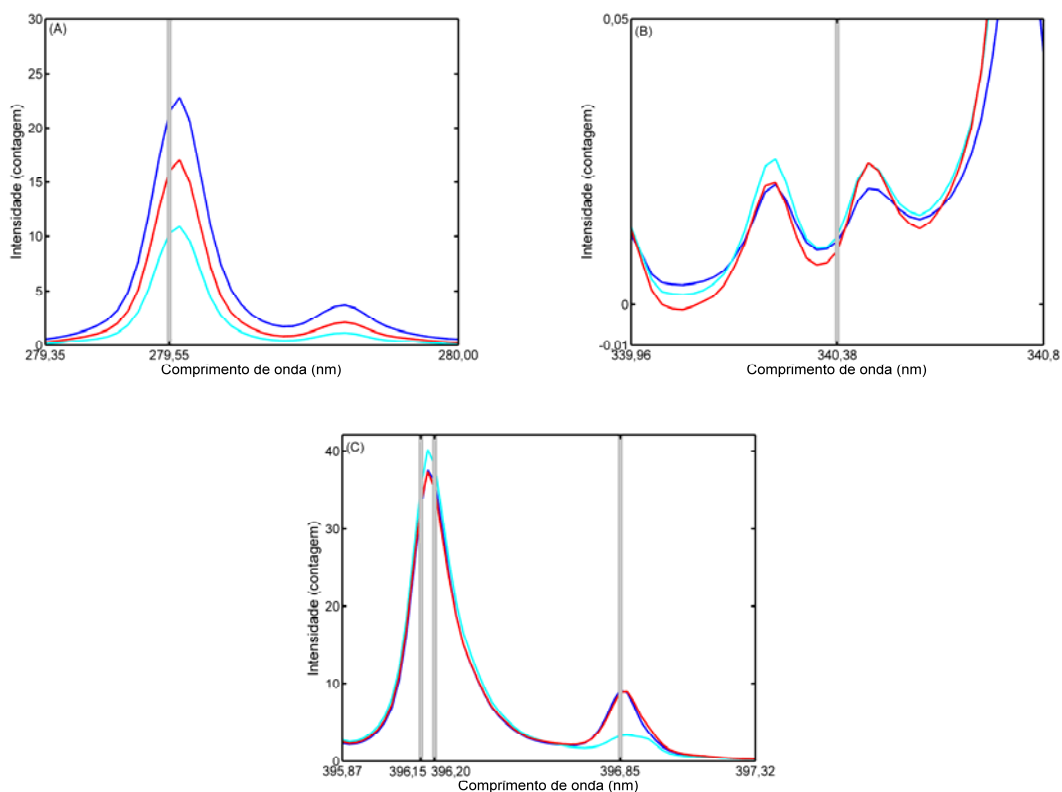


Figura 6.11. Espectros LIBS médio ampliados de cada classe com as variáveis selecionadas pelo SPA-LDA. Argissolo: —; Latossolo: — e Nitossolo: —.

As cinco variáveis selecionadas pelo SPA-LDA correspondem aos comprimentos de onda: 279,55; 340,38; 396,15; 396,20 e 396,85 nm.

Capítulo VI. Classificação de solos brasileiros

O primeiro comprimento de onda selecionado (em torno de 279 nm) pelo SPA pode estar associado à raia do manganês (Mn I). Nele, as três classes encontram-se com intensidade de emissão bem diferentes (**Figura 6.11a**). O SPA selecionou uma variável localizada próxima à linha de base do espectro (340,38 nm), com intensidade de emissão muito baixa. Mesmo assim, com uma ampliação nessa região, pode-se perceber que as intensidades de emissão, por mais que pequenas, são diferentes para a classe do Nitossolo. Curiosamente, o SPA selecionou dois comprimentos de onda localizados em regiões muito próximas (396,15 nm e 396,20 nm) que, possivelmente, devem estar associados à raia do alumínio (Al I). Finalmente, a última variável selecionada pelo SPA encontra-se localizada na região de 396,85 nm. Essa região pode estar associada à raia iônica do cálcio (Ca II), onde o perfil da média da classe dos Latossolo é bem diferente das demais ordens^[142].

O modelo LDA construído com os cinco comprimentos de onda selecionados resultou, assim como o SW-LDA, em seis erros de classificação para o conjunto de teste. O resultado de classificação para os três diferentes métodos de seleção de variáveis é apresentado na **Tabela 6.4**.

Tabela 6.4. Número de erros obtidos pelos modelos GA-LDA, SW-LDA e SPA-LDA em amostras de solos brasileiros do conjunto de teste.

Classe	GA-LDA (12)			SW-LDA (5)			SPA-LDA (5)		
	Arg.	Lat.	Nit.	Arg.	Lat.	Nit.	Arg.	Lat.	Nit.
Argissolo	6	2	4	3	2	1	4	1	3
Latossolo	2	2	0	0	2	2	0	1	1
Nitossolo	1	2	3	0	1	1	0	1	1

Os valores entre parênteses representam o número de variáveis selecionadas pelo GA-LDA, SW-LDA ou SPA-LDA.

É importante lembrar que, em LDA, o número de erros do Tipo I é idêntico ao do Tipo II, ou seja, se uma amostra deixar de ser classificada em sua classe verdadeira, obrigatoriamente será classificada em uma classe errada (erro do Tipo II). Com o SPA-LDA, quatro amostras de argissolo deixaram de ser classificadas em sua classe verdadeira. Dessas quatro, uma foi classificada na classe Latossolo e três no grupo dos Nitossolos. O SW-LDA e SPA-LDA apresentaram um desempenho semelhante e melhor do que o GA-LDA. Contudo, para otimizar o conjunto de variáveis pelo SW, torna-se necessário avaliar alguns valores de limiar.

Os resultados até aqui apresentados foram provenientes de um estudo de classificação envolvendo espectros LIBS com 26624 variáveis. A alta dimensionalidade apresentada nesses sinais provocou um elevado tempo

Capítulo VI. Classificação de solos brasileiros

computacional para a construção dos modelos. O tempo aproximado para executar o SPA-LDA foi de 7 h e 20 minutos. O computador utilizado foi um Pentium core 2 quad, 2,66 GHz, 3 Gb de RAM.

Um procedimento de compressão wavelet foi, então, realizado com intuito de reduzir a dimensionalidade dos dados e o tempo computacional. O resultado desse estudo será apresentado na próxima seção.

6.6.2. Classificação no domínio dos coeficientes wavelet

Assim como descrito na seção 6.5.4, 22 wavelets foram testadas (*Symlet 4-10*, *Daubechies 1-10* e *Coiflet 1-5*) para a compressão dos espectros LIBS. O critério adotado para a escolha da melhor wavelet foi a parcimônia, ou seja, escolheu-se aquela com o menor número de coeficientes retidos no processo de compressão. A **Tabela 6.5** apresenta os resultados, que são expressos em termos do número de coeficientes wavelet necessários para explicar 95% da variância explicada dos dados.

Tabela 6.5. Número de coeficientes wavelet necessários para explicar 95% da variância dos dados.

Wavelet	Nº. de coef. retidos	Wavelet	Nº. de coef. retidos	Wavelet	Nº de coef. retidos
Sym4	660	Db1	782	Coif1	674
Sym5	680	Db2	688	Coif2	674
Sym6	693	Db3	687	Coif3	697
Sym7	698	Db4	734	Coif4	715
Sym8	720	Db5	749	Coif5	747
Sym9	726	Db6	777		
Sym10	748	Db7	813		
		Db8	854		
		Db9	861		
		Db10	892		

O melhor desempenho foi obtido com a wavelet *symlet 4* (sym4), que comprimiu as 26624 variáveis espectrais em 660 coeficientes wavelet em um intervalo de tempo de aproximadamente 4 s. Este novo conjunto dados foi, então, utilizado para construção dos mesmos modelos de classificação abordados na seção 6.6.1.

