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

# TESE DE DOUTORADO

## Um novo método para transferência de modelos de calibração NIR e uma nova estratégia para classificação de sementes de algodão usando imagem hiperespectral NIR

**Sófacles Figueredo Carreiro Soares**

**João Pessoa – PB - Brasil**
**Junho/2016**

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

# TESE DE DOUTORADO

## Um novo método para transferência de modelos de calibração NIR e uma nova estratégia para classificação de sementes de algodão usando imagem hiperespectral NIR

**Sófacles Figueredo Carreiro Soares***

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

**Orientador: Prof. Dr. Mário César Ugulino de Araújo**

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

***Bolsista CNPq e CAPES**

**João Pessoa – PB - Brasil**
**Junho/2016**

S676u Soares, Sófacles Figueredo Carreiro.
Um novo método para transferência de modelos de calibração NIR e uma nova estratégia para classificação de sementes de algodão usando imagem hiperespectral NIR / Sófacles Figueredo Carreiro Soares.- João Pessoa, 2016.
121f. : il.
Orientadores: Mário César Ugulino de Araújo, Roberto Kawakami Harrop Galvão
Tese (Doutorado) - UFPB/CCEN
1. Química analítica. 2. Transferência de calibração.
3. Regressão robusta. 4. Correção univariada.
5. Espectrometria NIR. 6. Sementes de algodão.

UFPB/BC CDU: 543(043)

# Um novo método para transferência de modelos de calibração NIR e uma nova estratégia para classificação de sementes de algodão usando imagem hiperespectral NIR

Tese de Doutorado apresentada pelo aluno Sófacles Figueredo Carreiro Soares e aprovada pela banca examinadora em 20 de junho de 2016.

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

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

Prof. Dr. Everaldo Paulo de Medeiros
Examinador

Profa. Dra. Patricia Kaori Soares
Examinadora

Prof. Dr. Edvan Cirino da Silva
Examinador

Prof. Dr. Sherlan Guimarães Lemos
Examinador

## AGRADECIMENTOS


- Agradeço primeiramente a Deus, que me guiou pelo caminho da persistência e me deu discernimento para escolher a direção a ser seguida nos momentos mais difíceis;
- Ao meu pai Pedro, minha mãe Neide, e meus irmãos Pietros e Perla, por todo amor, confiança, apoio e compreensão;
- À Arnayra Silva por todo amor, apoio e paciência;
- Ao Professor Mário César Ugulino de Araújo pela orientação, incentivo e confiança;
- Ao Professor Roberto Kawakami Harrop Galvão pela orientação e ensinamentos;
- À EMBRAPA algodão pela cessão das amostras e equipamentos. Em especial Dr. Everaldo Paulo de Medeiros.
- Ao Professor Célio Pasquini por disponibilizar a estação de imagens hiperespectrais e sempre estar disposto a nos ajudar;
- À Professora Maria Fernanda Pimentel pela cessão dos conjuntos de dados de gasolinas;
- Aos amigos Stefani Yuri, Marcelo Batista e Inakã Barreto pela amizade, companheirismo e pelas constantes contribuições e discussões científicas;
- A todos os demais membros do LAQA, pela amizade, convivência e pelos bons momentos;
- Aos Professores Edvan Cirino da Silva e Sherlan Guimarães Lemos pelas excelentes sugestões na banca de qualificação;
- À Universidade Federal da Paraíba;
- Ao CNPq e CAPES pelo apoio financeiro.

**SUMÁRIO**

## LISTA DE FIGURAS

# LISTA DE SIGLAS E ABREVIATURAS

| | |
|---|---|
| DA | Discriminant Analysis (análise discriminante) |
| DS | Direct Standardization (padronização direta) |
| GA | Genetic Algorithm (algoritmo genético) |
| HSI | Hyperspectral Images (imagens hiperespectrais) |
| LDA | Linear Discriminant Analysis (análise discriminante linear) |
| LS | Least Squares (mínimos quadrados) |
| MLR | Multiple Linear Regression (regressão linear multipla) |
| NIR | Near-Infrared Spectroscopy (infravermelho próximo) |
| $N_{trans}$ | Número de amostras de transferência |
| PCA | Principal Component Analysis (análise de componentes principais) |
| PCR | Principal Component Regression (regressão em componentes principais) |
| PDS | Piecewise Direct Standardization (padronização direta por partes) |
| PLS | Partial least Squares (regressão em mínimos quadrados parciais) |
| RMSEP | Root Mean Squares Error of Prediction (raiz do erro quadrático médio de predição) |
| RON | Research Octane Number |
| RR | Regressão *ridge* |
| SM | Specific Mass (massa específica) |
| SPA | Successive Projection Algorithm (algoritmo das projeções sucessivas) |
| SQR | Soma quadrática residual |
| SVD | Singular Value Decomposition (decomposição em valores singulares) |
| TCC | Taxa de Classificação Correta |
| UVC | Univariate Correction (correção univariada) |

# LISTA DE TABELAS

## RESUMO

Este trabalho envolve o desenvolvimento de dois estudos, que são apresentados nos capítulos 2 e 3. No primeiro, um novo método para realizar a transferência de calibração foi concebido. Este método foi desenvolvido para fazer uso de variáveis isoladas em vez de usar todo o espectro ou janelas espectrais. Para realizar essa tarefa, um procedimento univariado é inicialmente usado para corrigir os espectros registrados no equipamento secundário, dado um conjunto de amostras de transferência. Uma técnica de regressão robusta é então usada para obter um modelo com pequena sensibilidade em relação aos resíduos da correção univariada. O novo método é então empregado em dois estudos de caso envolvendo análise espectrométrica NIR, em que foram determinados os parâmetros massa específica, RON (Research Octane Number) e teor de naftênicos em gasolina e os teores de água e óleo em amostras de milho. Os resultados do novo método foram melhores do que os obtidos usando o método PDS. No segundo, uma nova estratégia para classificação de sementes de algodão usando imagens hiperespectrais no NIR foi desenvolvido. Inicialmente as amostras de sementes de algodão foram registradas em uma estação de imagem HSI-NIR e em um equipamento NIR convencional. Após isso, as imagens foram segmentadas e os espectros médios de cada semente foram extraídos. Os modelos de classificação SPA-LDA e PLS-DA baseados nos espectros médios foram construídos para os dois conjuntos de dados. Os resultados SPA-LDA e PLS-DA para os modelos demonstraram que a classificação com os dados HSI-NIR foi alcançada com maior exatidão quando comparada aos modelos obtidos usando o NIR-convencional.

Palavras-chave: Transferência de calibração, regressão robusta, correção univariada, espectrometria NIR, Classificação, Sementes de algodão, Imagens Hiperespectrais.

## ABSTRACT

This work involves the development of two studies that are presented in chapters 2 and 3. At first, a new method to perform the calibration transfer was designed. This method was developed to make use of separate variables instead of using the full spectrum or spectral windows. To accomplish this task a univariate procedure is initially used to correct the spectra recorded in the secondary equipment, given a set of transfer samples. A robust regression technique is then used to obtain a model with small sensitivity with respect to the univariate correction. The proposed method is employed in two case studies involving near infrared spectrometric determination of specific mass, research octane number and naphtenes in gasoline, and moisture and oil in corn. In both cases, better calibration transfer results were obtained in comparison with piecewise direct standardization (PDS). In the second, a new strategy for cotton seed classification using near infrared (NIR) hyperspectral images (HSI) was developed. Initially the cotton seeds samples were recorded on a station HSI image-NIR and a conventional spectrometer NIR. Thereon, the images were segmented and the mean spectrum of each seed was extract. Classification models SPA-LDA e PLS-DA based on the mean spectral were developed for two data sets. The results for models SPA-LDA and PLS-DA showed that the classification with HSI-NIR data set has been achieved with greater accuracy when compared to models for the NIR-conventional data set.

Keywords: Calibration transfer, robust regression, univariate correction, NIR spectrometry, classification, cotton seeds, hyperspectral image.

# CAPÍTULO 1: APRESENTAÇÃO

Os métodos analíticos modernos normalmente utilizam técnicas instrumentais capazes de realizar medidas em amostras sólidas, líquidas ou gasosas, a partir de pouco tratamento químico, o que consequentemente, reduz o tempo de análise e a quantidade de resíduos gerados. Entretanto, os instrumentos empregados são formados por uma grande quantidade de canais analíticos no sistema de detecção, produzindo assim conjuntos de dados com um considerável número de variáveis. Exemplos incluem, Imagens Hiperespectrais (AMIGO; MARTÍ; GOWEN, 2013), Espectrometria de Emissão em Plasma Induzido por Laser (LIBS: *Laser-Induced Breakdown Spectroscopy*) (PASQUINI et al., 2007) e Infravermelho Próximo (NIR: *Near-Infrared Spectroscopy*) (PASQUINI, 2003), que permitem medir ao longo de um grande número de comprimentos de onda.

A informação produzida por esses instrumentos podem ser armazenadas em uma matriz, em que as linhas correspondem às medidas realizadas para cada amostra e a resposta em cada canal analítico é disposta ao longo das colunas. Para estas matrizes, somente a estatística univariada não é ser suficiente, sendo necessário o uso de ferramentas capazes de explorar adequadamente toda essa informação. Dessa forma, a quimiometria passou a ser usada para a obtenção de respostas cada vez mais rápidas na análise de dados químicos (FORINA; LANTERI; CASALE, 2007; HOPKE, 2003).

Com o avanço das metodologias analíticas, cada vez mais se exige dos métodos quimiométricos uma abordagem matemática e estatística compatível com o grau de resposta exigido. Assim sendo, o interesse em técnicas instrumentais mais precisas aliadas a novos métodos quimiometricos tem crescido (KUMAR et al., 2014), de modo que as informações obtidas lançam uma nova luz sobre a avaliação dos dados.

Com base no exposto, neste trabalho busca-se desenvolver um novo método para a transferência de calibração a partir da combinação da correção univariada e da regressão robusta, especificamente para trabalhar com variáveis isoladas, ao invés de janelas espectrais,

permitindo assim transferir modelos multivariados para equipamentos dedicados com maior robustez. Em uma segunda etapa, propõe-se o desenvolvimento de uma metodologia analítica, rápida, não destrutiva e não invasiva baseada no uso da espectroscopia de imagem hiperespectral no infravermelho próximo aliadas aos métodos de reconhecimento de padrões para a classificação de sementes individuais de algodão com respeito à variedade.

# CAPÍTULO 2: UM NOVO MÉTODO PARA TRANSFERÊNCIA DE CALIBRAÇÃO

## 2.1 Introdução

O desenvolvimento de modelos de calibração multivariada envolve diversas etapas, tipicamente inclui a coleta de amostras e registro dos sinais analíticos multivariados, seguidos pela construção e validação do modelo. Todos estes estágios são importantes para o modelo alcançar uma boa predição quando empregado na análise de novas amostras. Em particular, seria desejável eliminar ou minimizar fontes de variabilidade nos dados que não são relacionadas com as propriedades analíticas de interesse. No entanto, existem casos em que mudanças nas condições analíticas ocorrem após a calibração ter sido realizada, provocando efeitos adversos na habilidade de predição do modelo (BOUVERESSE; MASSART, 1996; FEUDALE et al., 2002). Tais mudanças podem estar associadas a características físicas/químicas das amostras (viscosidade, granulometria, textura da superfície e presença de espécies interferente) (BROWN, 2009), condições ambientais (temperatura e umidade, por exemplo), bem como em função da resposta do próprio instrumento.

Problemas associados à resposta do instrumento tipicamente surgem por causa dos efeitos do envelhecimento, deterioração de partes específicas ou intervenções de manutenção. Dificuldades também podem surgir se o instrumento utilizado para a aquisição de dados não é o mesmo utilizado para a construção do modelo de calibração (FEUDALE et al., 2002).

Estas alterações podem ser removidas pela recalibração do modelo usando um conjunto de dados adquirido sob as novas condições de análise (BROWN, 2009). No entanto, este procedimento pode ser caro, trabalhoso e demorado, pois todo o processo de construção do modelo teria de ser repetido. Alternativamente, os diferentes métodos têm sido desenvolvidos para compensar as alterações nas condições experimentais, sem a necessidade de um modelo completo de recalibração.

Entre os métodos mais recorrentes na literatura podemos destacar os métodos de padronização direta e por partes (WANG; VELTKAMP; KOWALSKI, 1991), correção de linha de

base (BEEBE; PELL; SEASHOLTZ, 1998), correção multiplicativa de sinal (NÆS, 2002), filtragem de resposta ao impulso finita (BLANK et al., 1996), correção de sinal ortogonal (FERNÁNDEZ PIERNA et al., 2001), decomposições wavelet (YOON; LEE; HAN, 2002), transferência por projeções ortogonais (ANDREW; FEARN, 2004), entre outros.

No entanto nos últimos anos têm sido propostos novos métodos para realizar transferência de calibração. Nesse contexto, Binfeng & Haibo (2015) propuseram uma nova estratégia de transferência de modelos de regressão em máquina de vetores de suporte (SVR: *Support Vector Regression*) com base em um método de transferência de calibração por aprendizagem. Neste método, a diferença entre os vetores de regressão obtidos pelo SVR na condição inicial e após o modelo sofrer interferência é penalizada pela norma $L_2$, que impedem o novo modelo SVR seja afetado demasiadamente pelas mudanças nas condições experimentais.

Zheng et al. (2014) aplicaram a análise de correlação canônica (CCA: *Canonical Correlation Analysis*) para transferir os componentes informativos extraídos de um conjunto de dados espectral. Tal estratégia é capaz de reduzir a interferência do ruído de fundo e propriedades não preditas. Este método emprega o PLS para extrair os componentes informativos e, em seguida, corrige os componentes informativos com base na CCA. O desempenho deste algoritmo foi testado utilizando três conjuntos de dados e os resultados mostraram que este método pode reduzir significativamente os erros de predição.

Chen et al. (2011) desenvolveram um novo método que realiza a transferência de calibração por meio de uma correção do erro de predição de forma sistemática. Para tal uma decomposição em valores singulares e o resíduo do modelo de calibração são usados. O desempenho do método foi testado em dois conjuntos de dados NIR (um com mudanças nas respostas instrumentais, o outro com variações nas condições experimentais) e os resultados foram satisfatórios quando comparados a outros métodos.

Cooper; Larkin & Abdelkader (2011) realizaram a transferência de calibração usando padrões virtuais e correção por *slope-bias*. No método apresentado, os padrões virtuais (amostras de transferência) são construídos de forma virtual pela combinação de espectros de solventes puros registrados no equipamento primário e no secundário.

Com objetivo de eliminar as diferenças espectrais provocadas pela mudança nas condições de medidas, Du et al. (2011) desenvolveram um novo método baseado na transformação do espaço espectral por meio de uma decomposição em valores singulares dos espectros do subconjunto de amostras de padronização medidas em dois equipamentos e em dois conjuntos com condições experimentais distintas.

Peng et al. (2011) recorreram a uma estratégia baseada em uma série de regressões gerando mudanças de base vetoriais de forma a realizar a transferência de calibração com dados NIR. Os resultados experimentais mostram um bom desempenho do método baseado em regressões, principalmente, quando o número de amostras do subconjunto de padronização aumenta.