6.6.2.1. Classificação SIMCA no domínio wavelet

A **Tabela 6.6** mostra o número de erros obtidos pelos modelos SIMCA construídos no domínio dos coeficientes wavelet.

Capítulo VI. Classificação de solos brasileiros

Tabela 6.6. Número de erros de classificação SIMCA (no domínio wavelet) obtidos para o conjunto de teste.

Modelo	Argissolo (7 PCs)				Latossolo (7 PCs)				Nitossolo (8 PCs)			
	1	5	10	25	1	5	10	25	1	5	10	25
Nível (%)	1	5	10	25	1	5	10	25	1	5	10	25
Argissolo	0	0	0	0	10	8	7	5	8	7	8	3
Latossolo	21	21	20	15	0	0	0	5	16	14	15	5
Nitossolo	4	4	4	3	4	4	4	2	0	0	0	1

Os resultados obtidos pelos modelos SIMCA no domínio wavelet são parecidos com aqueles obtidos no domínio espectral original (**Tabela 6.3**), ou seja, um número elevado de erros do Tipo II é apresentado para todos os níveis de significância do Teste-*F*.

6.6.2.2. GA-LDA, SW-LDA e SPA-LDA no domínio wavelet

A **Tabela 6.7** apresenta os resultados de classificação dos modelos GA-LDA, SW-LDA e SPA-LDA construídos no domínio wavelet.

Tabela 6.7. Número de erros de classificação dos modelos GA-LDA, SW-LDA e SPA-LDA construídos no domínio wavelet (conjunto de teste). Os valores entre parênteses indicam o número de coeficientes wavelet selecionados pelos modelos

Classe	GA-LDA (4)			SW-LDA (7)			SPA-LDA (3)		
	Arg.	Lat.	Nit.	Arg.	Lat.	Nit.	Arg.	Lat.	Nit.
Argissolo	3	2	1	3	2	1	3	2	1
Latossolo	0	1	1	0	2	2	0	2	2
Nitossolo	0	1	1	0	1	1	0	1	1

No domínio wavelet, o número de erros apresentados pelo modelo SW-LDA foi idêntico ao SPA-LDA, assim como nos dados originais. Com o GA-LDA, apenas cinco amostras foram incorretamente classificadas.

Para propósito de comparação, a **Tabela 6.8** é apresentada com o resumo dos erros de classificação obtidos no domínio wavelet e nos dados originais.

Tabela 6.8. Resumo final dos erros de classificação para os modelos SIMCA, GA-LDA, SW-LDA e SPA-LDA obtidos no domínio das variáveis originais e dos coeficientes wavelet.

	SIMCA [1%]	SIMCA [5%]	SIMCA [10%]	SIMCA [25%]	GA-LDA 12 {4}	SW-LDA 5 {7}	SPA-LDA 5 {3}
Tipo I	0 (0)	0 (0)	0 (0)	4 (6)	11 (5)	6 (6)	6 (6)
Tipo II	60 (63)	58 (61)	56 (58)	45 (33)	11 (5)	6 (6)	6 (6)
Total	60 (63)	58 (61)	56 (58)	49 (39)	22 (10)	12 (12)	12 (12)

Os valores apresentados entre parênteses indicam o número de erros obtido no domínio dos coeficientes wavelet. Entre chaves, encontram-se os coeficientes wavelet selecionados pelo GA-LDA, SW-LDA ou SPA-LDA.

Com base nos valores apresentados na **Tabela 6.8**, pode-se concluir que o uso da WC não comprometeu o desempenho dos modelos de classificação. O

número de erros para o conjunto de teste, tanto para o SW-LDA, como para o SPA-LDA, permaneceu constante. Contudo, o SPA-LDA foi mais parcimonioso, selecionando um número menor de variáveis (três coeficientes wavelet). Para o SIMCA (nível de significância de 25%) e o GA-LDA, os resultados obtidos após a WC foram ainda melhores do que quando aplicados aos espectros originais.

É importante ressaltar que o esforço computacional envolvido no processo de modelagem foi substancialmente reduzido pelo uso da WC em todas as estratégias estudadas. Com o SPA-LDA, por exemplo, foi possível reduzir de 7 h e 20 minutos para aproximadamente 4 min e 40 s, utilizando o mesmo computador especificado na seção 6.6.1.2.

6.7. Considerações Finais

Neste capítulo, foi apresentada uma nova metodologia analítica baseada na combinação dos espectros LIBS com o SPA-LDA para a classificação de solos brasileiros em três diferentes ordens (Argissolo, Latossolo e Nitossolo).

Os modelos LDA construídos com todas as três estratégias de seleção de variáveis (GA, SW e SPA) apresentaram um desempenho melhor do que o SIMCA. Especificamente, os melhores resultados foram obtidos com o SW-LDA e SPA-LDA. Contudo, o SW tem a desvantagem de buscar, entre vários valores de limiar, o mais adequado.

O procedimento de compressão wavelet proposto foi de grande utilidade na redução da dimensionalidade dos dados (26624 variáveis para 660 coeficientes wavelet) e do tempo computacional, sem comprometer o desempenho dos modelos.

A escolha da wavelet foi baseada no critério da parcimônia (menor número de coeficientes retidos no processo de compressão). Com intuito de melhorar os resultados de classificação, torna-se importante, em trabalhos futuros, investigar outros critérios de escolha, ou ainda, outras técnicas de compressão de dados.

Entre os quatro problemas de classificação apresentados nessa tese, este certamente foi o mais complicado, não apenas por se tratar de amostras complexas e com dificuldades pré-estabelecidas de padronização, mas também por utilizar dados de alta dimensionalidade provenientes de uma técnica analítica ainda em fase de desenvolvimento. Mesmo assim, o SPA-LDA obteve um índice de acerto de 84% (seis erros de classificação) para o conjunto de teste.

CAPÍTULO VII
CONCLUSÕES

7.0 CONCLUSÕES

Nesta tese, foi apresentada uma proposta para uso do SPA em problemas de classificação. Para essa finalidade, uma nova função de custo associada ao risco de classificação incorreta pela LDA foi concebida para guiar a escolha de variáveis.

O SPA-LDA foi validado em quatro problemas de classificação envolvendo três diferentes técnicas analíticas (espectrometria UV-VIS, NIR e LIBS).

O SPA-LDA se mostrou eficiente na seleção de subconjuntos representativos de variáveis e com um desempenho de classificação superior ao SIMCA, método bem estabelecido na literatura e apropriado para trabalhar com dados de alta dimensão.

A metodologia proposta apresentou um desempenho de classificação superior ou similar ao GA-LDA. Todavia, o SPA-LDA foi mais parcimonioso, selecionando sempre um número menor de variáveis.

Nos três primeiros problemas de classificação apresentados, foi realizado um estudo de sensibilidade e robustez dos modelos com respeito à adição de ruído extra aos espectros do conjunto de teste. O SPA-LDA foi menos sensível que o GA-LDA e o SIMCA em relação à presença do ruído. Naturalmente isso ocorreu porque, diferentemente do GA, as variáveis selecionadas pelo SPA localizavam-se em regiões mais informativas e com alta relação sinal/ruído.

No último problema de classificação, o SPA-LDA foi também comparado com o SW-LDA. O mesmo número de erros para um conjunto de teste foi obtido pelas duas estratégias. Entretanto, o SW-LDA apresenta a desvantagem de ter necessariamente que testar diferentes valores de limiar para encontrar o conjunto de variáveis apropriado.

Um procedimento de compressão wavelet foi também avaliado no último estudo de caso. Como o conjunto de dados apresentava alta dimensionalidade, o tempo computacional gasto para construir os modelos de classificação foi elevado. Com o procedimento de compressão, foi possível reduzir o número de variáveis de 26624 comprimentos de onda para 660 coeficientes wavelet. Conseqüentemente, o tempo de execução do cálculo foi drasticamente reduzido, sem comprometer o desempenho de classificação dos modelos.

Enfim, este trabalho mostrou que o SPA-LDA pode ser uma alternativa vantajosa para seleção de variáveis espectrais em problemas analíticos de classificação.

7.1. Propostas futuras

As propostas futuras para a continuidade deste trabalho são:

- *No estudo de classificação de solos, pretende-se aumentar o número de amostras e ordens envolvidas no problema.*
- *Avaliar o potencial do SPA-LDA em outros problemas de classificação e em outras técnicas analíticas.*

Referências Bibliográficas

- [1] SKOOG, D. A.; LEARY, J. J. ***Principles of instrumental analysis***. 6. ed. New York: Saunders College Publishing, 1992.
- [2] ALEIXO, B. L. M.; STEIN, E.; GODINHO, O. E. S. ***Introdução à semimicroanálise qualitativa***. 3. ed. Campinas: Editora Unicamp, 1990.
- [3] BERRUETA, L. A.; ALONSO-SALCES, R. M.; HÉBERGER, K., Supervised pattern recognition in food analysis, *Journal of Chromatography A*, **1158:196, 2007**.
- [4] GONZÁLEZ, A. G., Use and misuse of supervised pattern recognition methods for interpreting compositional data, *Journal of Chromatography A*, **1158:215, 2007**.
- [5] BEEBE, K. R.; PELL, R. J.; SEASHOLTZ, M. B. ***Chemometrics a practical guide***. New York: John Wiley Y Sons, 1998.
- [6] PEREIRA, A. F. C.; PONTES, M. J. C.; GAMBARRA NETO, F. F.; SANTOS, S. R. B.; GALVÃO, R. K. H.; ARAÚJO, M. C. U., NIR spectrometric determination of quality parameters in vegetable oils using iPLS and variable selection, *Food Research International*, **41: 341,2008**.
- [7] MASSART, D. L.; VANDEGINSTE, B. G. M.; BUYDENS, S. J.; Lewi, P. J.; Smeyers-Verbeke. ***Journal Handbook of Chemometrics and Qualimetrics: Parte B***, Amsterdam: Elsevier, 1997.
- [8] MOREDA-PINEIRO, A.; FISHER, A.; HILL, S.J., The classification of tea according to region of origin using pattern recognition techniques and trace metal data, *Journal of Food Composition and Analysis*, **16:196, 2003**.
- [9] MARQUARDT, B. J.; WOLD, J., P., Raman analysis of fish: a potential methodo for rapid quality screening, *Lebensmittel-Wissenschaft und-Technologie*, **37:1, 2004**.
- [10] OLIVEIRA, A. P.; GOMES NETO, J. A.; FERREIRA, M. M. C., Uso da análise exploratória de dados na avaliação de modificadores químicos para determinação

Referências Bibliográficas

direta e simultânea de metais em álcool combustível por GFASS, *Eclética Química*, **31:7, 2006**.

[11] CHOI, H-K.; KIM, K-H.; KIM, K. H.; KIM, Y-S.; LEE, M-W.; WHANG, W. K., Metabolomic differentiation of deer antlers of various origins by ¹H NMR spectrometry and principal components analysis, *Journal of Pharmaceutical and Biomedical Analysis*, **41:1047, 2006**.

[12] DRAGOVIC, S.; ONJIA, A., Classification of soil samples according to their geographic origin using gamma-ray spectrometry and principal component analysis, *Journal of Environmental Radioactivity*, **89:150, 2006**.

[13] KAROUI, R.; THOMAS, E.; DUFOUR, E., Utilization of a rapid technique based on front-face fluorescence spectroscopy for differentiating between fresh and frozen-thawed fish fillets, *Food Research International*, **39:349, 2006**.

[14] PONTES, M. J. C.; SANTOS, S. R. B.; ARAÚJO, M. C. U.; ALMEIDA, L. F.; LIMA, R. A. C.; GAIÃO, E. N.; SOUTO, U. T. C. P., Classification of distilled alcoholic beverages and verification of adulteration by near infrared spectrometry, *Food Research International*, **39:182, 2006**.

[15] DU, C.; LINKER, R.; SHAVIV, A., Identification of agricultural Mediterranean soils using mid-infrared photoacoustic spectroscopy, *Geoderma*, **143:85, 2008**.

[16] SAVITZKY, A.; GOLAY, M. J. E., Smoothing and differentiation of data by simplified least-squares procedures, *Analytical Chemistry*, **8:1627, 1964**.

[17] WOLD, S., Pattern recognition by means of disjoint principal component models, *Pattern Recognition*, **8:12, 1976**.

[18] CANDOLFI, A.; MAESSCHALCK, R. D.; MASSART, D. L.; HAILEY, P. A.; HARRINGTON, A. C. E., Identification of pharmaceutical excipients using NIR spectroscopy and SIMCA, *Journal of Pharmaceutical and Biomedical Analysis*, **19:923, 1999**.

Referências Bibliográficas

- [19] WOO, Y-A.; KIM, H-J.; CHO, J., Identification of herbal medicines using pattern recognition techniques with near-infrared reflectance spectra, *Microchemical Journal*, **63:61, 1999**.
- [20] HE, J.; RODRIGUEZ-SAONA, L. E.; GIUSTI, M. M., Mid infrared spectroscopy for juice authentication - Rapid differentiation of commercial juices, *Journal of Agricultural and Food Chemistry*, **55:4443, 2007**.
- [21] WANG, L.; LEE, F. S. C.; WANG, X.; HE, Y., Feasibility study of quantifying and discriminating soybean oil adulteration in camellia oils by attenuated total reflectance MIR and fiber optic diffuse reflectance NIR, *Food Chemistry*, **95:529, 2006**.
- [22] VOGT, N. B.; SJOEGREN, C. E., Investigation of chemical and statistical methods for oil-spill classification, *Analytica Chimica Acta*, **222:135, 1989**.
- [23] KRIZOVÁ, J.; MATEJKA, P.; BUDÍNOVÁ, G.; VOLKA, K., Fourier-transform Raman spectroscopy study of surface of Norway spruce needles, *Journal of Molecular Structure*, **480:547, 1999**.
- [24] LINDON, J. C.; HOLMES, E.; NICHOLSON, J. K., Toxicological applications of magnetic resonance, *Progress in Nuclear Magnetic Resonance Spectroscopy*, **45:109, 2004**.
- [25] CHARLTON, A. J.; ROBB, P.; DONARSKI, J. A.; GODWARD, J., Non-targeted detection of chemical contamination in carbonated soft drinks using NMR spectroscopy, variable selection and chemometrics, *Analytica Chimica Acta*, **618:196, 2008**.
- [26] RAO, K. R.; LAKSHMINARAYANAN, S., Partial correlation based variable selection approach for multivariate data classification methods, *Chemometrics and Intelligent Laboratory Systems*, **86:68, 2007**.
- [27] CANECA, A. R.; PIMENTEL, M. F.; GALVÃO, R. K. H.; MATTA, C. E.; CARVALHO, F. R.; RAIMUNDO JR, I. M.; PASQUINI, C.; ROHWEDDER, J. J. R.,

Referências Bibliográficas

Assessment of infrared spectroscopy and multivariate techniques for monitoring the service condition of diesel-engine lubricating oils, *Talanta*, **70:344, 2006**.

[28] MOREIRA, E. D. T.; PONTES, M. J. C.; GALVÃO, R. K. H.; ARAÚJO, M. C. U., NIR classification of cigarettes using the successive projections algorithm for variable selection, *Vibrational Spectroscopy*, **Artigo submetido, 2008**.