Martins; Galvão & Pimentel (2010) propuseram uma nova técnica para transferência de calibração, que combina o Algoritmo das Projeções Sucessivas (SPA: *Successive Projections Algorithm*) para seleção de variáveis robustas com a técnica de sub-amostragem e agregação de modelos conhecida como *subagging*. A técnica proposta tem por objetivo construir modelos de Regressão Linear Múltipla que sejam robustos a diferenças na resposta instrumental de dois espectrômetros (primário e secundário). A eficiência da técnica é demonstrada em dois estudos de casos e nesse caso verificou-se que o uso de *subagging* resultou em uma redução mais sistemática do erro de predição com a inclusão progressiva de amostras de transferência.

Fan et al. (2008) propuseram um método baseado na análise de correlação canônica (CCA) para a transferência de modelos de calibração. Um estudo comparativo entre o novo método e a padronização direta por partes (PDS) foi realizada. Nesse caso, os resultados de transferência obtidos com base no CCA foram melhores do que os obtidos por PDS.

Nos trabalhos acima apresentados percebe-se que no âmbito das técnicas de espectrometria, a transferência de calibração foi baseada em transformações matemáticas que envolvem o espectro total ou janelas de variáveis. Para este fim, o método PDS é muitas vezes usado como referência em estudos comparativos de transferência de calibração (FAN et al., 2008; WALCZAK; BOUVERESSE; MASSART, 1997). No PDS, o modelo de padronização relaciona cada variável do espectro primário com uma janela de variáveis do espectro secundário. Dessa forma, pode-se argumentar que métodos como o PDS não seriam adequados para uso em aplicações com instrumentos que monitoram apenas um pequeno conjunto de variáveis espectrais usando filtros (HAUSER; RUPASINGHE; CATES, 1995; MALINEN et al., 1998) e/ou *light emitting diodes* (LEDS) (CAPITÁN-VALLVEY; PALMA, 2011; DE LIMA, 2012; GAIÃO et al., 2008; GIOVENZANA et al., 2015). Em tal caso, não seria possível obter um modelo de padronização com base em janelas espectrais, porque as medidas estariam relacionadas com os comprimentos de onda isolados.

Honorato et al. (2005) propuseram uma estratégia para a seleção de variáveis robustas a diferenças instrumentais. Este método é uma adaptação do SPA, que minimiza uma função custo levando em conta a habilidade preditiva e a robustez com respeito às diferenças instrumentais, na escolha das variáveis. Embora este método trabalhe com variáveis isoladas, ele não pode ser usado para transferir modelos no qual as variáveis já foram escolhidas inicialmente.

Diante do exposto, nessa tese é proposto um novo método para transferência de calibração que emprega um procedimento univariado inicialmente para corrigir as medições

espectrais do instrumento secundário, dado um conjunto de amostras de transferência (FEUDALE et al., 2002). Em seguida uma técnica de regressão robusta é utilizada para construir um novo modelo de calibração multivariada com baixa sensibilidade em relação aos resíduos do processo de correção univariada. A partir da necessidade de avaliar o novo método de transferência de calibração, recorreu-se ao Algoritmo Genético (GA: *Genetic Algorithm*) para selecionar diferentes subconjuntos de variáveis.

## 2.2 Objetivos

### 2.2.1 Objetivo geral

Desenvolver um novo método para a transferência de calibração a partir da combinação da correção univariada e da regressão robusta, especificamente para trabalhar com variáveis isoladas, ao invés de janelas espectrais, permitindo assim transferir modelos multivariados para equipamentos dedicados com maior robustez.

### 2.2.2 Objetivos específicos

- Obter um modelo de calibração multivariada MLR, usando o algoritmo genético para a seleção de variáveis, e calcular os erros obtidos quando os espectros do equipamento secundário são usados diretamente no modelo;
- Construir o modelo de correção univariado para corrigir as medidas espectrais do instrumento secundário usando as amostras de transferência previamente selecionadas;
- Utilizar a técnica de regressão robusta para construir um novo modelo de calibração multivariada com baixa sensibilidade em relação aos resíduos do processo de correção univariada;
- Avaliar a influência dos parâmetros do algoritmo genético nos resultados obtidos com o novo método;
- Construir um modelo PLS com os espectros do equipamento primário e realizar a transferência de calibração usando o método PDS para fins de comparação com o novo método;
- Construir modelos de regressão *ridge* para a comparação com o novo método de transferência de calibração;

## 2.3 Fundamentação teórica

### 2.3.1 Calibração multivariada

A medida de uma única variável, em muitos casos, não é capaz de fornecer uma boa relação com a resposta, e consequentemente uma boa predição, sem a minimização das influências que o sinal pode ter de outras espécies (NÆS, 2002). Dessa forma, um modelo univariado só pode fornecer resultados precisos, se o sinal medido for proveniente apenas do analito em estudo ou se o seu sinal líquido puder ser obtido matematicamente (BRO, 2003).

O uso de dados multivariados é capaz de reduzir a quantidade de ruído, melhorando a precisão e fornecendo a capacidade de determinação simultânea de muitas propriedades de uma amostra através de um único sinal multivariado (BRO, 2003; KUMAR et al., 2014). Diante dessas vantagens, a calibração multivariada tornou-se uma ferramenta indispensável para a determinação quantitativa em química analítica. Na literatura são reportados diversos métodos usados para a construção de modelos de calibração multivariada, em que muitos deles essencialmente diferem pela obtenção da inversa generalizada para estimar os coeficientes de regressão. Nas seções que seguem, serão apresentadas as técnicas de calibração multivariada usadas nessa tese.

#### 2.3.1.1 Regressão linear múltipla

A regressão linear múltipla pode ser caracterizada como uma técnica para resolver uma série de equações simultâneas, em sistemas multicomponentes (GOODARZI et al., 2015). Dessa forma, inicialmente, vamos assumir que a propriedade de interesse $y$ deve estar relacionada com as $p$ variáveis espectrais $x_1$, $x_2$, ..., $x_p$ por um modelo empírico linear (KUMAR et al., 2014) da forma $y = \mathbf{xb} + e$, onde $\mathbf{x} = [x_1\ x_2\ \dots\ x_p]$, $\mathbf{b} = [b_1\ b_2\ \dots\ b_p]^T$ é o vetor de coeficientes a ser determinado, e $e$ denota o resíduo do modelo.

Se uma matriz **X** ($N \times p$) de respostas instrumentais e um vetor **y** ($N \times 1$) com os valores da propriedade de interesse estão disponíveis para $N$ amostras, a estimativa para o vetor de coeficientes $\mathbf{b} = [b_1\ b_2\ \ldots\ b_p]^T$ usando o método dos mínimos quadrados (LS) é a que minimiza a seguinte função de custo:

$$J(\mathbf{b}) = \| \mathbf{Xb} - \mathbf{y} \|^2 \quad \textbf{(1)}$$

que corresponde ao quadrado da norma-2 dos resíduos do modelo. A solução para essa função custo é dada por:

$$\mathbf{b_{LS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \quad \textbf{(2)}$$

Todavia, a obtenção da matriz $\mathbf{b_{LS}}$ pela expressão acima está sujeita as seguintes restrições:

- A obtenção de um sistema indeterminado quando $p > N$;
- Alta colinearidade que ocorre ao longo das colunas de **X**. Se isso ocorre, $\mathbf{X^TX}$ está mal condicionada e a sua inversa não pode ser calculada com precisão.

Para contornar tais problemas no uso do MLR, em sinais multivariados, recorre-se aos métodos de seleção de variáveis.

#### 2.3.1.1.1 O algoritmo genético para seleção de variáveis

O algoritmo genético (GA: *Genetic Algorithm*) proposto por John H. Holland é uma técnica que simula matematicamente os mecanismos de seleção natural e a teoria da evolução das espécies de Charles R. Darwin (LEARDI, 2001; LEARDI; SEASHOLTZ; PELL, 2002).

A implementação desse algoritmo para seleção de variáveis é feita pela representação de cada variável como um gene (GALVÃO; ARAÚJO, 2009). Genes que tem valor 1 indicam que a variável será incluída na fase de avaliação, enquanto genes com valor 0 indicam que as variáveis não serão avaliadas. O conjunto de genes binários, que representam os indivíduos, formam os cromossomos. Na **Figura 2.1** é apresentado um esquema de codificação das variáveis em uma matriz quadrada de dados com cinco variáveis e amostras. Nessa figura, três e duas variáveis são incluídas, respectivamente, no primeiro e segundo cromossomos.

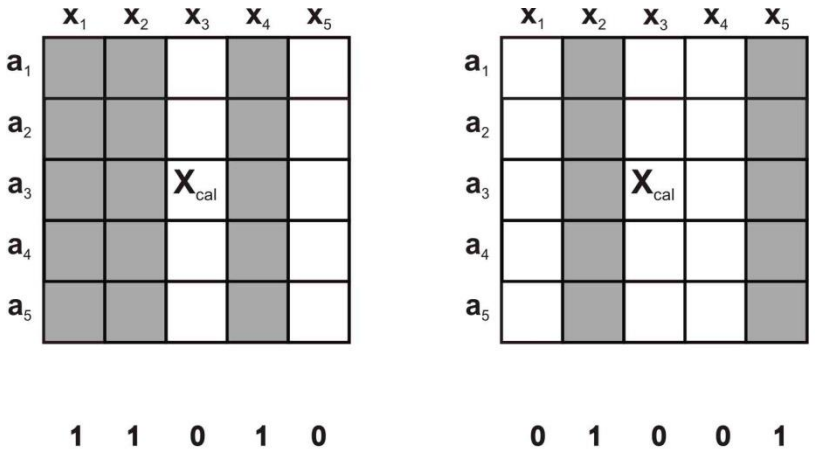



**Figura 2.1.** Codificação binária usada na seleção de variáveis.

No início do processo, o GA utiliza um gerador aleatório para criar uma população inicial de cromossomos, evitando-se a influência tendenciosa na construção dessa população, caso o analista tenha uma prévia informação de alguma variável, ele pode incluí-la na população inicial (GALVÃO; ARAÚJO, 2009).

A avaliação dos cromossomos é feita com base na aptidão, que é o parâmetro que indicará a habilidade de um indivíduo sobreviver. Matematicamente, quanto maior a aptidão de um indivíduo, melhor a resposta produzida (menor erro). O cálculo da aptidão pode ser realizado construindo modelos MLR ou PLS, baseados nos comprimentos de onda indicados em cada cromossomo. Para isso, o valor da aptidão é calculado como sendo o inverso do

PRESS (*Predictive Residual Sum of Squares*) obtido no conjunto de validação. O subconjunto de variáveis que produzir o menor PRESS (e assim uma maior aptidão) é então adotado como o resultado do GA.

A seleção dos indivíduos que gerarão descendentes é realizada de modo que os mais aptos tenham maior probabilidade de serem os escolhidos para a reprodução. O procedimento usado para esse fim é conhecido como método da roleta. Neste método a probabilidade de cada indivíduo é dada pela **Equação (3)**.

$$P(i) = \frac{Aptidão(i)}{\sum_{i=1}^{N} Aptidão(i)} \quad \textbf{(3)}$$

Uma nova população é formada cruzando pares de cromossomos aleatoriamente e gerando filhos que possuem material genético dos pais. O cruzamento pode ser realizado por ruptura, que pode ocorrer em mais de um ponto, seguida de recombinação.

Em uma pequena parcela da população é promovida a mutação (que são alterações no código genético) de modo a obter uma maior variabilidade genética. Na representação binária, a mutação é realizada pela troca de 1 por 0 ou vice-versa em um dos genes do cromossomo. O processo de cruzamento e mutação é ilustrado na **Figura 2.2**

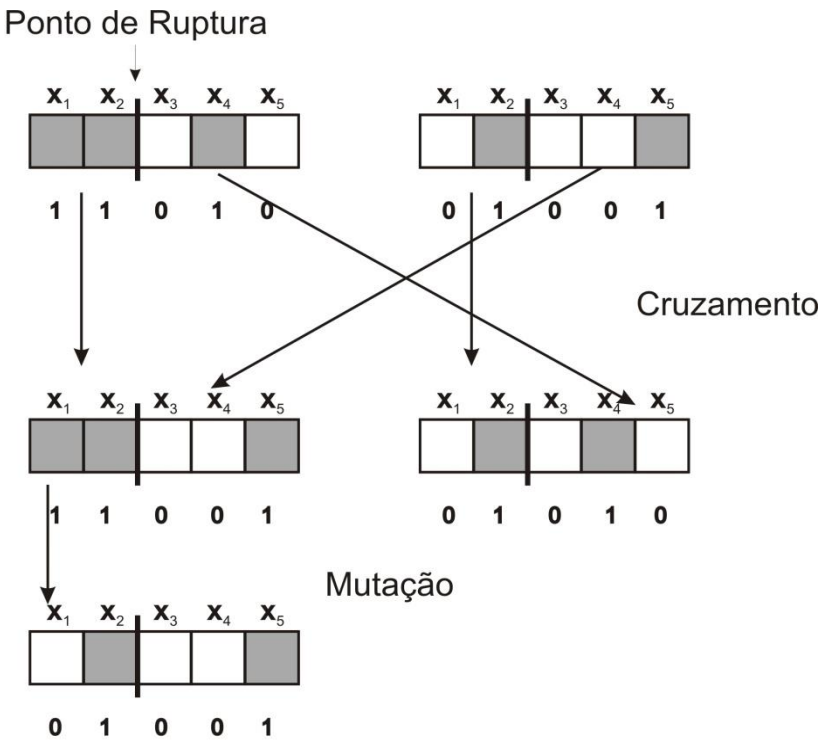



**Figura 2.2.** Esquema de cruzamento e de mutação no algoritmo genético.

### 2.3.1.2 Regressão *ridge*

A regressão *ridge* (RR) foi introduzida em 1970 (HOERL; KENNARD, 1970b) com a finalidade de reduzir os efeitos da alta colinearidade ao longo das variáveis independentes. Esse método consiste em promover uma perturbação em $(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}$ pela adição de uma pequena constante ao longo da diagonal principal da matriz $\mathbf{X}^{\mathbf{T}}\mathbf{X}$. Essa simples estratégia é capaz de minimizar os efeitos provocados pelo mal-condicionamento desta matriz. Na regressão *ridge* a estimativa dos coeficientes de regressão é a que minimiza:

$$J(\mathbf{b}_{rr}) = (\| \mathbf{X}\mathbf{b}_{rr} - \mathbf{y} \|^2 + \lambda \| \mathbf{b}_{rr} \|^2) \quad \textbf{(4)}$$

que é similar a função custo no LS (*Least Squares*) restrita a $\| \mathbf{b}_{rr} \|^2 \leq c$ (ver **Figura 2.3**). A solução com essa restrição é dada por:

$$\mathbf{b}_{rr} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \quad \textbf{(5)}$$

em que: $\lambda \geq 0$ é a constante de perturbação (parâmetro *ridge*), **I** é a matriz identidade de dimensões *p* x *p*. Se **X** estiver autoescalonada ao longo das colunas, a matriz $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})$ será equivalente ao parâmetro *ridge* adicionado em cada elemento da diagonal da matriz de correlação de **X**.