[29] GAMBARRA NETO, F. F.; MARINO, G.; ARAÚJO, M. C. U.; GALVÃO, R. K. H.; PONTES, M. J. C.; MEDEIROS, E. P.; LIMA, R. S., Classification of edible vegetable oils using square wave voltammetry with multivariate data analysis, *Talanta*, **In press, 2008. (doi:10.1016/j.talanta.2008.10.003)**

[30] PONTES, M. J. C.; CORTEZ, J.; GALVÃO, R. K. H.; PASQUINI, C.; COELHO, R. M.; CHIBA, M. K.; ABREU, M. F.; MADARI, B. E., Classification of brazilian soils by using libs and variable selection in the wavelet domain, *Analytica Chimica Acta*, **Artigo submetido, 2008**.

[31] SOUTO, U. T. C. P.; PONTES, M. J. C.; SILVA, E. C.; GALVÃO, R. K. H.; ARAÚJO, M. C. U.; SANCHES, F. A. C.; CUNHA, F. A. S.; OLIVEIRA, M. S. R., UV-Vis spectrometric classification of coffees by SPA-LDA, *Food Chemistry*, **Artigo submetido, 2008**.

[32] FISHER, R. A., The use of multiple measurements in taxonomic problems, *Annales Eugenics*, **7:179, 1936**.

[33] SOLA-LARRAÑAGA, C.; NAVARRO-BLASCO, I., Chemometric analysis of minerals and trace elements in raw cow milk from the community of Navarra, Spain, *Food Chemistry*, **112:189, 2009**.

[34] CÂMARA, J. S.; ALVES, M. A.; MARQUES, J. C., Classification of Boal, Malvazia, Sercial and Verdelho wines base don terpenoid patterns, *Food Chemistry*, **101: 475, 2007**.

Referências Bibliográficas

- [35] SANDERCOCK, P. M. L.; PASQUIER, E. D., Chemical fingerprinting of unevaporated automotive gasoline samples, *Forensic Science International*, **134:1,2003**.
- [36] NAES, T.; MEVIK, B. H., Understanding the collinearity problem in regression and classification, *Journal of Chemometrics*, **15:413, 2001**.
- [37] MALLET, Y.; COOMANS, D.; DE VEL, O., Recent developments in discriminant analysis on high dimensional spectra data, *Chemometrics and Intelligent Laboratory Systems*, **35:157, 1996**.
- [38] PARTRIDGE, M.; CALVO, R. A., Fast dimensionality reduction and simple PCA, *Intelligent Data Analysis*, **2:203, 1998**.
- [39] INDAHL, U. G.; SAHNI, N. S.; KIRKHUS, B.; NAES, T., Multivariate strategies for classification based on NIR-spectra with application to mayonnaise, *Chemometrics and Intelligent Laboratory Systems*, **49:19, 1999**.
- [40] YAN, J.; ZHANG, B.; YAN, S.; LIU, N.; YANG, Q.; CHENG, Q.; LI, H.; CHEN, Z.; MA, W-Y., A scalable supervised algorithm for dimensionality reduction on streaming data, *Information Sciences*, **176:2042, 2006**.
- [41] QUNXIONG, Z. H. U.; CHENGFEI, L. I., Dimensionality reduction with input training neural network and its application in chemical process modelling chinese, *Journal of Chemical Engineering*, **14:597, 2006**.
- [42] TAN, C.; LI, M.; QIN, X., Study of the feasibility of distinguishing cigarettes of different brands using an Adaboost algorithm and near-infrared spectroscopy, *Analytical and Bioanalytical Chemistry*, **389:667, 2007**.
- [43] 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*, **78:11, 2005**.

Referências Bibliográficas

- [44] BALDOVIN, A.; WU, W.; CENTNER, V.; JOUAN-RIMBAUD, D.; MASSART, D. L.; FAVRETTO, L.; TURELLO, A., Feature selection for the discrimination between pollution types with partial least squares modeling, *Analyst*, **121:1603, 1996**.
- [45] DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern Classification*, New York: John Wiley, 2001.
- [46] WU, W.; GUO, Q.; RIMBAUD, D. J.; MASSART, D. L., Using contrasts as data pretreatment method in pattern recognition of multivariate data, *Chemometrics and Intelligent Laboratory Systems*, **45:39, 1999**.
- [47] GUYON, I.; ELISSEEFF, A., An introduction to variable and feature selection, *Journal of Machine Learning Research*, **3:1157, 2003**.
- [48] GOLUB, T. R.; SLONIM, D. K.; TAMAYO, P.; HUARD, C.; GAASENBEEK, M.; MESIROV, J. P.; COLLIER, H.; LOH, M. L.; DOWNING, J. R.; CALIGIURI, M. A.; BLOOMFIELD, C. D.; LANDER, E. S., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, **286:531, 1999**.
- [49] FUREY, T. S.; CRISTIANINI, N.; DUFFY, N.; BEDNARSKI, D. W.; SCHUMMER, M.; HAUSSLER, D., Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, **16:906, 2000**.
- [50] PAVLIDIS P.; WESTON J.; CAI J.; NOBLE, W. S., Learning gene functional classifications from multiple data types, *Journal Computational Biology*, **9:401, 2002**.
- [51] METROPOLIS, N.; ROSENBLUTH, W. W.; ROSENBLUTH, M. N.; TELLER, A. H.; TELLER, E., Equation of state calculations by fast computing machines, *Journal of Chemical Physics*, **21:1087,1953**.
- [52] KIRKPATRICK, S.; GELATT JR., C. D.; Vecchi, M. P., Optimization by simulated annealing, *Science*, **220:671, 1983**.

Referências Bibliográficas

- [53] LLOBET, E.; GUALDRÓN, O.; VINAIXA M.; EL-BARBRI, N.; BREZMES, J.; VILANOVA, X.; BOUCHIKHI, B.; GÓMEZ, R.; CARRASCO, J. A.; CORREIG, X., Efficient feature selection for mass spectrometry based electronic nose applications, *Chemometrics and Intelligent Laboratory Systems*, **85:253, 2007**.
- [54] CENTNER, V.; MASSART, D. L.; NOORD, O. E.; JONG, S.; VANDEGINSTE, B. M.; STERNA, C., Elimination of uninformative variables for multivariate calibration, *Analytical Chemistry*, **68:3851, 1996**.
- [55] LEARDI, R., Genetic algorithms in chemometrics and chemistry: a review, *Journal of Chemometrics*, **15:559, 2001**.
- [56] LEARDI, R.; GONZALEZ, A. L., Genetic algorithms applied to feature selection in PLS regression: how and when to use them, *Chemometrics and Intelligent of Laboratory System*, **41:195, 1998**.
- [57] LEARDI, R., Application of genetic algorithm-PLS for feature selection in spectral data sets, *Journal of Chemometrics*, **14:643, 2000**.
- [58] KEMSLEY, E. K., A hybrid classification method: discrete canonical variate analysis using a genetic algorithm. *Chemometrics and Intelligent of Laboratory Systems*, **55:39, 2001**.
- [59] DHARMARAJ, S.; JAMALUDIN, A. S.; RAZAK, H. M.; VALLIAPPAN, R.; AHMAD, N. A.; HARN, G. L.; ISMAIL, Z., The classification of *Phyllanthus niruri* Linn. according to location by infrared spectroscopy, *Vibrational Spectroscopy*, **41:68, 2006**.
- [60] AVCI, E.; SENGUR, A.; HANBAY, D., An optimum feature extraction method for texture classification, *Expert Systems with Applications*, **Artigo In Press**. doi:10.1016/j.eswa.2008.06.076,2008.

Referências Bibliográficas

- [61] SOLA-LARRAÑAGA, C.; NAVARRO-BLASCO, I., Chemometric analysis of minerals and trace elements in raw cow milk from the community of Navarra, Spain, *Food Chemistry*, 112:189, 2009.
- [62] OLIVEROS, C. C.; BOGGIA, R.; CASALE, M.; ARMANINO, C.; FORINA, M., Optimisation of a new headspace mass spectrometry instrument discrimination of different geographical origin olive oils, *Journal of Chromatography A*, **1076:7, 2005**.
- [63] FORINA, M.; LANTERI, S.; CASALE, M.; OLIVEROS, M. C. C., Stepwise orthogonalization of predictors in classification and regression techniques: An “old” technique revisited, *Chemometrics and Intelligent Laboratory Systems*, **87:252, 2007**.
- [64] ARAÚJO, M. C. U.; SALDANHA, T. C. B.; GALVÃO, R. K. H.; YONEYAMA, T.; CHARME, H. C.; VISANI, V., The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, *Chemometrics and Intelligent Laboratory Systems*, **57:65, 2001**.
- [65] MARTENS, H.; NAES, T. ***Multivariate Calibration***. Chichester, England: John Wiley & Sons, 1989.
- [66] GALVÃO, R. K. H.; PIMENTEL, M. F.; ARAÚJO, M. C. U.; YONEYAMA, T.; VISANI, V., Aspects of the successive projections algorithm for variable selection in multivariate calibration applied to plasma emission spectrometry, *Analytica Chimica Acta*, **443:107, 2001**.
- [67] DANTAS FILHO, H. A.; SOUZA, E. S. O. N.; VISANI, V.; BARROS, S. R. R. C.; SALDANHA, T. C. B.; ARAÚJO, M. C. U.; GALVÃO, R. K. H., Simultaneous spectrometric determination of Cu²⁺, Mn²⁺ and Zn²⁺ in polivitaminic/polimineral drug using SPA and GA algorithms for variable selection, *Journal of Brazilian Chemical Society*, **16:58, 2005**.
- [68] DANTAS FILHO, H. A.; GALVÃO, R. K. H.; ARAÚJO, M. C. U.; SILVA, E. C.; SALDANHA, T. C. B.; JOSÉ, G. E.; PASQUINI, G. E.; RAIMUNDO JR.; I. M.;

Referências Bibliográficas

ROHWEDDER, J. J. R., A strategy for selecting calibration samples for multivariate modeling, *Chemometrics and Intelligent Laboratory Systems*, **72:83, 2004**.

[69] HONORATO, F. A.; GALVÃO, R. K. H.; PIMENTEL, M. F.; BARROS NETO, B.; ARAÚJO, M. C. U.; CARVALHO, F. R., Robust modeling for multivariate calibration transfer by the successive projections algorithm, *Chemometrics and Intelligent Laboratory Systems*, **76:65, 2005**.

[70] BREITKREITZ, M. C.; RAIMUNDO JR.; I. M., ROHWEDDER, J. J. R., PASQUINI, C.; DANTAS FILHO, H. A.; JOSÉ, G. E.; ARAÚJO, M. C. U., Determination of total sulfur in diesel fuel employing NIR spectroscopy and multivariate calibration, *Analyst*, **128:1204, 2003**.

[71] GALVÃO, R. K. H.; ARAÚJO, M. C. U.; SILVA, E. C.; JOSÉ, G. E.; SOARES, S. F. C.; PAIVA, H. M., Cross-validation for the selection of spectral variables using the successive projections algorithm, *Journal of Brazilian Chemical Society*, **18:1580, 2007**.

[72] GALVÃO, R. K. H.; ARAÚJO, M. C. U.; FRAGOSO, W. D.; SILVA, E. C.; JOSÉ, G. E.; SOARES, S. F. C.; PAIVA, H. M., A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm, *Chemometrics and Intelligent Laboratory Systems*, **92:83, 2008**.

[73] PONTES, M. J. C. *Classificação de bebidas alcoólicas destiladas e verificação de adulteração usando espectrometria NIR e quimiometria*. João Pessoa, Programa de Pós-Graduação em Química, UFPB, 2005. Dissertação de Mestrado.

[74] BARNES, R. J.; DHANOA, M. S.; LISTER, S. J., Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, *Applied Spectroscopy*, **43:772, 1989**.

[75] CHEN, Q.; ZHAO, J.; CHAITEP, S.; GUO, Z., Simultaneous analysis of main catechins contents in green tea (*Camellia sinensis* (L.)) by fourier transform near infrared reflectance (FT-NIR) spectroscopy, *Food Chemistry*, **113:1272, 2009**.

Referências Bibliográficas

- [76] BUREAU, S.; RUIZ, D.; REICH, M.; GOUBLE, B.; BERTRAND, D.; AUDERGON, J-M.; RENARD, C. M. G.C., Rapid and non-destructive analysis of apricot fruit quality using FT-near-infrared spectroscopy, *Food Chemistry*, **113:1323**, **2009**.
- [77] XIA, J-F.; LI, X-Y.; LI, P-W.; MA, Q.; DING, X-X., Application of wavelet transform in the prediction of navel orange vitamin c content by near-infrared spectroscopy, *Agricultural Sciences in China*, **6:1067**, **2007**.
- [78] CERQUEIRA, E. O.; POPPI, R. J.; KUBOTA, L. T., Utilização de filtro de transformada de fourier para a minimização de ruídos em sinais analíticos, *Química Nova*, **23:690**, **2000**.
- [79] WOLD, S; ESBENSEN, K.; GELADI, P., Principal component analysis, *Chemometrics and Intelligent Laboratory Systems*, **2:37**, **1987**.
- [80] CAMO S.A. Manual do Usuário. UNSCRAMBLER 7.5. Noruega. 1998.
- [81] VANDEGINSTE, B. G. M.; SIELHORST, C.; GERRITSEN, M., The nipals algorithm for the calculation of the principal components of a matrix, *TRAC-Trends In Analytical Chemistry*, **7:286**, **1988**.
- [82] MAESSCHALCK, R. D.; JOUAN-RIMBAUD, D.; MASSART, D. L., The Mahalanobis distance, *Chemometrics and Intelligent Laboratory Systems*, **50:1**, **2000**.
- [83] DASH, M.; LIU, H., Feature selection for classification, *Intelligent Data Analysis*, **1:131**, **1997**.
- [84] SIEDLECKI, W.; SKLANSKY, J., On automatic feature selection, *International Journal of Pattern Recognition and Artificial Intelligence*, **2:197**, **1998**.

Referências Bibliográficas

- [85] GALVÃO, R. K. H.; ARAÚJO, M. C. U., Linear regression modelling: variable selection. In: WALCZAK, B., FERRÉ, R. T., BROWN, S., **Comprehensive chemometrics**, 2009.
- [86] GAMBARRA NETO, F. F. *Classificação de óleos vegetais utilizando voltametria de onda quadrada e métodos quimiométricos*. João Pessoa, Programa de Pós-Graduação em Química, UFPB, 2008. Dissertação de Mestrado.
- [87] MORETTO, E.; FEET, R. **Tecnologia de Óleos e Gorduras vegetais na Indústria de Alimentos**. São Paulo: Varela Editora e Livraria, 1998.
- [88] Agência Nacional de Vigilância Sanitária (ANVISA). Resolução nº. 482, RDC 492/99. Disponível em: http://www.anvisa.gov.br/legis/resol/482_99.htm. Acesso em 20 de Dezembro de 2007.
- [89] PEREIRA, A. F. C. *Determinação simultânea de acidez, índice de refração e viscosidade em óleos vegetais usando espectrometria NIR, calibração multivariada e seleção de variáveis*. João Pessoa, Programa de Pós-Graduação em Química, UFPB, 2007. Dissertação de Mestrado.
- [90] Empresa Brasileira de Pesquisa Agropecuária EMBRAPA). Soja na alimentação. Disponível em: <http://www.cnpso.embrapa.br/soja/alimentacao/index.php?pagina=23>. Acesso : 15 de outubro de 2007.
- [91] HOURANT, P.; BAETEN, V.; MORALES, M. T.; MEURENS, M.; APARICIO, R., Oil and fat classification by selected bands of near-infrared spectroscopy, *Applied Spectroscopy*, **54:1168, 2000**.
- [92] LAI, Y. W.; KEMSLEY, E. K.; WILSON, R. H., Potential of fourier transform infrared spectroscopy for the authentication of vegetable oils, *Journal of Agricultural and Food Chemistry*, **42: 1154, 1994**.