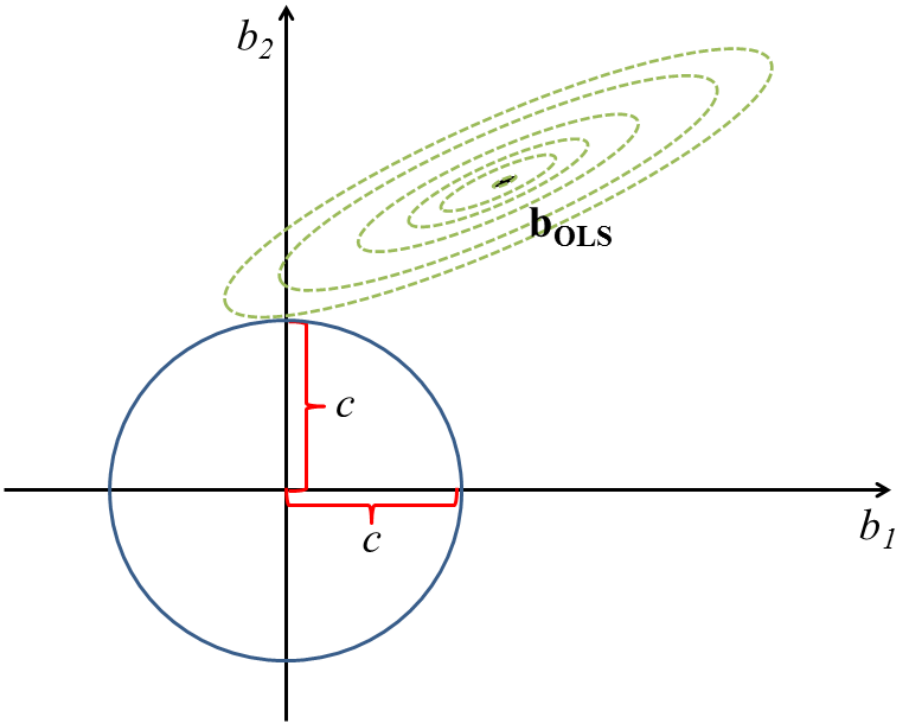



**Figura 2.3.** Espaço bidimensional ($b_1$ vs $b_2$). $b_{OLS}$ é o valor obtido pelo método dos mínimos quadrados ordinários. A restrição *ridge* é representada pelo circulo azul. Adaptado de Rasmussen & Bro (2012)**.**

Na **Figura 2.3** a elipse representada em cada contorno corresponde a um valor diferente da soma quadrática residual (SQR) e o ponto interno é o menor valor de SQR que equivale à solução obtida pelo método dos mínimos quadrados. Nessa representação, percebe-se o aumento de SQR ao passo que $b_1$ e $b_2$ ficam mais distantes da solução $\mathbf{b}_{OLS}$. O círculo azul corresponde à representação da restrição *ridge* imposta aos coeficientes de regressão. Dessa

forma, a solução *ridge* é obtida a partir da minimização da área da elipse e do círculo simultaneamente.

O parâmetro de ajuste λ, apresentado na **Equação (5)**, guarda uma relação inversa com o valor de *c*. Pois quanto maior λ é, mais próximo de zero **b** deve estar. Por outro lado, no caso extremo, quando λ = 0, os valores de **b** são iguais a solução $\mathbf{b_{OLS}}$. Por esse motivo, deve-se assumir um compromisso entre valores de $\lambda$ e SQR, uma vez que quando a restrição *ridge* é imposta o valor de SQR é aumentado em relação à solução LS. A seleção do parâmetro de ajuste λ é essencial para a obtenção do modelo adequado, pois a escolha de um parâmetro demasiadamente grande ou pequeno pode ocasionar, respectivamente, em subajuste ou sobreajuste (FORRESTER; KALIVAS, 2004).

Alguns autores tem sugerido escolher o valor do parâmetro *ridge* a partir da variação do RMSEC (NÆS, 2002) ou RMSECV (GOLUB; HEATH; WAHBA, 1979) em função de λ, ou de acordo com a estabilidade dos coeficientes de regressão (HANSEN, 1992; HOERL; KENNARD, 1970a) medido pela variação da norma dos coeficientes de regressão em função de λ. Como alternativa Forrester & Kalivas (2004) propuseram um critério em que o parâmetro *ridge* é determinado a partir de uma abordagem harmônica. Recentemente Kalivas; Heberger & Andries (2015) selecionaram de forma automática o melhor parâmetro de ajuste do modelo a partir da soma do ranking de diferenças.

#### 2.3.1.3 Regressão em mínimos quadrados parciais

A abordagem original do PLS (*Partial Least Squares*) foi proposta em torno de 1975 por Herman Wold e aprimorada em anos posteriores (WOLD; SJÖSTRÖM; ERIKSSON, 2001). O PLS é um método de calibração inversa que usa os escores de **X**, representado aqui por **T,** como os preditores de **y** buscando um bom modelo linear entre **X** e **y**. Nessa busca, os escores são rotacionados de forma a maximizar a covariância entre **X** e **y**.

Os coeficientes da combinação linear (representados por $\mathbf{W}^{*}$) entre **T** e **X** recebem a nomenclatura de *loading weights* (NÆS, 2002). Em termos matriciais, a relação acima é descrita por:

$$\mathbf{T} = \mathbf{X}\mathbf{W}^{*} \tag{6}$$

Os escores de **X** quando multiplicados pelos pesos **P,** deve satisfazer a **Equação (7).** Nessa equação os resíduos $\mathbf{E_x}$ são minimizados.

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

Da mesma forma, **u** os escores de **y** são multiplicados por **c** (pesos de **y**) de modo que os resíduos $\mathbf{e_y}$, na **Equação (8)**, sejam também minimizados.

$$\mathbf{y} = \mathbf{u}\mathbf{c}^{\mathbf{t}} + \mathbf{e_y} \tag{8}$$

A matriz **T** deve ser um bom preditor de **y**, assim sendo, a **Equação (9)** deve ser satisfeita.

$$\mathbf{y} = \mathbf{T}\mathbf{c}^{\mathbf{t}} + \mathbf{F} \tag{9}$$

Os resíduos, **F** expressam os desvios entre as respostas observadas e modeladas. Usando as **Equações (6) e (9)**, pode-se obter a seguinte relação para um modelo de regressão multivariado:

$$\mathbf{y} = \mathbf{X}\mathbf{W}^{*}\mathbf{c}^{\mathbf{t}} + \mathbf{F} = \mathbf{X}\mathbf{b} + \mathbf{F} \tag{10}$$

Assim sendo, os coeficientes de regressão PLS, $\mathbf{b_{PLS}}$, podem ser escritos como:

$$\mathbf{b_{PLS}} = \mathbf{W}^{*}\mathbf{c}^{\mathbf{t}} \quad \textbf{(11)}$$

Quando apenas um parâmetro **y** é usado, o PLS é definido como PLS1. Contudo, todas as equações podem ser reescritas considerando os vetores $\mathbf{y_1}$ e $\mathbf{y_2}$ em uma matriz **Y**. Se o cálculo é realizado assim, o PLS recebe a nomenclatura de PLS2.

Os *loadings* $\mathbf{W}^{*}$ e **c** fornecem informações a respeito de como as variáveis se combinam para formar a relação quantitativa entre **X** e **y**, proporcionando assim uma interpretação dos escores, **T** e **u**. Assim, os *loadings* são indispensáveis para compreensão de quais variáveis de **X** fornecem informações mais importantes para o modelo.

#### 2.3.1.4 Decomposição em valores singulares

A decomposição em valores singulares (SVD: *Singular-value decomposition*) tem sido usada extensamente para a obtenção da pseudoinversa nos métodos de regressão multivariados (KALIVAS, 2009). No SVD a matriz **X** pode ser reescrita como:

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}} \quad \textbf{(12)}$$

em que: **U** ($N \times N$) é a matriz de autovetores para $\mathbf{XX^T}$, **V** ($p \times p$) representa a matriz de autovetores para $\mathbf{X^TX}$ e $\Sigma$ ($N \times p$) é uma matriz diagonal, em que os valores diferentes de zero são iguais aos autovalores $\sigma_{kk}$. Os autovalores de **X** ($\sigma_{kk}$) são iguais à raiz quadrada dos autovalores de $\mathbf{XX^T}$ e $\mathbf{X^TX}$.

As colunas de **U** formam uma base ortonormal do conjunto de vetores que abrangem o espaço coluna de **X**, ou seja, uma coluna de **X** pode ser escrita como uma combinação linear das colunas de **U**. Similarmente, as colunas de **V** formam uma base ortonormal do conjunto

de vetores que abrangem o espaço linha de **X**, ou seja, uma linha em **X** pode ser escrita como uma combinação linear das colunas em **V**.

Os valores singulares são dispostos em ordem decrescente em $\mathbf{\Sigma}$ e as colunas em **U** e **V** estão na mesma ordem relativa. As magnitudes dos valores singulares transmitem a informação da quantidade ou da variabilidade nas respectivas direções dos autovetores (KALIVAS, 2009). Após realizar a decomposição, o SVD pode ser usado na redução da dimensionalidade da matriz **X**. Isso é alcançado reconstruindo **X** com as matrizes **U**, $\Sigma$ e **V** limitadas a um número determinado de dimensões. Tipicamente, se a matriz é limitada a *A* autovalores (geralmente *A* é o *rank* de X) as matrizes **U**, $\Sigma$ e **V** passam a ter dimensões *N* x *A*, *A* x *A* e *p* x *A*, respectivamente.

O cálculo dos coeficientes de regressão nos métodos apresentados nesta tese podem ser estimados em função da decomposição SVD (KALIVAS, 1999; KALIVAS, 2009). Assim sendo, de forma generalizada temos que:

$$\hat{\mathbf{b}} = \mathbf{V}\mathbf{F}\Sigma^{-1}\mathbf{U}\mathbf{y} = \sum_{k=1}^{p} f_k \frac{\mathbf{u}_k^{\mathrm{T}}\mathbf{y}}{\sigma_k}\mathbf{v}_k \qquad \textbf{(13)}$$

A diferença entre os métodos de regressão concentra-se na matriz de filtros **F** (FORRESTER; KALIVAS, 2004; KALIVAS, 1999; KALIVAS, 2009)**.** No método LS os filtros $f_k$ são todos iguais a um. No PLS e RR os filtros podem variar em uma faixa de valores $0 \leq f_k \leq 1$ e $0 \leq f_k \leq \infty$, respectivamente.

### 2.3.2 Transferência de modelos de calibração multivariada

A transferência de calibração tem por objetivo corrigir os efeitos provocados pela mudança nas condições analíticas após a calibração ter sido realizada. Tais mudanças podem estar associadas a diversos fatores (alguns deles já foram citados na introdução). Um fator

clássico que provoca esse tipo de problema é o uso de um equipamento na etapa de predição diferente do usado na calibração. Quando isso ocorre as diferenças entre a resposta instrumental dos dois equipamentos podem ser observadas a partir das mudanças de intensidade e deslocamento nos eixos das variáveis (ARAÚJO, 2006).

#### 2.3.2.1 Minimização das diferenças instrumentais

Os efeitos na diferença de resposta instrumental podem ser minimizados antes da construção do modelo de calibração, basicamente de duas maneiras. Na primeira, estratégias são tomadas de modo que os equipamentos sejam mantidos sob as mesmas condições de análise (BAKEEV; KURTYKA, 2005; BROWN, 2009; SWIERENGA et al., 1998) ou pode-se, posteriormente ao registro das medidas, recorrer às técnicas de pré-processamentos dos dados de forma a eliminar as fontes de diferenças entre os equipamentos (FEUDALE et al., 2002; PEREIRA et al., 2008). Em uma segunda tentativa, é possível incluir os efeitos da variação de fatores, previamente definidos com base na diferença instrumental, de forma implícita no modelo de calibração. Para isso, pode-se recorrer ao uso de um planejamento experimental para promover a variação dos fatores de forma mais satisfatória (BROWN, 2009; FLATEN; WALMSLEY, 2003).

É possível perceber que essas estratégias são úteis, porém oferecem como dificuldades a identificação e correção de todas as fontes de variação na resposta instrumental. Assim, se tornam muito difíceis de serem realizadas para a maioria das aplicações devido à complexidade do processo. Diante de tais problemas, o modelo de calibração multivariada em análise de rotina requer outras maneiras de corrigir as diferenças instrumentais como fonte de viés na predição. Para tal, a padronização das respostas instrumentais vem sendo bastante usada.

### 2.3.2.2 Padronização das respostas espectrais

A padronização ocorre de forma que uma amostra medida em um equipamento "secundário" seja corrigida para se assemelhar à resposta obtida no equipamento "primário". Nesse contexto, a padronização das respostas espectrais é geralmente baseada em transformações matemáticas que envolvem o uso do espectro total ou janelas de variáveis dentro do espectro. Essa estratégia é capaz de corrigir diferenças de intensidade, de *background*, desalinhamento de comprimentos de onda e alargamento de picos (BROWN, 2009).

Para este fim, um conjunto contendo $N_{trans}$ "amostras de transferência" medidas em ambos os instrumentos são usados para construir um modelo de padronização na forma:

$$\mathbf{X^P = X^S B_{DS}} \quad \textbf{(14)}$$

em que, as matrizes $\mathbf{X}^P$ ($N_{trans} \times p$) e $\mathbf{X}^S$ ($N_{trans} \times p$) são formadas pelos espectros das amostras de transferência adquiridas nos instrumentos primário e secundário, respectivamente. A matriz de transformação $\mathbf{B_{DS}}$ ($p \times p$) é usada na padronização do novo espectro $\mathbf{x}^S$ ($1 \times p$) adquirido no equipamento secundário.

$$\hat{\mathbf{x}}^P = \mathbf{x}^S \mathbf{B_{DS}} \quad \textbf{(15)}$$

A fim de gerar uma estimativa $\hat{\mathbf{x}}^P$ ($1 \times p$) do espectro $\mathbf{x}^P$ no equipamento primário. A **Equação (15)** pode também ser escrita:

$$\hat{x}_k^P = \mathbf{x}^S \mathbf{b}_k \text{ , } k = 1, 2, ..., p \quad \textbf{(16)}$$

em que, $\hat{x}_k^P$ denota a *k*-ésima variável do espectro padronizado e $\mathbf{b}_k$ a *k*-ésima coluna da matriz de transformação.

#### 2.3.2.2.1 Padronização direta

Esse método foi proposto por Wang et al. (1991) para corrigir as diferenças entre os espectros registrados no equipamento secundário e primário. Neste método, as matrizes $\mathbf{X}^{\boldsymbol{P}}$ e $\mathbf{X}^{\boldsymbol{S}}$ são descritas pelo modelo expresso na **Equação (14)**. Assim sendo, a obtenção da matriz **F** é realizada pela multiplicação da **Equação (14)** pela inversa da matriz $\mathbf{X}^S$ (HONORATO et al., 2007).

$$\mathbf{B_{DS}} = (\mathbf{X}^S)^{-1}\mathbf{X}^P \qquad \textbf{(17)}$$

Semelhante às dificuldades apresentadas na calibração MLR, no cálculo de $\mathbf{B_{DS}}$ também é necessário que $N_{trans}$ (número de amostras de transferência) seja maior ou igual a *p* (número de variáveis na matriz $\mathbf{X}^S$) para que $\mathbf{X}^S$ seja invertível. Com intuito de corrigir esses problemas, a matriz de transformação $\mathbf{B_{DS}}$ é tipicamente obtida por regressão em componentes principais ou regressão pelo método dos mínimos quadrados parciais (FEUDALE et al., 2002). Mesmo usando os métodos PCR e PLS para o cálculo de $\mathbf{B_{DS}}$, deve-se assumir que as mudanças nas respostas ocorrem pela diferença instrumental. Portanto, se houver variação na composição química da amostra, essa também será incorporada no modelo fazendo com que $\mathbf{B_{DS}}$ não represente adequadamente as diferenças instrumentais (BROWN, 2009).