Referências Bibliográficas

- [93] BAETEN, V.; MEURENS, M.; MORLES, M. T.; APARICIO, R., Detection of virgin olive oil adulteration by fourier transform raman spectroscopy, *Journal of Agricultural and Food Chemistry*, **44:2225, 1996**.
- [94] GUIMET, F.; FERRÉ, J.; BOQUÉ, R.; RIUS, F. X., Application of unfold principal component analysis and parallel factor analysis to the exploratory analysis of olive oils by means of excitation-emission matrix fluorescence spectroscopy, *Analytica Chimica Acta*, **515:75, 2004**.
- [95] KENNARD, R. W.; STONE, L. A., Computer-aided design of experiments, *Technometrics*, **11:137, 1969**.
- [96] DANTAS FILHO, H. A. *Desenvolvimento de técnicas quimiométricas de compressão de dados e de redução de ruído instrumental aplicadas a óleo diesel e madeira de eucalipto usando espectroscopia NIR*. Campinas, Programa de Pós-Graduação em Química, UNICAMP, 2007. Tese de Doutorado.
- [97] ANP. Portaria N° 310, DE 27.12.2001 - DOU 28.12.2001. Disponível em: http://www.anp.gov.br/doc/petroleo/P310_2001.pdf. Acesso em setembro de 2008.
- [98] American Society for Testing and Material (ASTM), D3120-92.
- [99] American Society for Testing and Material (ASTM), D4294-90.
- [100] Navas, M. J., Jimenez, A. M., Chemiluminescent methods in petroleum products analysis, *Critical Reviews in Analytical Chemistry*, **30:153, 2000**.
- [101] GALVÃO, R. K. H.; JOSÉ, G. E.; DANTAS FILHO, H. A.; ARAÚJO, M. C. U.; SILVA, E. C., PAIVA, H. M.; SALDANHA, T. C. S.; SOUZA, E. S. O. N., Optimal wavelet filter construction using x and y data, *Chemometrics and Intelligent Laboratory Systems*, **70:1, 2004**.
- [102] PASQUINI, C., Fundamentals, practical aspects and analytical applications, *Journal of Brazilian Chemical Society*, **14:198, 2003**.

Referências Bibliográficas

- [103] BOKOBZA, L., Near Infrared Spectroscopy, *Journal of Near Infrared Spectroscopy*, **6:3, 1998**.
- [104] ABIC (Associação Brasileira de Indústria de Café). Disponível em: http://www.abic.com.br/scafe_historia.html. Acessado em 1 de outubro de 2008.
- [105] RIESSELMANN, B.; ROSENBAUM F.; ROSCHER, S.; SCHNEIDER, V. Fatal caffeine intoxication, *Forensic Science International*, **103:S49, 1999**.
- [106] FORMAN, J.; AIZER, A.; YOUNG, C. R., Myocardial infarction resulting from caffeine overdose in an anorectic woman. *Annals of Emergency Medicine*, **29:178, 1997**.
- [107] KERRIGAN, S.; LINDSEY, T., Fatal caffeine overdose: two case reports. *Forensic Science International*, **153:67, 2005**.
- [108] LAFUENTE-LAFUENTE, C.; MOULY, S.; DELCEY, V.; MAHE, I.; DIEMER, M.; CHASSANY, O.; CAULIN, C.; BERGMANN, J. P., Effects of an evening single cup of coffee on the quality of sleep in caffeine-sensitive patients: a randomised cross-over trial, *Fundamental & Clinical Pharmacology*, **22:60, 2008**.
- [109] MYERS, M. G. Effects of caffeine on blood pressure. *Archives of Internal Medicine*, **148:1189, 1998**.
- [110] VLAJINAC, H. D.; PETROVIC, R. R.; MARINKOVIC, J. M.; SIPETIC, S. B., Effects of caffeine intake during pregnancy on birth wight, *American Journal of Epidemiology*, **145:335, 1997**.
- [111] CARDELLI, C.; LABUZA, T. P., Application of wibull hazard analysis to the determination of the shelf life of roasted and ground coffee, *Lebensm.-Wiss. U.-Technol.*, **34:273, 2001**.

Referências Bibliográficas

- [112] GONZÁLEZ, A. G.; PABLOS, F.; MARTIN, M. J.; LEÓN-CAMACHO, M.; VALDENEbro, M. S., HPLC analysis of tocopherols and triglycerides in coffee and their use as authentication parameters, *Food Chemistry*, **73:93, 2001**.
- [113] AGREStI, P. D. C. M.; FRANCA, A. S.; OLIVEIRA, L. S.; AUGUSTI, R., Discrimination between defective and non-defective brazilian coffee beans by their volatile profile, *Food Chemistry*, **106:787, 2008**.
- [114] FERNANDES, A. P.; SANTOS, M. C.; LEMOS, S. G.; FERREIRA, M. M. C; NOGUEIRA, A. R. A.; NÓBREGA, J. A., Pattern recognition applied to mineral characterization of brazilian coffees and sugar-cane spirits, *Spectrochimica Acta Part B*, **70:717, 2005**.
- [115] CHARLTON, A. J.; FARRINGTON, W. H. H.; BRERETON, P., Application of ¹H NMR and multivariate statistics for screening complex mixtures: Quality control and authenticity of instant coffee, *Journal of Agricultural and Food Chemistry*, **50:3098, 2002**.
- [116] BRIANDET, R.; KEMSLEY, K.; WILSON, R. H., Discrimination of arabica and robusta in instant coffee by fourier transform infrared spectroscopy and chemometrics, *Journal of Agricultural and Food Chemistry*, **44:170, 1996**.
- [117] LÓPEZ-MARTÍNEZ, L.; LÓPEZ-DE-ALBA, P. L.; GARCÍA-CAMPOS, R.; LEÓN-RODRÍGUEZ, L. M. D., Simultaneous determination of methylxantines in coffees and teas by UV-Vis spectrophotometry and partial least squares, *Analytica Chimica Acta*, **493:83, 2003**.
- [118] MOORES, R. G.; DOROTHY, L.; MCDERMOTT, L.; WOOD, T. R., Determination of chlorogenic acid in coffee, *Analytical Chemistry*, **20:620, 1948**.
- [119] ILLY, A.; VIANNI, R. *Espresso coffee: the chemistry of quality*. London: Academic Press, 1995.

Referências Bibliográficas

- [120] ANVISA (Agência Nacional de Vigilância Sanitária). DOU, 26 de Abril de 1999. Nº. 377. Disponível em: http://www.anvisa.gov.br/legis/portarias/377_99.htm, acesso em: 1 de outubro de 2008).
- [121] VITORINO, M. D.; FRANÇA, A. S.; OLIVEIRA, L. S.; BORGES, M. L. A., Metodologia de obtenção de extrato de café visando à dosagem de compostos não voláteis, *Revista Brasileira de Armazenamento*, **3:17, 2001**.
- [122] MARTINS, V. L.; ALMEIDA, L. F.; CASTRO, S. L.; GALVÃO, R. K. H.; ARAÚJO, M. C. U.; SILVA, E. C., A multiscale wavelet data treatment for reliable localization of inflection points for analytical purposes. *Journal of Chemical Information and Computer Sciences*, **43:1725, 2003**.
- [123] SIBCS (Sistema Brasileiro de Classificação de Solos). Rio de Janeiro: EMBRAPA, Centro Nacional de Pesquisa de Solo, 1999.
- [124] BALDWING, M.; KELLOGG, C. E.; THORP, J. Soil classification: In: ***Soils and men***. Washington, 1938.
- [125] THORP, J.; SIMITH, G. D., Higher categories of soil classification: order, suborder and great groups, *Soil Science*, **67:177, 1949**.
- [126] TITTONELL, P.; SHEPHERD, K. D.; VANLAUWE, B.; GILLER, K. E., Unravelling the effects of soil and crop management on maize productivity in smallholder agricultural systems of western Kenya – An application of classification and regression tree analysis, *Agriculture, Ecosystems & Environment*, **123:137, 2008**.
- [127] PHILLIPS, J. D.; MARION, D. A., Soil geomorphic classification, soil taxonomy, and effects on soil richness assessments, *Geoderma*, **141:89, 2007**.
- [128] ZAGATTO, E. A. G. *Análises químicas multielementares em sistemas FIA-ICP-GSAM e classificações dos solos do estado de São Paulo*, Campinas, Programa de Pós-Graduação em Química, UNICAMP, 1981, Tese de Doutorado.

Referências Bibliográficas

- [129] DEMATTÊ, J. A. M.; CAMPOS, R. C.; ALVES, M. C.; FIORIO, P. R.; NANNI, M. R., Visible-NIR reflectance: a new approach on soil evaluation, *Geoderma*, **121:95**, 2004.
- [130] MOUAZEN, A. M.; KAROUI, R.; BAERDEMAEKER, J.; RAMON, H. J., Classification of soil texture classes by using soil visual near infrared spectroscopy and factorial discriminant analysis techniques, *Near Infrared Spectroscopy*, **13:231**, 2005.
- [131] SANTOS JUNIOR, D.; TARELHO, L. V. G.; KRUG, F. J.; MILOR, D. M. B.; MARTIN NETO, L.; VIEIRA JUNIOR, N. D., Espectrometria de emissão óptica com plasma induzido por laser (LIBS) – Fundamentos, aplicações e perspectivas, *Revista Analytica*, **24:72**, 2006.
- [132] PASQUINI, C.; CORTEZ, J.; SILVA, L. M. C.; GONZAGA, F. B., laser induced breakdown spectroscopy, *Journal of Brazilian Chemical Society*, **18:463**, 2007.
- [133] CORTEZ, J. *Construção e avaliação de um instrumento para espectroscopia de emissão em plasma induzido por laser (LIBS): Aplicação em ligas metálicas*, Campinas, Programa de Pós-Graduação em Química, UNICAMP, 2007. Dissertação de Mestrado.
- [134] BOUSQUET, B.; SIRVEN, J.-B.; CANIONI, L., Towards quantitative laser-induced breakdown spectroscopy analysis of soil samples, *Spectrochimica Acta Part B*, **62:1582**, 2007.
- [135] CAI, C.; HARRINGTON, P. B., Different discrete wavelet transforms applied to denoising analytical data, *Journal of Chemical Information and Computer Sciences*, **38:1161**, 1998.
- [136] COELHO, C. J.; GALVÃO, R. K. H.; ARAÚJO, M. C. U.; PIMENTEL, M. F.; SILVA, E. C., A linear semi-infinite programming strategy for constructing optimal wavelet transforms in multivariate calibration problems, *Journal of Chemical Information and Computer Sciences*, **43:928**, 2003.

Referências Bibliográficas

[137] GALVÃO, R. K. H.; DANTAS FILHO, H. A.; MARTINS, M. N.; ARAÚJO, M. C. U.; PASQUINI, C., Sub-optimal wavelet denoising of coaveraged spectra employing statistics from individual scans, *Analytica Chimica Acta*, **581:159, 2007**.

[138] VETTERLI, M.; KOVACEVIC, J. **Wavelets and Subband Coding**. New Jersey: Prentice-Hall, 1995.

[139] JOSÉ, G. E., *Uso do método MDL para filtragem de ruído instrumental empregando a transformada wavelet*, João Pessoa, Programa de Pós-Graduação em Química, UFPB, 2008. Dissertação de Mestrado.

[140] SANTOS, R. N. F.; GALVÃO, R. K. H.; ARAÚJO, M. C. U.; SILVA, E. C., Improvement of prediction ability of PLS models employing the wavelet packet transform: A case study concerning FT-IR determination of gasoline parameters, *Talanta*, **71:1136, 2007**.

[141] GAO, L.; REN, S., Simultaneous spectrophotometric determination of four metals by two kinds of partial least squares methods, *Spectrochimica Acta Part A*, **61:3013, 2005**.

[142] NIST (National Institute of Standards and Technology). Disponível em: http://physics.nist.gov/PhysRefData/ASD/lines_form.html. Acessado em 1 de outubro de 2008.

Anexos

O código-fonte do programa SPA-LDA, escrito em linguagem MATLAB 6.5, é apresentado a seguir:

```
function [I,R,Lopt] = SPA_LDA(Train,Group_Train,Val,Group_Val,Test,Group_Test,N1,N2)

% SPA-LDA - Seleção de Variáveis para Classificação baseada no Algoritmo das Projeções
% Sucessivas (SPA) empregando a built-in function qr do Matlab
%
% [I,R,Lopt] = SPA-LDA(Train,Group_Train,Val,Group_Val,Test,Group_Test,N1,N2)
%
% VARIÁVEIS DE ENTRADA
%
% *** Matrizes correspondentes as respostas instrumentais ***
%     Train  -> Conjunto de treinamento;
%     Val    -> Conjunto de validacao;
%     Test   -> Conjunto externo para teste;
%
% *** Vetores correspondentes aos indices de classes ***
%
%     Group_Train -> Indice de classes para amostras de treinamento;
%     Group_Val   -> Indice de classes para amostras de validação;
%     Group_Test  -> Indice de classes para amostras de teste.
%
%*** Numero mínimo e Maximo de variáveis a serem selecionadas ***
%     N1 -> Numero mínimo de variáveis;
%     N2 -> Numero Maximo de variáveis.

% VARIÁVEIS DE SAIDA
%
%     I --> Resultado da seleção (melhor cadeia de variáveis)
%     R --> Risco médio G de uma classificação incorreta pela LDA em função do número de
% variáveis usadas;
%     Lopt --> Matriz contendo as cadeias de variáveis associadas a R
%
% Autor: Roberto Kawakami Harrop Galvão - ITA/IEES
% Revisão: 10 de outubro de 2008 p/Marcio J. C. Pontes - UFPB.

Ntrain = size(Train,1);
Nlambdas = size(Train,2);

if N2 > Ntrain
    error('O APS nao pode selecionar mais variaveis do que amostras de treinamento');
end

% Indices de classes
C = max(Group_Train); % Numero de classes do problema
Xpooled = []; % Matriz a ser usad para o Calculo da Matriz de Covariancia Conjunta (Pooled)
for i=1:C
    index{i} = find(Group_Train==i); % Objetos de treinamento pertencentes a i-esima classe
    Traini = Train(index{i},:);
    media{i} = mean(Traini);
    Trainic = Traini - repmat(media{i},size(Traini,1),1);
    Xpooled = [Xpooled;Trainic]; % Xpooled contem os objetos centrados nas medias das respectivas
classes
end
```