Uma vez que cada variável do equipamento primário é predita a partir de todo o espectro do equipamento secundário, esses modelos são calculados por meio de uma grande quantidade de variáveis, gerando assim muitos coeficientes de regressão e consequentemente

um maior risco de sobreajuste. O procedimento para obtenção dos coeficientes de padronização direta é apresentado na **Figura 2.4**.

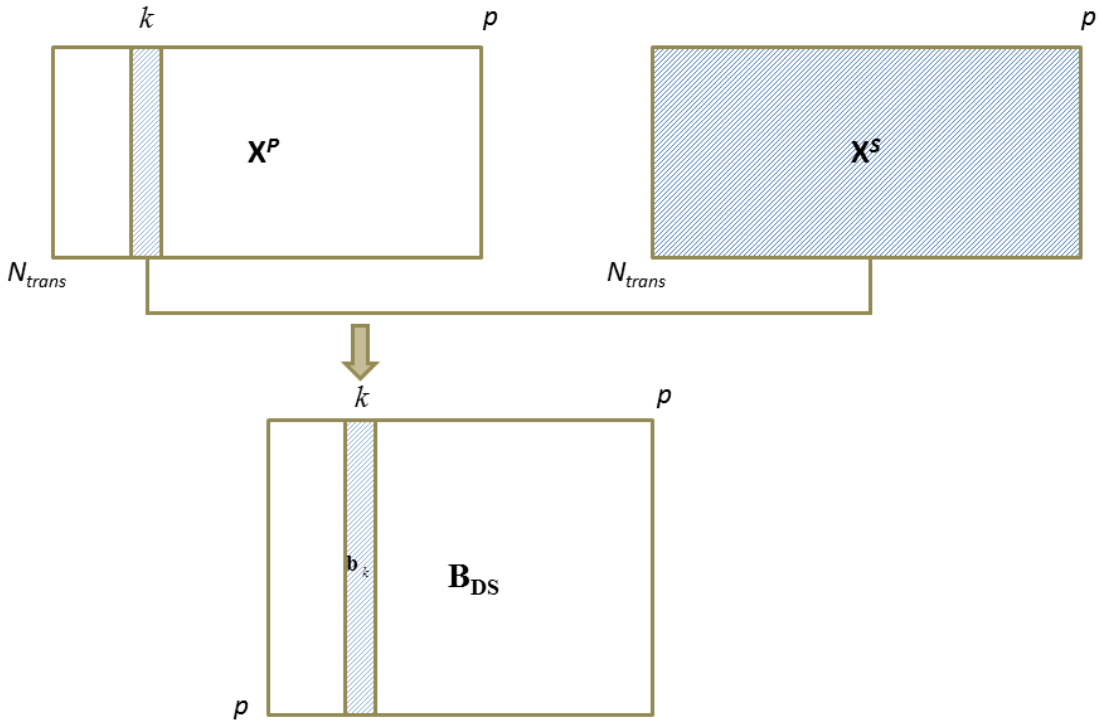



**Figura 2.4.** Obtenção dos coeficientes de padronização direta. Adaptado de Swierenga et al. (1998)**.**

### 2.3.2.2.2 Padronização direta por partes

No método PDS, também proposto por Wang et al. (1991), cada variável do equipamento primário é relacionada a uma janela espectral móvel de medidas realizadas no equipamento secundário. Esta simples estratégia permite reduzir o risco de sobreajuste que ocorre no método DS, uma vez que o número de parâmetros estimados em cada regressão é diminuído significativamente.

Em uma padronização PDS tem-se $p$ ($k$ = 1, 2, ..., $p$) regressões realizadas. Para cada passo $k$ da janela móvel, um vetor de coeficientes de regressão $\mathbf{b}_k$ associado à predição da $k$-ésima variável é obtido por PCR ou PLS (ver **Figura 2.5**).

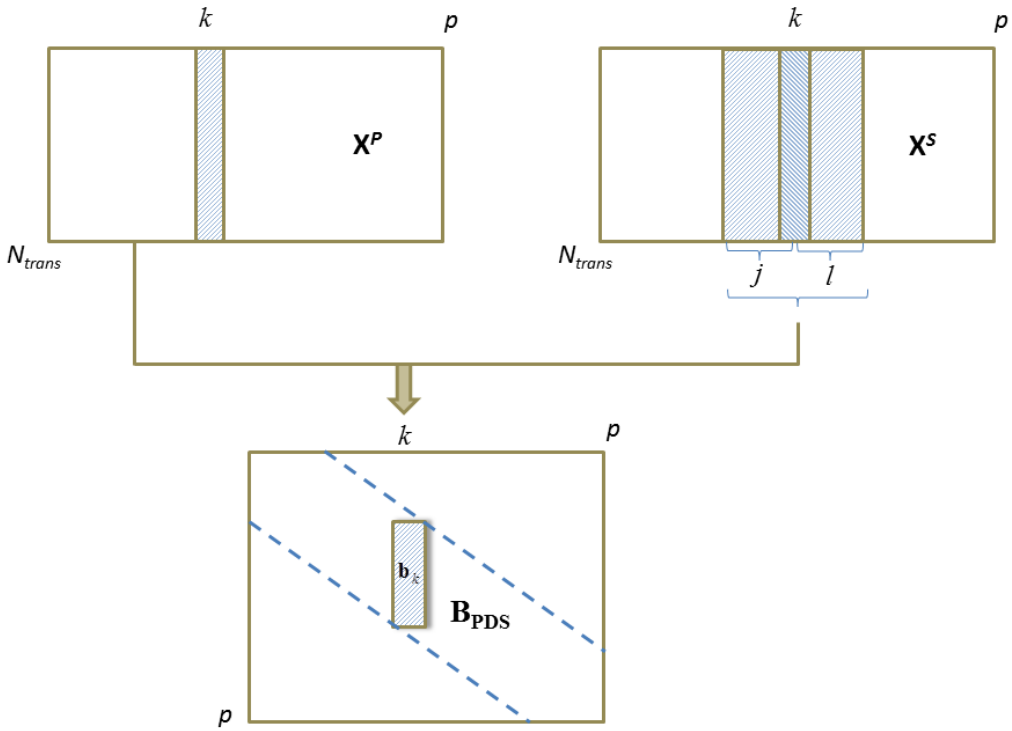



**Figura 2.5.** Obtenção dos coeficientes de padronização direta por partes. Adaptado de Swierenga et al. (1998).

Os vetores $\mathbf{b}_k$ podem ser organizados em uma matriz $\mathbf{B}_{\mathbf{PDS}}$, **Equação (18)**, de modo que os sinais analíticos do equipamento secundário possam ser usados para predizer os do equipamento primário.

$$\mathbf{B}_{\mathbf{PDS}} = \mathrm{diag}(\mathbf{b}_1^{\mathrm{T}}, \mathbf{b}_2^{\mathrm{T}} \cdots, \mathbf{b}_k^{\mathrm{T}}) \qquad \textbf{(18)}$$

Os vetores $\mathbf{b}_k^{\mathrm{T}}$ possuem os elementos com índices *j*-*k* a *l*+*k* diferentes de zero. Em que *j*+*l* corresponde ao tamanho da janela móvel que deve ser otimizada no processo PDS. Neste caso, a matriz de transformação $\mathbf{B}_{\mathbf{PDS}}$ tem uma estrutura esparsa, porque cada coluna $\mathbf{b}_k^{\mathrm{T}}$ só terá valores diferentes de zero dentro da janela em torno do *k*-ésimo elemento. A janela espectral pode ser simétrica ou assimétrica em torno de cada variável *k*. Quando o formato simétrico é escolhido, é preferível que desalinhamentos espectrais sejam negligenciáveis. No caso contrário, deve-se optar por uma janela assimétrica (BROWN, 2009).

#### 2.3.2.3 O novo método

O novo método de transferência de calibração aqui apresentado, e recentemente descrito em Galvão et al. (2015), é uma combinação da correção univariada e de uma regressão robusta. Dessa forma, nas seções seguintes, essas duas metodologias serão descritas.

##### 2.3.2.3.1 Correção univariada

Essa abordagem foi inicialmente desenvolvida por Shenk et al. (1985) e depois reportada por Nørgaard (1995) como padronização em comprimento de onda único. A correção univariada (UVC: *Univariate Correction*) é um procedimento empregado para padronizar a resposta instrumental do equipamento secundário em cada variável registrada. Como descrito em Feudale et al. (2002), esta correção pode ser expressa de acordo com a equação:

$$\hat{x}_k^P = \alpha_k x_k^S + \beta_k \quad \textbf{(19)}$$

em que $x_k^S$, $\hat{x}_k^P$ denota o sinal original e os valores padronizados da *k*-ésima variável para o equipamento secundário. O coeficiente $\alpha_k$ é usado para corrigir a diferença de intensidade na resposta entre o equipamento primário e secundário, enquanto que o coeficiente $\beta_k$ responde pelas mudanças na linha de base que pode ocorrer em todo espectro. Os coeficientes $\alpha_k$ e $\beta_k$ podem ser obtidos pela regressão com o método dos mínimos quadrados empregando o conjunto de amostras de transferência. A notação $\hat{x}_k^P$ indica que esta é uma estimativa de $x_k^P$, que é o valor de $x_k$ medido no equipamento primário.

Alguns refinamentos desse método foram realizados ao longo dos anos com objetivo de melhorar a correção de intensidade quando as amostras de transferência não são da mesma

natureza das amostras a serem analisadas (BOUVERESSE; MASSART; DARDENNE, 1995) e corrigir problemas de deslocamento de picos (SHENK; WESTERHAUS, 1989).

#### 2.3.2.3.2 Regressão robusta

O desenvolvimento matemático apresentado nesta seção é uma adaptação da formulação apresentada por Hindi & Boyd (1998). Para tal, suponha que o vetor **x** de resposta instrumental pode ser afetado por uma perturbação estocástica **Δx** de média zero e covariância conhecida. Neste caso, uma solução mais robusta para o problema de regressão pode ser obtida pela minimização da seguinte função custo (GALVÃO et al., 2015):

$$J(\mathbf{b}) = E \parallel (\mathbf{X} + \Delta\mathbf{X})\mathbf{b} - \mathbf{y} \parallel^2 \qquad \textbf{(20)}$$

em que $E$ denota o operador esperança nas variáveis aleatórias. Nesta equação, $\Delta\mathbf{X}$ ($N \times p$) é um termo de perturbação com $E[\Delta\mathbf{X}] = \mathbf{0}$ e $E[\Delta\mathbf{X}^{\mathrm{T}}\Delta\mathbf{X}] = \mathbf{R}$, onde $\mathbf{R}$ é uma matriz ($p \times p$) conhecida. Esta função custo pode ser reescrita como:

$$\begin{aligned} J(\mathbf{b}) &= E[(\mathbf{X} + \Delta\mathbf{X})\mathbf{b} - \mathbf{y}]^T[(\mathbf{X} + \Delta\mathbf{X})\mathbf{b} - \mathbf{y}] \\ &= \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{b} + 2\mathbf{b}^T\mathbf{X}^T E[\boldsymbol{\Delta}\mathbf{X}]\mathbf{b} - 2\mathbf{b}^T\mathbf{X}^T\mathbf{y} + \mathbf{b}^T E[\Delta\mathbf{X}^T\Delta\mathbf{X}]\mathbf{b} - 2\mathbf{b}^T E[\Delta\mathbf{X}^T]\mathbf{y} + \mathbf{y}^T\mathbf{y} \\ &= \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{b} - 2\mathbf{b}^T\mathbf{X}^T\mathbf{y} + \mathbf{b}^T\mathbf{R}\mathbf{b} + \mathbf{y}^T\mathbf{y} \\ &= \mathbf{b}^T(\mathbf{X}^T\mathbf{X} + \mathbf{R})\mathbf{b} - 2\mathbf{b}^T\mathbf{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{y}. \end{aligned} \qquad \textbf{(21)}$$

Uma vez que $(\mathbf{X}^T\mathbf{X} + \mathbf{R})$ é simétrica, segue-se que o vetor gradiente $\partial J / \partial \mathbf{b}$ do custo em relação à **b** é dado por:

$$\frac{\partial J}{\partial \mathbf{b}} = 2(\mathbf{X}^T\mathbf{X} + \mathbf{R})\mathbf{b} - 2\mathbf{X}^T\mathbf{y} \qquad \textbf{(22)}$$

Portanto, a solução $\mathbf{b_{rob}}$, que corresponde ao valor de **b** para o qual $\partial J / \partial \mathbf{b} = 0$ é:

$$\mathbf{b_{rob}} = (\mathbf{X}^T\mathbf{X} + \mathbf{R})^{-1}\mathbf{X}^T\mathbf{y} \quad \textbf{(23)}$$

desde que a inversa indicada exista. Isto deverá ser garantido se $\mathbf{X}^T\mathbf{X}$ for invertível porque $\mathbf{R} = E[\Delta\mathbf{X}^T\Delta\mathbf{X}]$ sempre é uma matriz positiva definida ou pelo menos positiva semidefinida.

A interpretação direta do significado da matriz **R** pode ser entendida como segue. Inicialmente, assuma a matriz de perturbação **ΔX** como:

$$\Delta\mathbf{X} = \begin{bmatrix} \Delta X_{1,1} & \Delta X_{1,2} & \cdots & \Delta X_{1,p} \\ \Delta X_{2,1} & \Delta X_{2,2} & \cdots & \Delta X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta X_{N,1} & \Delta X_{N,2} & \cdots & \Delta X_{N,p} \end{bmatrix} \quad \textbf{(24)}$$

e também assuma que em cada variável espectral (os elementos de cada coluna de Δ**X**) a perturbação possui média nula e variáveis aleatórias com distribuições idênticas. Neste caso, a covariância entre as perturbações das colunas *i* e *j* será $E\left[\Delta X_{n,i}\Delta X_{n,j}\right] = \sigma_{i,j}, \forall n.$ Os elementos (*i, j*) da matriz **R** serão então dados por:

$$R_{i,j} = E\left[\sum_{n=1}^{N}\Delta X_{n,i}\Delta X_{n,j}\right] = \sum_{n=1}^{N}E\left[\Delta X_{n,i}\Delta X_{n,j}\right] = \sum_{n=1}^{N}\sigma_{i,j} = N\sigma_{i,j} \quad \textbf{(25)}$$

Portanto, **R** pode ser interpretado como a matriz de covariância das perturbações, multiplicada pelo número de amostras *N*.

Vale a pena notar que a **Equação (23)** é similar à solução obtida em uma Regressão *Ridge* (DRAPER; SMITH, 1998; NGO; KEMÉNY; DEÁK, 2003). Contudo, em uma regressão *Ridge* convencional, o termo de regularização é $\lambda\mathbf{I}$, onde **I** é uma matriz identidade e $\lambda > 0$ é um parâmetro que necessita ser ajustado. Em contraste, na solução da regressão robusta, o

termo de regularização é dado pela matriz de covariância **R**, seguindo o desenvolvimento teórico apresentado acima, sem a necessidade de ajuste de parâmetros.