Anexos

```
clc
disp('- Seleção de variáveis para classificação via APS -')
disp(' ')
disp('Escolha uma opção e tecle <ENTER>')
disp('1 - Gerar cadeias de variáveis via APS')
disp('2 - Carregar cadeias de variáveis previamente geradas')
opcao = input('Opção: ');

if (opcao == 1) % Gerar

    disp(' ')
    disp('O APS sera utilizado para gerar cadeias de variáveis.')
    disp(' ')
    disp('Entre com o nome de um arquivo para salvar as cadeias')
    filename = input('(entre apóstrofos e sem extensão): ');

    % No MATLAB, a inicializacao previa de vetores torna a execucao mais rápida
    L = zeros(N2,Nlambdas);
    Xcaln = zeros(Ntrain,Nlambdas);
    X = zeros(Ntrain,Nlambdas);

    % Auto-Escalonamento das colunas de Xpooled

    opcao = input('Deseja fazer auto-escalonamento para fins de selecao ? (Sim: 1 - Nao: 0)');
    if (opcao == 1)
        disp('Sera feito auto-escalonamento!')
        for i=1:Nlambdas
            x = Xpooled(:,i);
            Xcaln(:,i)=x/std(x);
        end
    else
        disp('Nao sera feito auto-escalonamento!')
        Xcaln = Xpooled;
    end

    espera = input('Tecla algo para continuar...');

    % Aplicacao do APS
    clc

    normas = sum(Xcaln.^2); % Norma ao quadrado de cada coluna de Xcaln
    normamax = max(normas); % Maior norma

    t0 = clock; % Comando usado para estimar quanto tempo levará a otimização
    for i=1:Nlambdas

        X = Xcaln;
        X(:,i) = X(:,i)*2*normamax/normas(i); % forca o qr a comecar pela i-esima coluna de Xcaln
        [dummy1,dummy2,ordem] = qr(X,0); % ordem contem os comprimentos de onda selecionados a
partir do i-esimo
        L(:,i) = ordem(1:N2); % Guarda apenas N2 comprimentos de onda

        % Apresenta na Tela uma estimativa do tempo restante para o fim da otimização
        if (rem(i,10) == 0)
            tempo = etime(clock,t0);
            home
            disp(['Aplicando SPA. Tempo restante = ',num2str(round((Nlambdas-
i)*tempo/10)), 's',blanks(10)]);
            t0 = clock;
        end
    end
end
```

Anexos

```
end

save(filename,'L')

else % Carregar

    filename = input('Entre com o nome do arquivo (entre apóstrofos e sem extensão): ');
    load(filename)

end

% Determinação do custo de classificação p/ todas as cadeias de variáveis
disp('Determinando o custo de classificação para determinar a melhor cadeia de variáveis ...')
disp(' ')

R = zeros(1,N2);
Lopt = zeros(N2,N2);
custo = zeros(N2,Nlambdas);

N2
for N = N1:N2 % Para a cadeia de comprimento N
    N
    for i = 1:Nlambdas % partindo da variável i
        lambdas = L(1:N,i); % Variáveis da cadeia
        % Respostas instrumentais nas variáveis da cadeia
        S = cov(Xpooled(:,lambdas),1);
        invS = inv(S);
        Val2 = Val(:,lambdas);
        custoaux = 0;
        for j=1:size(Val,1) % Para cada objeto de Val
            x = Val2(j,:);
            grupox = Group_Val(j); % Classificação correta do objeto
            for k=1:C % Para cada classe
                mu = media{k};
                mu = mu(lambdas);
                r(k) = (x - mu)*invS*(x - mu)';
            end
            num = r(grupox); % Distancia de Mahalanobis a classe correta
            remaining = setdiff([1:C],grupox); % Demais classes
            den = min(r(remaining)); % Menor Distancia de Mahalanobis as classes restantes
            custoaux = custoaux + num/den;
        end
        % Calculo do custo associado a cadeia
        custo(N,i)= custoaux/size(Val,1); % Custo medio
    end
    [R(N) imin] = min(custo(N,:));
    Lopt(1:N,N)=L(1:N,imin);
end

[Rbest,Nbest] = min(R(N1:N2));
Nbest = Nbest+N1-1;
custoopt = custo(Nbest,:);
l = (Lopt(1:Nbest,Nbest));

disp(['Menor custo obtido no conjunto de validação: ' num2str(Rbest)])
disp(['Número ideal de variáveis: ' num2str(Nbest)])

figure,plot([N1:N2],R(N1:N2)),grid
axis([N1 N2 -inf inf])
```

Anexos

```
xlabel('Número de variaveis')  
ylabel('Custo - Validação')
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%  
% Apresentação dos resultados de classificação usando LDA  
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
Train2 = Train(:,l);  
Val2 = Val(:,l);  
Test2 = Test(:,l);
```

```
[class_Val,errors_Val] = multilda(Val2,Train2,Group_Train,Group_Val);  
[class_Test,errors_Test] = multilda(Test2,Train2,Group_Train,Group_Test);
```

```
disp(['Erros para o conjunto de validação: ' num2str(errors_Val)])  
disp(['Erros para o conjunto de teste: ' num2str(errors_Test)])
```

Programas Auxiliares do SPA-LDA

- multilda.m
- mahal2.m

1) multilda.m

```
function [class,errors] = multilda(sample,training,group,Group_Test)
%MULTILDA Multi-Class Linear discriminant analysis.
% [class,errors] = multilda(sample,training,group,Group_Test) classifies each row
% of the data in SAMPLE into one of the values of the vector
% GROUP. GROUP contains integers from one to the number of
% groups in the training set, which is the matrix, TRAINING.
%
% SAMPLE and TRAINING must have the same number of columns.
% TRAINING and GROUP must have the same number of rows.
% CLASS is a vector with the same number of rows as SAMPLE.
%
% The biased Maximum-Likelihood estimate for the covariance matrix is employed: cov(X,1)
%
% B.A. Jones 2-05-95
% Copyright 1993-2000 The MathWorks, Inc.
% $Revision: 2.9 $ $Date: 2000/05/26 17:28:36 $

[gr,gc] = size(group);
if min(gr,gc) ~= 1
    error('Requires the third argument to be a vector.');
```

```
end

if gc ~= 1,
    group = group(:);
    gr = gc;
end

if any(group - round(group)) | any(group < 1)
    error('The third input argument must be positive integers.');
```

```
end
maxg = max(group);

[tr,tc] = size(training);
```

Anexos

```
if tr ~= gr,
    error('The number of rows in the second and third input arguments must match.');
```

```
end
```

```
[sr,sc] = size(sample);
if sc ~= tc
    error('The number of columns in the first and second input arguments must match.');
```

```
end
```

```
d = zeros(sr,maxg);
% Calculo de S
grouptemp = [];
for k = 1:maxg
    groupk = training(find(group == k),:);
    groupkc = groupk - repmat(mean(groupk),size(groupk,1),1);
    grouptemp = [grouptemp;groupkc];
end
S = cov(grouptemp,1);
```

```
for k = 1:maxg
    groupk = training(find(group == k),:);
    d(:,k) = mahal2(sample,mean(groupk),S); % S eh a matriz de covariancia conjunta de todas as
classes
end
[tmp, class] = min(d');
class = class';
```

```
if nargin == 4 % If the correct classification of the test set is provided, the number of errors is
calculated
    e = class - Group_Test;
    errors = length(find(e~=0));
else
    errors = [];
end
```

Programas Auxiliares do SPA-LDA

2) mahal2.m

```
function d = mahal2(X,mi,S);  
%  
d = zeros(size(X,1),1);  
for i = 1:size(X,1)  
    d(i) = (X(i,:)-mi)*inv(S)*(X(i,:)-mi)';  
end
```