Se o modelo expandido com o termo linear $b_0$, isto é:

$$y = b_0 + b_1 x_1 + \cdots + b_p x_p + e = \mathbf{b}[1 \ \mathbf{x}] + e \tag{26}$$

a solução pelo método dos mínimos quadrados $\mathbf{b_{LS}}$ é obtida pela **Equação (2)** usando a matriz **X** expandida com a coluna inicial de uns. Usando a notação $\mathbf{X_a} = [\mathbf{1}_{N\times 1} \ \mathbf{X}]$, segue-se que $\mathbf{b_{LS}} = (\mathbf{X_a}^T \mathbf{X_a})^{-1} \mathbf{X_a}^T \mathbf{y}$. No caso da regressão robusta, a solução torna-se:

$$\mathbf{b_{rob}} = (\mathbf{X_a}^T \mathbf{X_a} + \mathbf{R_a})^{-1} \mathbf{X_a}^T \mathbf{y} \tag{27}$$

em que $\mathbf{R_a}$ é uma matriz de dimensões $(p + 1) \times (p + 1)$ dado por:

$$\mathbf{R_a} = E\{[\mathbf{0}_{N\times 1} \ \Delta\mathbf{X}]^T [\mathbf{0}_{N\times 1} \ \Delta\mathbf{X}]\} = \begin{bmatrix} 0 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{p\times 1} & \mathbf{R} \end{bmatrix} \tag{28}$$

Nesta expressão, os zeros surgem porque o coeficiente linear $b_0$ na **Equação (26)** é multiplicado por um valor constante igual a um, que não é afetado pelo termo de perturbação.

## 2.4 Experimental

### 2.4.1 O método de transferência de calibração

O novo método pressupõe a disponibilidade de um conjunto de calibração de $N$ amostras com as respostas espectrais registradas no instrumento primário sobre $p$ variáveis, bem como um conjunto de amostras de transferência $N_{trans}$ com os sinais registrados nos instrumentos primários e secundários. O novo método nesta tese compreende os seguintes passos que serão apresentados a seguir, conforme o esquema representado na **Figura 2.6.**

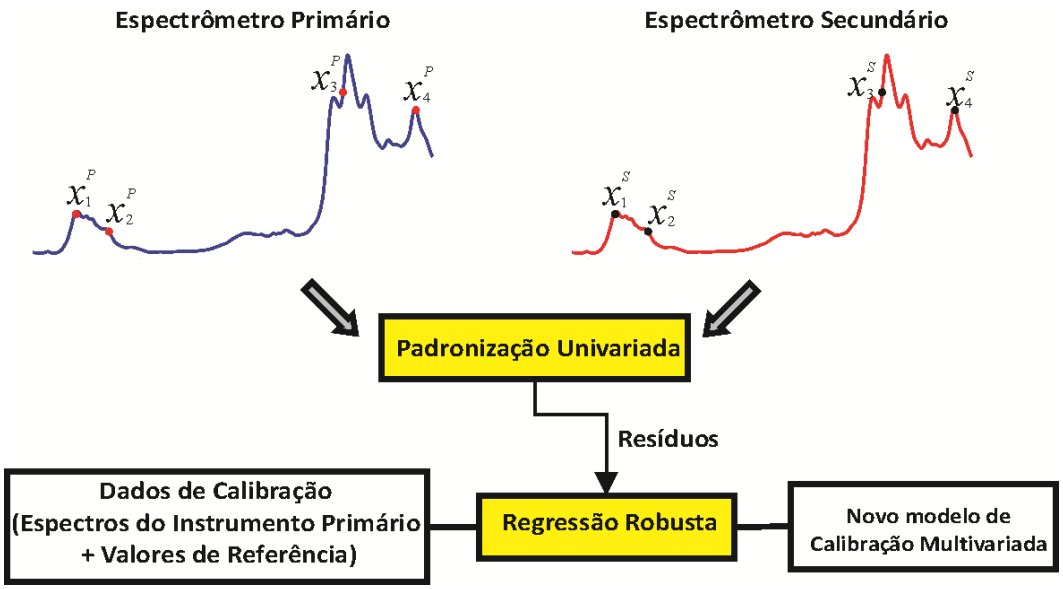



**Figura 2.6.** Representação esquemática dos principais passos para realizar a transferência de calibração usando o método aqui proposto.

#### 2.4.1.1 Etapa 1 (correção univariada)

A correção univariada é um procedimento empregado para padronizar a resposta instrumental do equipamento secundário em cada variável registrada. Esse procedimento é descrito na **Seção 2.3.2.3.1**. Após o sinal obtido no equipamento secundário ser corrigido, os resíduos deixados por esse método são calculados por $e_k = \hat{x}_k^P - x_k^P$ nas seções que se seguem.

#### 2.4.1.2 Etapa 2 (regressão robusta)

A formulação da regressão robusta descrita na **seção 2.3.2.3.2** é empregada por considerar que o modelo MLR, a qual foi calibrada com os dados $x_k^P$, $k$ = 1,…, $p$, do equipamento primário, deverá ser aplicado aos valores da resposta padronizada $\hat{x}_k^P$, $k$ = 1, …, $p$. Assim sendo, o resíduo $e_k$ pode ser considerado como uma perturbação que afeta a resposta instrumental. Neste caso, se o resíduo $e_n$ para $n$-ésima amostra de transferência é denotada por $e_{n,k}$, o elemento ($i$, $j$) da matriz **R** pode ser estimado como:

$$\hat{R}_{i,j} = \frac{N}{N_{trans} - 2} \sum_{n=1}^{N_{trans}} e_{n,i} e_{n,j} \qquad \textbf{(29)}$$

Nesta equação, ($N_{trans}$ − 2) é empregado porque dois graus de liberdade são consumidos no procedimento de correção univariada. Vale a pena considerar que, pelo menos, três amostras de transferência (ou seja, $N_{trans} \geq 3$) são necessárias, para estimar os coeficientes $\alpha_k$ e $\beta_k$ para a correção de cada variável $x_k$. O fator $N$ no numerador é incluído porque a definição $\mathbf{R} = E[\Delta\mathbf{X}^T\Delta\mathbf{X}]$ não foi normalizada em relação ao número de amostras de calibração $N$. Finalmente, a estimativa da matriz $\hat{\mathbf{R}}$ assim obtida é empregada no procedimento de regressão robusta usando o conjunto de dados de calibração.

Neste contexto, o uso de $\hat{\mathbf{R}}$ na regressão robusta contabiliza as imperfeições no procedimento de correção univariada. Mais especificamente, presume-se que a resposta padronizada $\hat{x}_k^P$ calculada de medidas futuras no instrumento secundário deve ter desvios das medidas $x_k^P$ que seriam obtidas usando o equipamento primário. Estes desvios desempenham a função das perturbações estocásticas na formulação da regressão robusta apresentada na **Seção 2.3.2.3.2**. Os resíduos de regressão obtidos no procedimento de correção univariada (isto é, as diferenças entre $\hat{x}_k^P$ e $x_k^P$ no conjunto de amostras de transferência) são

considerados como realizações dessas perturbações, que são utilizados para estimar a matriz de covariância **R**.

#### 2.4.1.3 Etapa 3 (predição de novas amostras)

Após adquirir o sinal de uma nova amostra no equipamento secundário, a **Equação (19)** é aplicada para corrigir os valores da medida, e então o modelo robusto MLR obtido na **etapa 2** é empregado para predizer a propriedade de interesse.

### 2.4.2 Avaliação dos modelos

Na seção de resultados e discussão, o novo método (correção univariada e regressão robusta) deverá ser comparado com o PLS-PDS, bem como com o MLR usando a correção univariada (MLR-UVC). Os resultados foram avaliados em termos do RMSEP (*Root Mean Squares Error of Prediction*). A notação $RMSEP_A^B$ deverá ser usada para representar o valor de RMSEP obtido no instrumento B predito pelo modelo calibrado no instrumento A. Se algum procedimento de transferência é empregado, a notação $RMSEP_{A\text{-}T}^B$ deverá ser adotada. Os instrumentos primário e secundário são representados pelas letras P e S, respectivamente.

### 2.4.3 Dados usados no estudo

Dois conjuntos de dados foram empregados para ilustrar o novo método. O primeiro conjunto de dados é formado por espectros de absorbância em amostras de gasolina registradas em dois espectrômetros FT-NIR, bem como os valores de referência de massa específica (SM: *Specifc Mass*), RON (*Research Octane Number*) e teor de naftênicos. A massa específica é rotineiramente usada para monitorar a qualidade de combustíveis, assim como para checar a conformidade com padrões exigidos pelas agências reguladoras. Os outros dois parâmetros foram incluídos no estudo para representar as classes relacionadas à composição e propriedades físicas, como apresentado por Pereira et al. (2008).

O segundo conjunto de dados, que é de domínio público, é composto por espectros de reflectância difusa NIR em amostras de milho registrados em dois espectrômetros, e os valores de referência para o teor de água e óleo.

### 2.4.4 Conjunto de dados de gasolina

O primeiro conjunto de dados compreende 155 amostras de gasolina coletadas em postos de combustíveis da região do Nordeste brasileiro. As gasolinas contêm ± 25% (v/v) de etanol, em conformidade com os padrões definidos pela Agência Nacional de Petróleo, Gás Natural e Biocombustíveis (ANP). As amostras foram armazenadas em frascos âmbar sob-refrigeração a 5 ºC.

Os valores de referência dos três parâmetros considerados neste trabalho, massa especifica, teor de naftênicos e RON foram obtidos de acordo com a ASTM (*American Society for Testing and Materials*) D1319 e D2700.

Os espectros de cada amostra foram adquiridos na faixa de 1600-2500 nm em intervalos de 2 nm. Os espectrômetros primário e secundário foram um FT-IR Perkin Elmer Spectrum GX e um FT-NIR ABB Bomen MB160D, respectivamente. Para remover a linha de base, a primeira derivada foi calculada nos espectros usando um filtro Savitzky-Golay com um polinômio de 2ª-ordem e uma janela de 13 pontos. Maiores detalhes em relação às condições experimentais podem ser encontrados em Pereira et al. (2008).

### 2.4.5 Conjunto de dados de milho

O segundo conjunto de dados é formado por espectros de reflectância difusa NIR, e os teores de água e óleo de 80 amostras de milho (disponível publicamente em www.eigenvector.com/Data/Corn/), que já foram utilizados em estudos anteriores de transferência de calibração (ANDREW; FEARN, 2004; FAN et al., 2008; HONORATO et al., 2005). Os espectros foram adquiridos na faixa de 1100-2498 nm nos instrumentos m5 e mp5,

que foram adotados como instrumentos primário e secundário, respectivamente. Para corrigir o problema de variação sistemática na linha de base, a primeira derivada foi calculada nos espectros. Para este fim, um filtro Savitzky-Golay com um polinômio de 2ª-ordem e uma janela de 21 pontos foi usado (RINNAN; BERG; ENGELSEN, 2009).

### 2.4.6 Seleção de amostras

Os dados foram divididos em conjuntos de calibração, validação e predição pela aplicação do algoritmo clássico Kennard–Stone (FEUDALE et al., 2002; KENNARD; STONE, 1969) aos espectros do instrumento primário. A divisão de amostra resultante também foi adotada no instrumento secundário, a fim de assegurar uma comparação justa dos resultados nos dois espectrômetros.

Os conjuntos de validação e predição para os dados de gasolina compreendem 30 amostras cada, enquanto que o conjunto de dados de milho foi dividido em 20 amostras para validação e 20 para a predição. As amostras remanescentes em ambos os conjunto de dados foram usadas para calibração.

O conjunto de validação foi empregado na escolha de um número apropriado de variáveis latentes no modelo PLS e para guiar a seleção de variáveis no algoritmo genético para uso no modelo de regressão linear múltipla. O conjunto de predição foi empregado para comparar o desempenho dos modelos PLS e MLR em termos do RMSEP.

O algoritmo Kennard-Stone também foi usado para selecionar um subconjunto de $N_{trans}$ amostras de transferência do conjunto de calibração. O efeito de variar $N_{trans}$ de 3 a 20 foi investigado.

### 2.4.7 Procedimento de modelagem

Todos os cálculos foram realizados usando o software Matlab R2010b. O número de variáveis latentes no modelo PLS foi determinado com base no erro no conjunto de validação

usando o critério do teste F com $\alpha = 0,25$, como sugerido por Haaland & Thomas (1988). As variáveis para o modelo MLR foram selecionadas usando o algoritmo genético (GALVÃO; ARAÚJO, 2009; LEARDI, 2001). Para este propósito, subconjuntos de variáveis foram codificados em cromossomos binários, com genes "1" indicando as variáveis a serem incluídas no modelo.

A aptidão foi calculada como o inverso do RMSEV, na predição do conjunto de validação, obtido pelo modelo MLR restrito ao subconjunto de variáveis codificadas no cromossomo. A probabilidade de um cromossomo qualquer ser selecionado para a população de cruzamento foi proporcional à sua aptidão.

Cada par de indivíduos da população de cruzamento gera dois descendentes. Neste processo os cromossomos dos pais podem ser combinados por cruzamento em um ponto de ruptura. Se o cruzamento não ocorreu os descendentes são simplesmente copias dos seus pais. A mutação consistiu na mudança de um valor de um gene escolhido aleatoriamente. O tamanho da população foi fixado, com a geração anterior substituída por uma nova. O tamanho da população no GA foi definido em 100 indivíduos, as probabilidades de mutação e cruzamento foram atribuídas a 5% e 60%, respectivamente e o número de gerações foi 80.

A largura da janela do PDS foi variada de 3 a 15 pontos e o melhor e o pior resultado para cada subconjunto de amostras de transferência foi anotado. Estas janelas englobam a faixa de valores empregados em estudos anteriores de transferência de calibração, envolvendo os conjuntos de dados sob consideração (HONORATO et al., 2007; HONORATO et al., 2005; MARTINS et al., 2010; PEREIRA et al., 2008). O PDS foi realizado usando o PLS Toolbox (versão 3.5) para o Matlab.

## 2.5 Resultados e discussão

Os resultados serão discutidos usando o conjunto de dados compostos por medidas de absorbância NIR em gasolinas para a determinação dos parâmetros massa específica, RON e naftênicos. Uma discussão semelhante será realizada no conjunto de dados de reflectância NIR em amostras de milho para a determinação dos teores de água e óleo.

### 2.5.1 Avaliação dos conjuntos de dados

Na **Figura 2.7a,** são apresentados os espectros das 155 amostras de gasolina, medidas nos dois instrumentos, e na **Figura 2.7c,** são apresentados os espectros de 80 amostras de milho também registradas em dois diferentes instrumentos. Diferenças de intensidade e linha de base ficam evidentes, permitindo que dois grupos correspondentes aos espectros primários e secundários sejam observados nos dados de gasolina. Nos dados de milho é observado apenas uma tendência de separação.

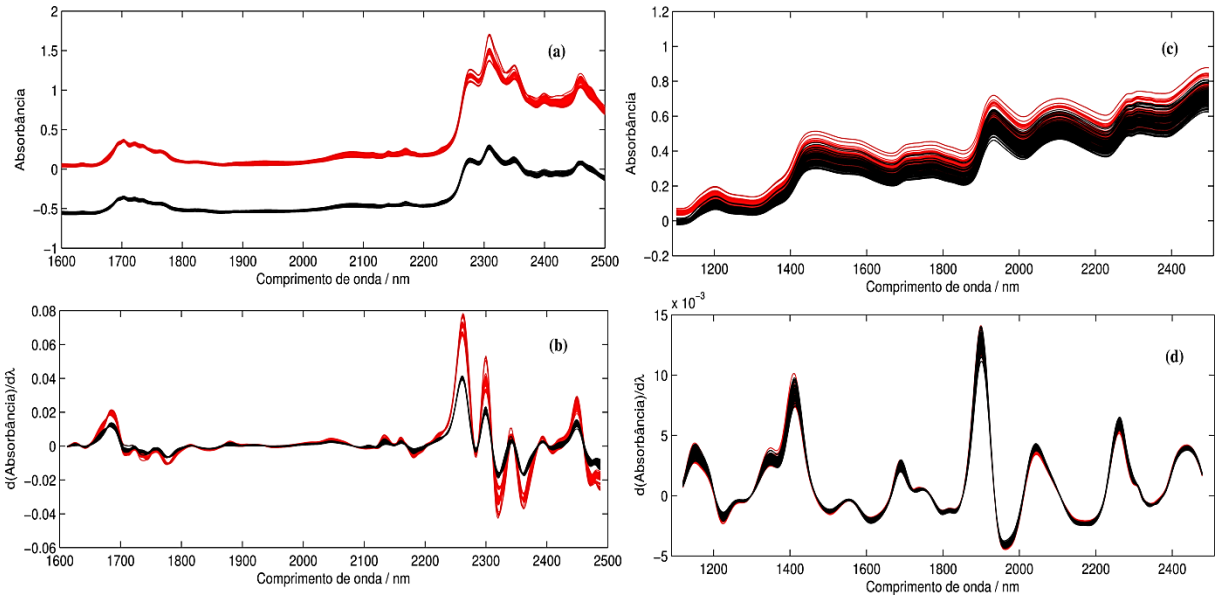



**Figura 2.7.** Espectros NIR registrados nos dois equipamentos, em vermelho o equipamento primário e em preto o equipamento secundário. (a) Espectros de gasolina sem pré-processamento, (b) espectros de gasolina derivativos, (c) espectros de milho sem pré-processamento e (d) espectros de milho derivativos.

Na **Figura 2.7b**, são apresentados os espectros de gasolina após o pré-processamento ser realizado. Isso é necessário para diminuir os efeitos da linha de base no desempenho dos modelos obtidos. O mesmo pré-processamento também foi realizado nos espectros do equipamento secundário. Como pode ser visto, apenas o pré-processamento não foi capaz de eliminar as diferenças entre os sinais dos dois instrumentos. De forma semelhante, na **Figura 2.7d** os espectros de milho após o pré-processamento são apresentados. Como é possível perceber, o pré-processamento diminuiu substancialmente (quando comparado com os dados de gasolina) as diferenças entre os dois conjuntos de espectros.

Na **Tabela 2.1** são apresentados os valores de RMSEP obtidos usando os modelos PLS e MLR calibrados para o instrumento primário, sem o uso de técnicas de transferência. Na comparação dos valores de $RMSEP_P^P$ com $RMSEP_P^S$, a performance de predição é deteriorada quando os modelos são aplicados aos dados do instrumento secundário. Esse fato motiva o uso de técnicas de transferência de calibração. Pode-se ressaltar que mesmo com uma menor diferença (observadas na **Figura 2.7b**) entre os espectros dos instrumentos primário e secundário, quando comparado com os espectros de gasolina, o modelo também foi bastante afetado pela diferença instrumental.

**Tabela 2.1.** Valores de $RMSEP_P^P$ e $RMSEP_P^S$, obtidos com os modelos desenvolvidos no equipamento primário. O número de fatores PLS e variáveis MLR são indicados entre parênteses.

| Model | RMSEP | SM* (kg $m^{-3}$) | RON* (% v/v) | Naftênicos (% v/v) | Água (% m/m) | Óleo (% m/m) |
|---|---|---|---|---|---|---|
| PLS | $RMSEP_P^P$ | 1,4 (9) | 0,34 (5) | 0,56 (4) | 0,026 (12) | 0,056 (9) |
| MLR | $RMSEP_P^P$ | 1,1 (15) | 0,22 (9) | 0,49 (15) | 0,174 (10) | 0,090 (10) |
| PLS | $RMSEP_P^S$ | 54,7 (9) | 5,57 (5) | 6,65 (4) | 1,606 (12) | 0,577 (9) |
| MLR | $RMSEP_P^S$ | 12,1 (15) | 15,54 (9) | 16,22 (15) | 1,544 (10) | 0,128 (10) |

*SM = *Specific Mass*; RON = *Research Octane Number*.

### 2.5.2 Modelos de transferência de calibração

Na **Figura 2.8** são apresentadas as curvas de $RMSEP_{P\text{-}T}^{S}$ em função do número de amostras de transferência ($N_{trans}$) para as três técnicas de transferência de calibração sob consideração. Para o PLS-PDS, os limites da área sombreada correspondem ao melhor e ao pior resultado obtido pela variação do tamanho da janela. Ao comparar os valores de RMSEP desta figura com os apresentados na **Tabela 2.1** é possível perceber que o uso das técnicas de transferência de calibração melhorou as predições obtidas com os espectros do equipamento secundário.

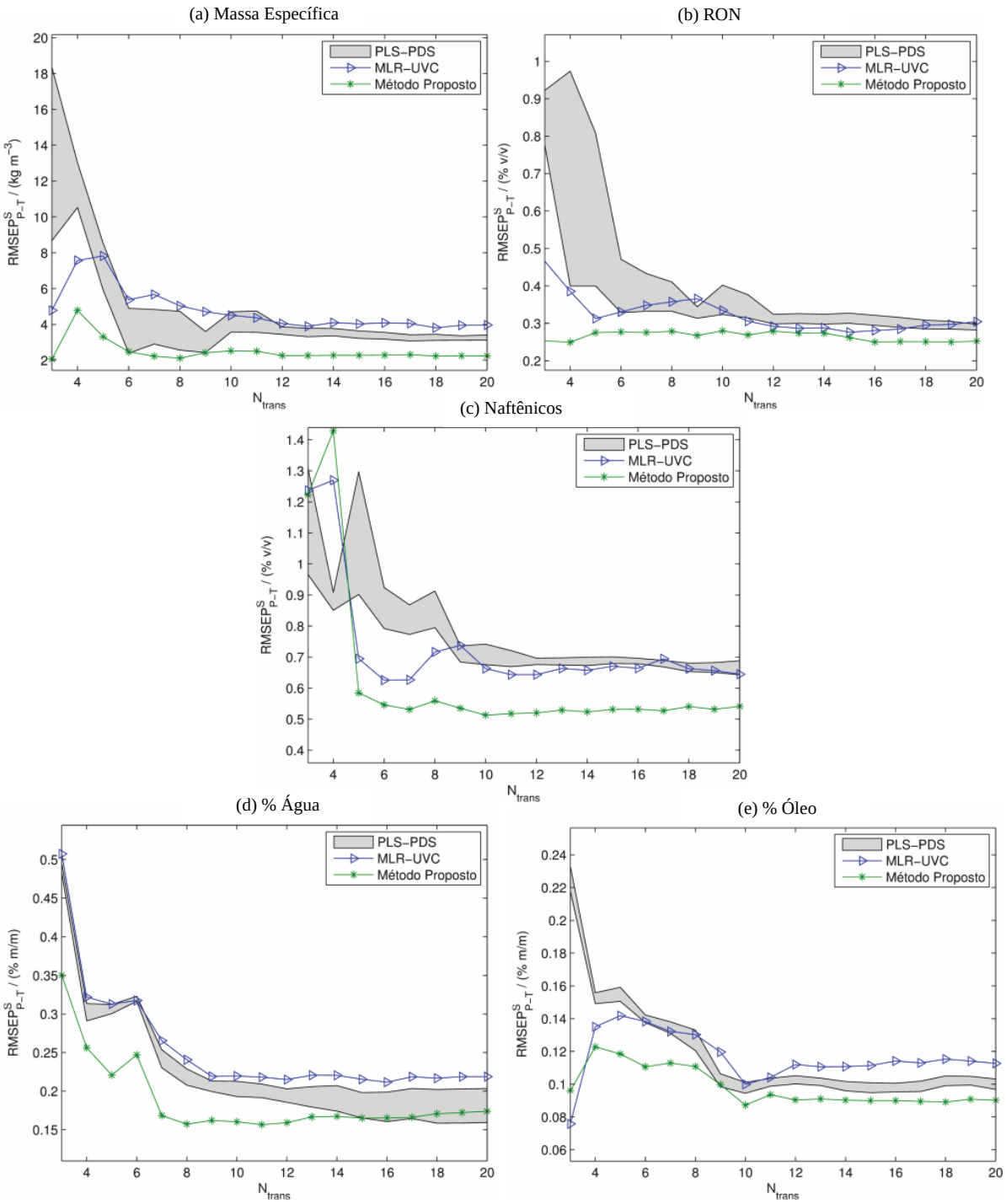



**Figura 2.8.** Valores de $RMSEP_{P\text{-}T}{}^{S}$ *versus* o número $N_{trans}$ de amostras de transferência para (a) SM, (b) RON, (c) Naftênicos, (d) % Água e (e) % Óleo.

Ao adotar pelo menos cinco amostras de transferência, o novo método forneceu menores valores de $RMSEP_{P\text{-}T}{}^{S}$ em todos os casos, quando comparado com o MLR-UVC (ver **Figura 2.8**). Os resultados indicam que o procedimento de regressão robusta foi de fato

importante para levar em conta a estatística associada aos resíduos deixados pelo procedimento de correção univariada. Além disso, no geral o novo método superou o PLS-PDS, mesmo considerando a melhor escolha de janela em cada caso (isto é, o limite inferior para a área sombreada na **Figura 2.8**).

É importante perceber que o método de transferência de calibração proposto, que compreende o procedimento de correção univariada seguido pela regressão robusta, independe da técnica adotada para a seleção de variáveis e de parâmetros de ajuste de tal técnica. De fato, o novo método poderia também ser aplicado se medidas espectrais fossem restritas a um pequeno conjunto de variáveis escolhidas, *a priori*, como no caso dos instrumentos dedicados usando filtros (HAUSER et al., 1995; MALINEN et al., 1998) e/ou *light emitting diodes* (CAPITÁN-VALLVEY; PALMA, 2011; GAIÃO et al., 2008).

### 2.5.3 Avaliação das variáveis selecionadas pelo GA

A característica estocástica do algoritmo genético foi usada, a fim de avaliar o efeito produzido por diferentes subconjuntos de variáveis no novo método. Para tal, um estudo sobre os parâmetros do algoritmo genético foi efetuado. Um planejamento fatorial $3^4$ foi empregado variando os níveis dos parâmetros do GA conforme apresentados na **Tabela 2.2.**

**Tabela 2.2.** Parâmetros do algoritmo genético e níveis usados no planejamento fatorial $3^4$.

| População (Nº de indivíduos) | Probabilidade de mutação | Cruzamento | Número de gerações |
|---|---|---|---|
| 50 | 1% | 40% | 40 |
| 100 | 5% | 60% | 80 |
| 200 | 10% | 80% | 160 |

Na **Figura 2.9** são apresentadas as frequências de seleção das variáveis pelo algoritmo genético nos 81 experimentos realizados para os quatro parâmetros estudados no GA. Esses gráficos demostram que todas as regiões dos espectros foram contempladas no modelo e os mais diversos tipos de variáveis foram usados para avaliar a metodologia proposta.

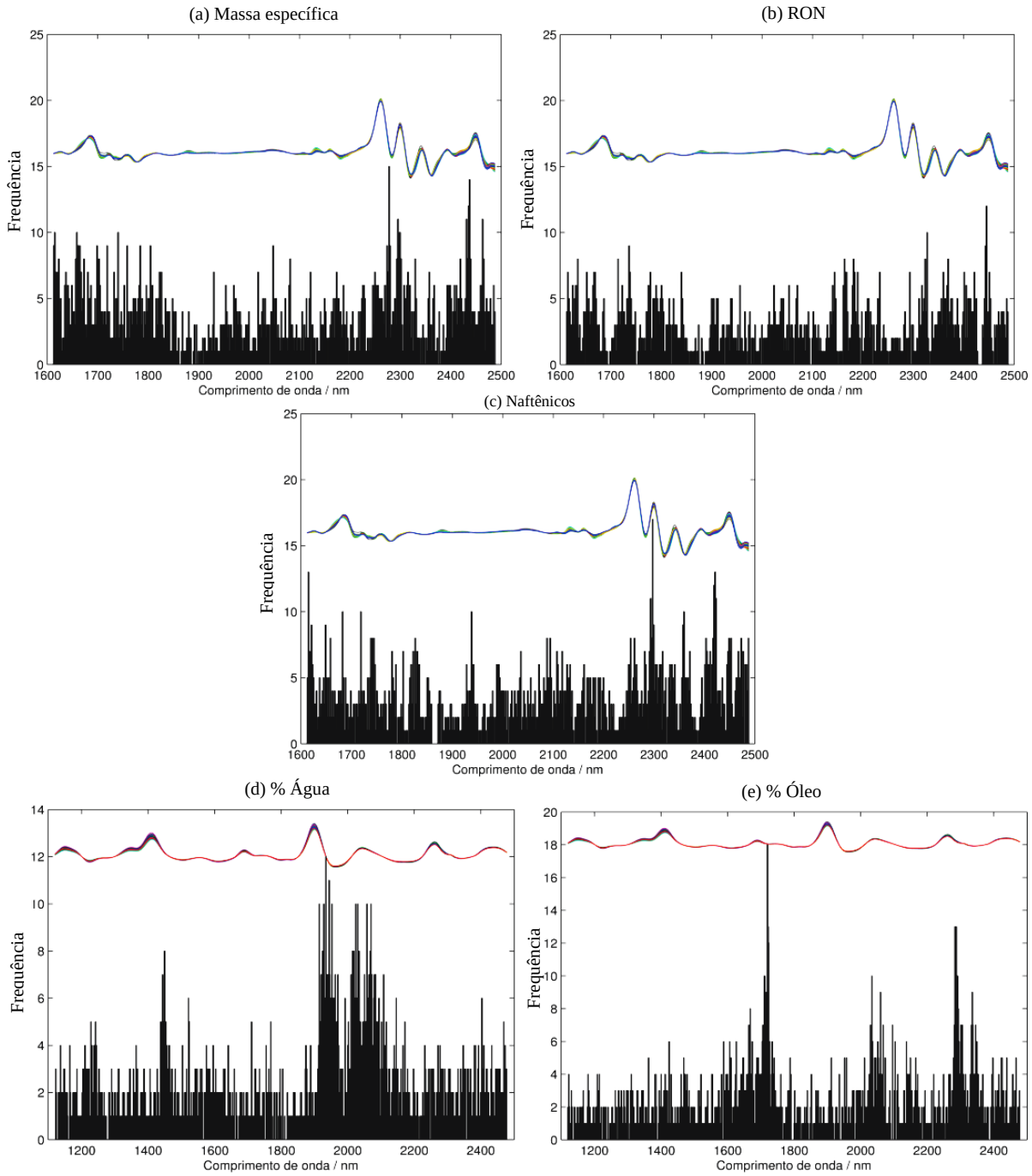



**Figura 2.9.** Frequência com que as variáveis foram selecionadas pelo GA nos 81 experimentos do planejamento $3^4$. (a) SM, (b) RON, (c) Naftênicos, (d) % Água e (e) % Óleo.

Após as variáveis serem selecionadas em cada experimento pelo GA, os resultados foram estabelecidos em termos dos valores de RMSEP obtidos no equipamento secundário ($RMSEP_{P\text{-}T}^{S}$ com o novo método e $RMSEP_{P}^{S}$ sem o novo método). Com esse parâmetro de desempenho, a análise de variância (ANOVA) foi usada para avaliar as influências dos fatores do GA e do uso do novo método na resposta (RMSEP) por meio de um teste F.

A probabilidade calculada a partir do valor do ponto F, para os efeitos principais e as interações entre os fatores, foi obtida para avaliar a significância dos parâmetros e do novo método na resposta. Como as probabilidades calculadas para todos os fatores (exceto o novo método) foram superiores ao valor de probabilidade crítico 0,05, pode-se concluir que para todos os casos (SM, ROM, naftênico, água e óleo) a análise de variância indica que somente o fator com influência significante nos resultados foi o do novo método.

### 2.5.4 Comparação entre o novo método e a regressão *ridge*

Uma investigação adicional foi realizada para comparar o novo método com a regressão *Ridge*, em vista pela similaridade da solução obtida para a regressão robusta apresentada na **Equação (27)**.

Na **Figura 2.10** os valores singulares das matrizes $\mathbf{X_a}^T\mathbf{X_a}$ e $(\mathbf{X_a}^T\mathbf{X_a}+\mathbf{R_a})$ são comparados. Os valores de $\mathbf{R_a}$ são obtidos pelo procedimento de correção univariado com $N_{trans}$ = 10 amostras de transferência. Como pode ser visto, o uso de $\mathbf{R_a}$ resulta em um pequeno aumento dos valores singulares, que é um efeito similar ao introduzido por $\lambda\mathbf{I}$ no termo regularização na RR. Portanto, isto deve ser considerado como um efeito indireto, porque a regressão robusta não foi especificamente desenvolvida para aumentar os valores singulares da matriz $\mathbf{X_a}^T\mathbf{X_a}$. Na verdade, o seu principal propósito é proporcionar robustez contra imperfeições do procedimento de correção univariado.

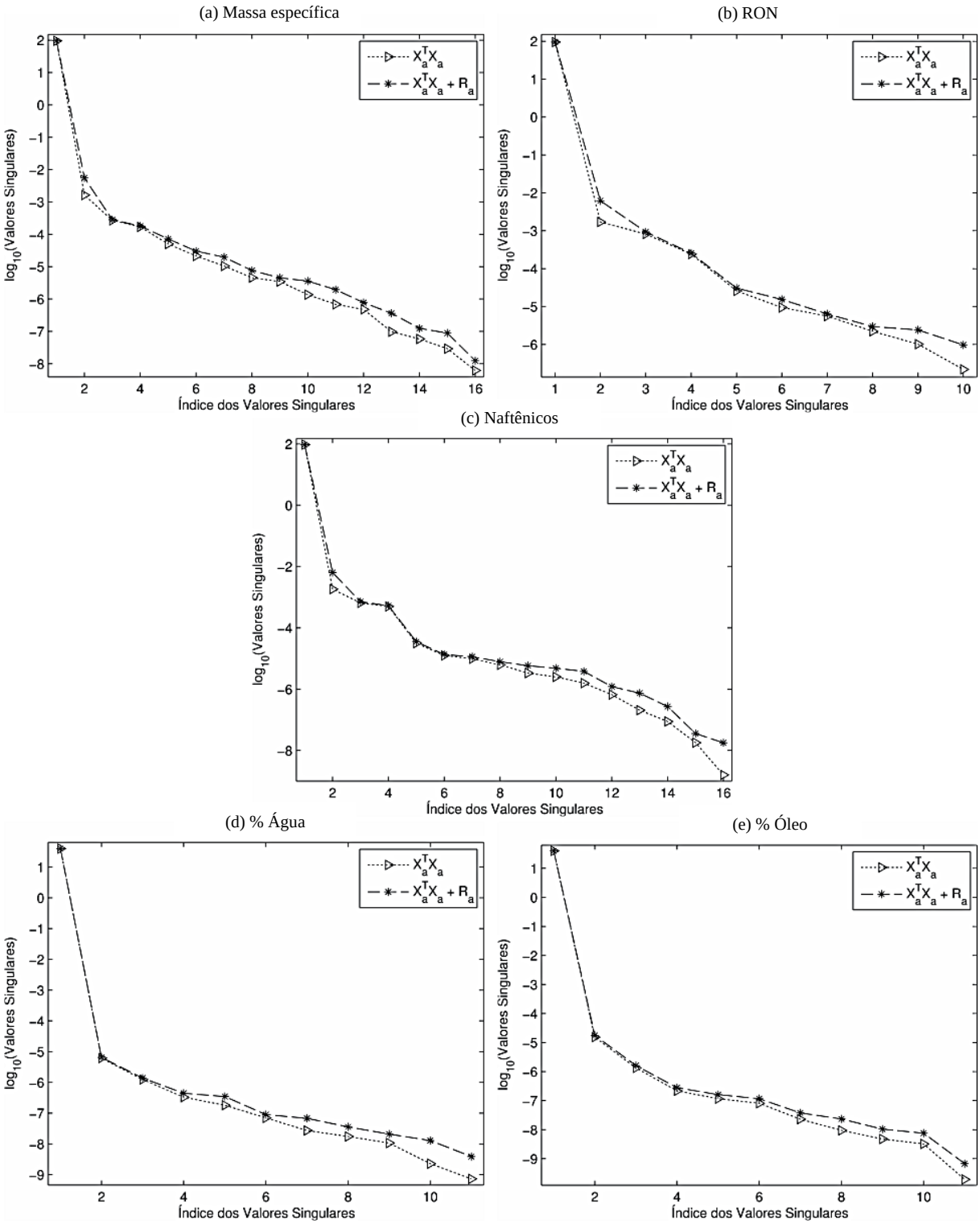



**Figura 2.10.** Valores singulares de $\mathbf{X_a}^T \mathbf{X_a}$ e ($\mathbf{X_a}^T \mathbf{X_a} + \mathbf{R_a}$) para (a) SM, (b) RON, (c) Naftênicos, (d) % Água e (e) % Óleo. A matriz $\mathbf{R_a}$ foi obtida usando dez amostras de transferência.

Na **Figura 2.11** são comparados os resultados do novo método (usando $N_{trans}$ = 10 amostras de transferência) e a regressão *ridge* (com e sem o procedimento de correção univariado). Neste caso, o parâmetro $\lambda$ da RR foi variado a fim de observar o melhor valor por meio dos $RMSEP_P^S$ resultantes (sem UVC) ou $RMSEP_{P\text{-}T}^S$ (com UVC) no instrumento secundário.

Em todos os casos, os melhores resultados da RR foram obtidos pelo uso da correção univariada (ponto mínimo da curva pontilhada), que está próximo aos resultados produzidos pelo novo método (linha horizontal). Portanto, o resultado da RR exibe uma substancial variação pela mudança no parâmetro $\lambda$ e pode realmente tornar-se muito pior em comparação ao novo método. De fato, o valor de $\lambda$ não deve ser escolhido com base nos valores resultantes de $RMSEP_{P\text{-}T}^S$, porque na prática somente as amostras de transferência possuem medidas disponíveis no equipamento secundário. Neste contexto, o novo método tem a vantagem de dispensar a necessidade de ajustar um parâmetro de regularização, uma vez que a matriz $\mathbf{R_a}$ é obtida de maneira sistemática pela correção univariada das amostras de transferência.

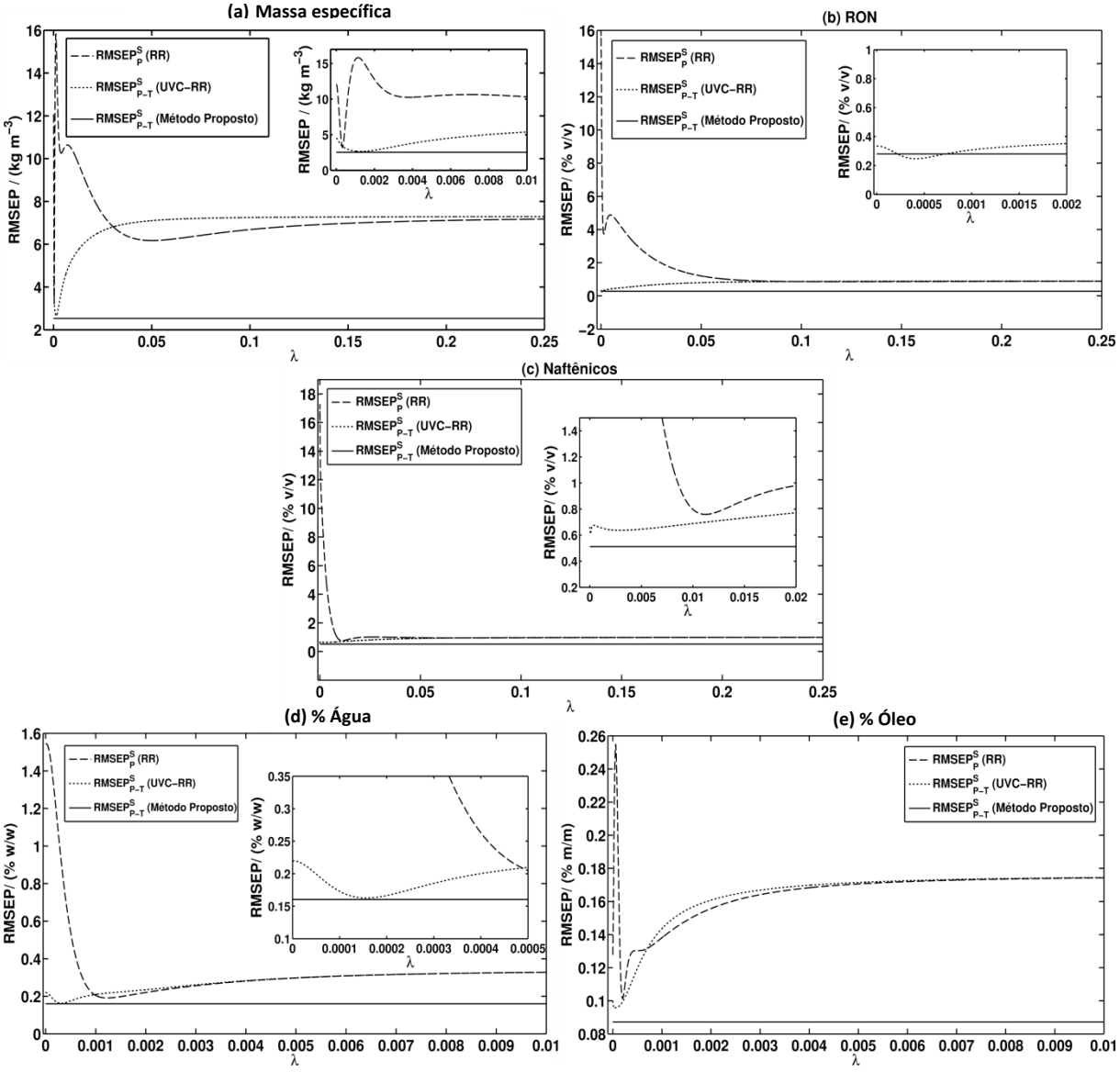



**Figura 2.11.** Resultados da regressão *Ridge* em função de λ para (a) SM, (b) RON, (c) Naftênicos, (d) % Água e (e) % Óleo. O procedimento de correção univariada foi realizado usando dez amostras de transferência.

## 2.6 Conclusões

Este estudo apresentou um novo método para transferência de calibração, que combinou o uso da correção univariada com a regressão robusta. Os resultados obtidos com o novo método demostraram uma melhoria significativa e sistemática em relação ao método de correção univariada sem a regressão robusta e aos obtidos com o método PLS usando os espectros completos com a transferência de calibração por meio do PDS.

O estudo realizado com os parâmetros do GA permitiu avaliar as variáveis ao longo de todo o espectro, corroborando com a eficiência dos diversos tipos de variáveis selecionadas pelo GA.

Na comparação realizada com a regressão *ridge*, o novo método de transferência de calibração apresentou resultados superiores mesmo considerando os diversos parâmetros de ajuste $\lambda$ usados. O novo método tem ainda a vantagem de dispensar a necessidade de ajustar um parâmetro de regularização, uma vez que a matriz $\mathbf{R_a}$ é obtida de maneira sistemática pelo uso da correção univariada das amostras de transferência.

# CAPÍTULO 3: CLASSIFICAÇÃO DE SEMENTES DE ALGODÃO

## 3.1 Introdução

O algodão é uma *commodity* agrícola de grande relevância econômica no cenário mundial. Produzido em mais de 60 países, o setor envolve uma grande quantidade de pessoas e gera cerca de U$$ 12 bilhões anuais (ABRAPA, 2016). Neste cenário, o Brasil possui um lugar de destaque, ocupando a quinta posição entre os países produtores (ABRAPA, 2016; BELTRÃO; AZEVEDO, 2008). Diante de dados tão expressivos que reforçam a grande demanda por algodão, o desenvolvimento de formas de aumentar a produtividade se torna cada vez mais necessário e dependente de novas tecnologias.

O uso de tecnologias genéticas tem sido explorado visando melhorar algumas características das plantas como: maior tolerância à seca (BOHNERT; JENSEN, 1996), a herbicidas (BAYLEY et al., 1992) e ao ataque de pragas (SHOWALTER et al., 2009; TORRES; RUBERSON; WHITEHOUSE, 2010). Consequentemente, estes melhoramentos reduzem as aplicações de defensivos agrícolas (TORRES et al., 2010) e aumentam o rendimento, garantindo a oferta de produtos para uma população em expansão (CHEN et al., 2016).

A obtenção de sementes de alta qualidade genética é umas das primeiras etapas para o aumento do rendimento da lavoura. A germinação, o vigor e as características especiais das fibras no algodão são alguns dos parâmetros usados para avaliar a qualidade destas sementes (HAN et al., 2014). Dessa forma, o desenvolvimento de sementes e registro nos órgãos competentes atendendo as leis de proteção intelectual e comercial (CARVALHO; BIANCHETTI; REIFSCHNEIDER, 2009) têm sido um dos principais desafios.

A fim de realizar a identificação dessas sementes, os desenvolvedores incorporam vários descritores que permitem identificá-las em casos de litígio ou em uma provável contaminação de lotes como, por exemplo, sementes convencionais contaminadas por sementes transgênicas. Frequentemente, informações de características da planta como: germinação, vigor, informações de produtividade, formato de folhas, cor das flores, tamanho

da planta, composição química, dentre outros (BRITO et al., 2014; OLIVEIRA; DIAS; DANTAS, 2011) são usados como descritores para auxiliar na identificação. Além desses recursos são empregadas ferramentas avançadas de biologia molecular associando marcadores específicos como proteínas, enzimas, sequências de aminoácidos específicos e DNA genômico (DONG et al., 2014; LEVI et al., 2008; MYLONAS et al., 2014; SIMON et al., 2012; WANG et al., 2014). Diante do exposto, percebe-se que estas técnicas apresentam alguns inconvenientes como: não preservar a amostra e ter um elevado tempo de análise. Além disso, as que utilizam ferramentas de biologia molecular possuem alto custo e estão disponíveis apenas em poucos laboratórios especializados, geralmente distantes dos centros de produção.

Uma alternativa a tais métodos pode ser encontrada no emprego da espectrometria de reflectância difusa na região do infravermelho próximo em conjunto com técnicas de modelagem apropriadas. Alguns estudos empregando tal estratégia vêm sendo relatados na literatura.

Yang & Sun (2012) classificaram a dureza de sementes de *Glycyrrhiza uralensis Fisch* empregando o método de aprendizagem semi-supervisionado em máquina de vetores de suporte (SVM: *Support Vector Machine*) e a espectroscopia no infravermelho próximo. O modelo proposto utiliza um método de programação robusta com base na diferença de funções convexas. Os resultados obtidos foram satisfatórios e melhores quando comparados com o do método SVM.

Lee & Choung (2011) desenvolveram um método para classificar soja geneticamente modifica e não modificada usando espectros NIR adquiridos em sementes individuais. O registro dos espectros foi realizado em um suporte que permitiu que a radiação refletida pelos lados da semente fosse direcionada para o detector. Após isso, uma análise de componentes principais (PCA) foi realizada para a avaliação das amostras e em seguida um modelo PLS-

DA (análise discriminante por mínimos quadrados parciais) foi construído. Os resultados obtidos pelo PLS-DA proporcionaram uma taxa de classificação correta de 97%.

Vitale et al. (2013) classificaram sementes de pistache (*Pistacia vera* L.) com relação a seis diferentes origens geográficas a partir de modelos SIMCA (modelagem independente flexível por analogias de classe) e PLS-DA. Foram analisadas 483 amostras e os espectros foram registrados entre 10.000 e 4000 $cm^{-1}$, em sementes cortadas ao meio de forma longitudinalmente, no modo de reflectância. Os resultados demonstraram que mais de 95% das amostras de validação foram corretamente classificadas utilizando o PLS-DA. Resultados similares foram obtidos com a técnica SIMCA.

Santos et al. (2014) desenvolveram um método usando a espectrometria NIR para a discriminação não-destrutiva de sementes de mamona para as principais cultivares brasileiras (BRS Nordestina e BRS Paraguaçu). Para esta finalidade, duas técnicas de classificação são comparadas, o SIMCA e o PLS-DA. Os melhores resultados foram obtidos usando PLS-DA, que classificou corretamente todas as amostras de teste. Estes resultados sugerem que o novo método é promissor na identificação de genótipos de sementes de mamona.

Como é possível perceber, em alguns desses trabalhos a classificação foi realizada após o uso de estratégias para aquisição dos espectros de uma área maior da semente (LEE; CHOUNG, 2011) ou registrando espectros em mais de um ponto na semente (SANTOS et al., 2014). Isso ocorre, porque o NIR convencional não permite: realizar uma varredura de uma grande área da semente sem uso de acessórios, analisar diversas sementes individuais ao mesmo tempo, monitorar compostos e doenças que ocorrem em áreas específicas das sementes (ESTEVE AGELET; HURBURGH, 2014).

As tecnologias de imagens digitais apresentam-se como alternativas para a classificação e identificação não destrutivas de cultivares (VILAR et al., 2014). Nas imagens digitais em um sistema RGB (*Red, Green e Blue*), por exemplo, a cor associada a um determinado pixel é

usada para fins de classificação. Enquanto que, a tecnologia de imagens hiperespectrais NIR (HSI: *Hyperspectral Imaging ou Hyperspectral Image*) adiciona em cada pixel um atributo de sua informação espectral. A HSI-NIR pode ser vista como uma técnica que combina as vantagens da espectroscopia e das imagens digitais (GAO et al., 2013). A estrutura hiperespectral é formada por um espectro em cada pixel e uma imagem para cada comprimento de onda. Este sistema pode fornecer características de composição, construção de um objeto e suas distribuições espaciais.

Diante dessas vantagens, esta técnica vem sendo aplicada com sucesso em estudos qualitativos envolvendo sementes, tais como, na classificação de trigo em relação aos tipos de sementes (MAHESH et al., 2008), identificação da origem geográfica (GAO et al., 2013), detecção de impurezas em aveia (SERRANTI et al., 2013), danos causados por insetos (RIDGWAY; CHAMBERS, 1998; SINGH et al., 2009; SINGH et al., 2010), defeito nas cores dos grãos de trigo e caracterização dos grãos e sementes individuais (RODRÍGUEZ-PULIDO et al., 2013).

Com objetivo de identificar cultivares de arroz chinesas usando imagens hiperespectrais NIR, Kong et al. (2013) recorreram a técnicas de classificação usando os métodos PLS-DA, SIMCA, K-vizinhos mais próximos (KNN: *K-Nearest Neighbor*), SVM e um algoritmo de aprendizado chamado de *Random Forest* (RF). Os modelos PLS-DA e KNN obtiveram uma taxa de classificação correta de 80%, enquanto os modelos SIMCA, SVM e RF proporcionaram 100% de classificação correta. Doze comprimentos de onda foram selecionados pelos coeficientes de regressão do modelo PLS-DA. Com base apenas nestes comprimentos de onda, o PLS-DA, KNN, SVM e modelos RF foram construídos. Todos os modelos restritos as variáveis selecionadas (exceto PLS-DA) produziram taxas de classificação superiores a 80%.

Zhang et al. (2012) desenvolveram um método para discriminar diferentes variedades de sementes de milho usando imagem hiperespectral no visível e no infravermelho próximo (Vis-NIR). As imagens hiperespectrais de 330 amostras de seis variedades de sementes de milho foram adquiridas na faixa de 380-1.030 nm. Após isso uma análise de componentes principais e a análise de componentes principais em *kernel* (KPCA) foram utilizadas para realizar uma triagem inicial dos dados. Três comprimentos de onda foram selecionados com base na PCA. Em seguida quatro variáveis de textura incluindo contraste, homogeneidade, energia e correlação foram extraídas a partir dos níveis de cinza da matriz de co-ocorrência (GLCM: *gray level co-occurrence matrix*) de cada imagem obtida para cada comprimento de onda. Com base nesses dados, vários modelos para identificação de sementes de milho foram construídos utilizando os métodos LS-SVM e redes neurais. A taxa de classificação alcançada no modelo PCA-GLCM-LS-SVM foi de 98,89%.

Neste contexto, o potencial da imagem hiperspectral no infravermelho próximo para a classificação de sementes de algodão foi demonstrada em um estudo envolvendo a discriminação de quatro variedades de sementes de algodão. Os resultados serão comparados com os obtidos pelo NIR convencional

## 3.2 Objetivos

### 3.2.1 Objetivo geral

Este trabalho propõe o desenvolvimento de uma metodologia analítica, rápida, não destrutiva e não invasiva baseada no uso da espectroscopia de imagem hiperespectral no infravermelho próximo aliada aos métodos de reconhecimento de padrões para a classificação de sementes individuais com respeito à variedade.

### 3.2.2 Objetivos específicos

- Realizar a seleção e processamento das imagens hiperespectrais;
- Comparar os modelos de classificação SPA-LDA e PLS-DA usando as medidas de imagens hiperespectrais;
- Comparar os modelos de classificação SPA-LDA e PLS-DA usando as medidas de espectros no NIR convencional;
- Comparar os resultados obtidos no NIR convencional com os obtidos usando as imagens hiperespectrais no NIR.

## 3.3 Fundamentação teórica

### 3.3.1 Imagens hiperespectrais na região do infravermelho próximo

Uma imagem digital pode ser considerada como uma representação de um conjunto de dados em um mapa que guarda a transcrição da informação espacial (SOLOMON; BRECKON, 2010). Esse mapa é capaz de representar com detalhes um objeto pertencente ao espaço 3D, em um plano bidimensional, a partir de cores que são obtidas por meio da combinação de três componentes univariados. As características destas componentes dependem do sistema de cor no qual a imagem digital foi adquirida. Alguns exemplos de sistemas de cores são o RGB, HSV (*Hue, Saturation and Value*), CMYK (*Cyan, Magenta, Yellow and Black*), CIELAB, e outros (GELADI; GRAHN; BURGER, 2007; VIDAL; AMIGO, 2012).

O sistema RGB baseia-se no mecanismo de formação de cores do olho humano (GELADI et al., 2007; VIDAL; AMIGO, 2012), em que combinações em diferentes níveis, das radiações de comprimentos de onda vermelho, verde e azul fornecem diferentes cores. Em um sistema digital de 8 bits, a intensidade de cada cor pode ser representada por valores inteiros entre 0 e 255. Dessa forma, a combinação dos canais vermelho, verde e azul produz 16,78 x $10^6$ possibilidades de cores.

As imagens hiperespectrais NIR podem ser vistas como a junção das imagens digitais com a espectroscopia NIR (CHENG; SUN, 2014). Neste caso o número de combinações entre os canais é infinitamente maior que no sistema RGB, uma vez que a imagem pode ser formada por mais de três canais. De fato, Geladi et al. (2007) definem que uma imagem hiperespectral deve conter muitos comprimentos de ondas, e frequentemente esse número deve ser maior que 100. Estes canais podem ainda assumir infinitos valores (por exemplo: absorbância e reflectância), não ficando restrito a uma faixa.

No registro dos sinais as informações espectral e espacial são gravadas simultaneamente, e como resultado é criada uma estrutura tridimensional (x, y, λ) (SILVA et al., 2014), também denominada por cubo hiperespectral, imagem química ou imagem espectral (CHENG; SUN, 2014; SILVA et al., 2014). As dimensões x e y acomodam a informação espacial e a terceira dimensão z fornece a informação espectral (uma representação é apresentada na **Figura 3.1a**). Cada imagem é formada por um conjunto de pontos conhecidos por pixels (a multiplicação entre x e y fornece a quantidade deles na imagem) e quanto maior a quantidade de pixels em uma determinada área maior a resolução dessa imagem.

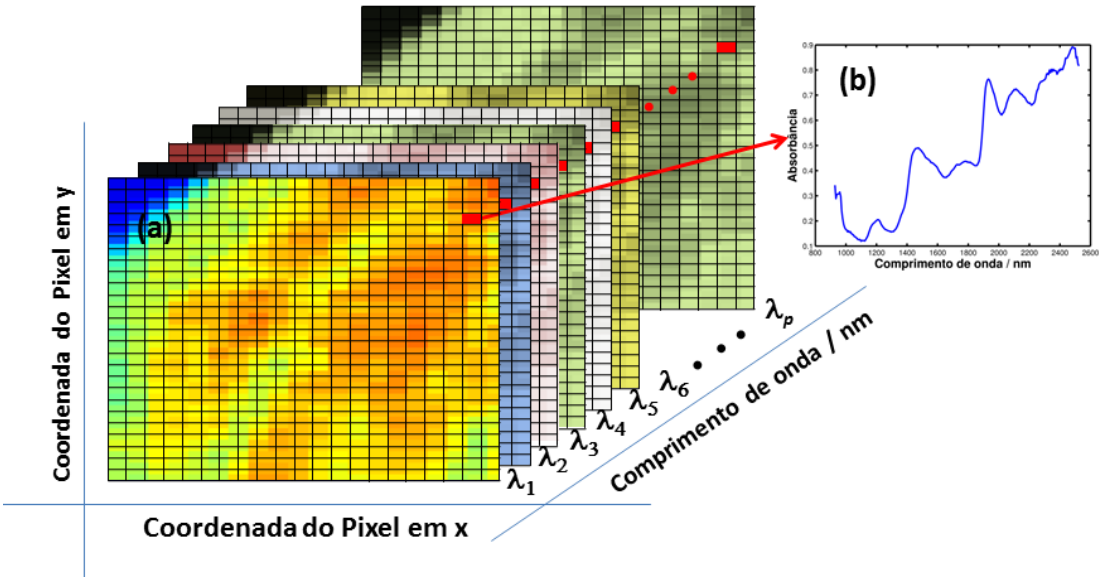



**Figura 3.1.** Estrutura tridimensional do cubo hiperespectral. (a) Em cada comprimento de onda uma imagem pode ser obtida e (b) cada pixel pode ser representado por um espectro.

O cubo de dados pode ser visto como uma série de espectros (um em cada pixel) ou como uma linha de imagens (uma para cada λ). A seleção de um único pixel (com coordenada x, y) ao longo da dimensão z irá mostrar o espectro gravado nesta localização espacial específica, que fornece a assinatura espectral dos componentes químicos presentes nessa parte

exata da amostra (**Figura 3.1b**). Se for selecionado um plano de imagem (xy) e um comprimento de onda específico (valor z), será mostrado os valores de intensidade para todos os pixels e assim será formada uma imagem de um único comprimento de onda.

O cubo de dados hiperespectrais pode ser visualizado como uma imagem utilizando uma escala de cinza ou uma escala de cor para representar a intensidade (GELADI et al., 2007), Por exemplo: o valor médio da intensidade da absorbância em cada pixel. Tal imagem em pseudocores não revela por si só nada sobre a composição química da amostra.

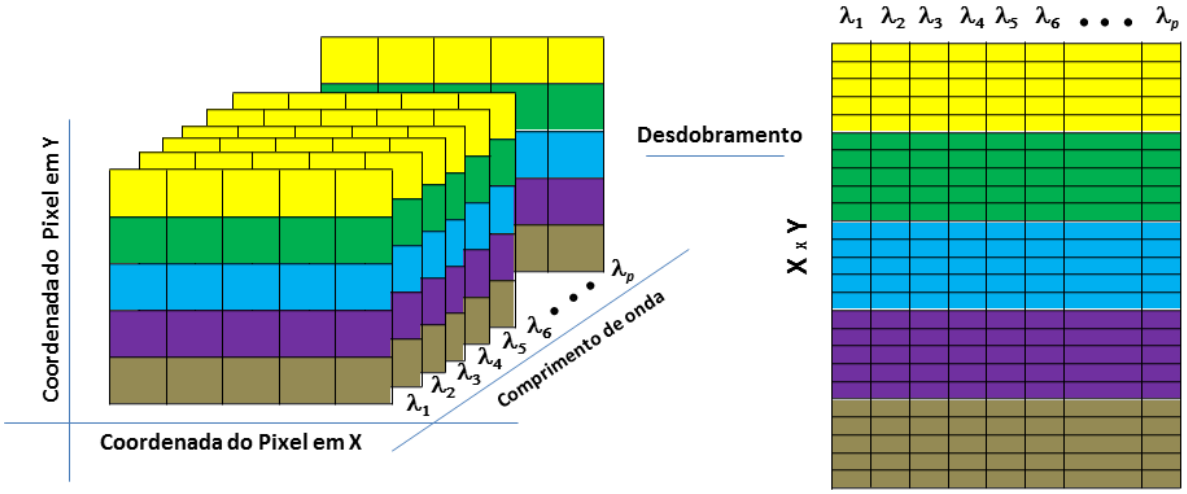



**Figura 3.2.** Ilustração do desdobramento de uma imagem hiperspectral de 25 pixels e *p* comprimento de ondas. Adaptado de Amigo et al. (2013).

As imagens hiperspectrais podem ser desdobradas (ver **Figura 3.2**), ao longo de cada comprimento de onda, de modo que cada pixel é acomodado em uma linha de uma nova matriz de dimensões iguais a xy objetos e *p* variáveis. Essa simples estratégia permite que métodos de primeira ordem, tais como a PCA, possam ser utilizados para extrair informações de uma dada imagem hiperspectral.

#### 3.3.1.1 Instrumentação do espectrômetro para imagens HSI-NIR

Os espectrômetros convencionais são formados por uma fonte de luz, um monocromador ou sistema de filtros, uma unidade de apresentação da amostra e um sistema de detecção (GELADI et al., 2007). A forma de obtenção dos espectros nos equipamentos NIR hiperespectrais é muito semelhante aos espectrômetros convencionais. Dessa forma, a instrumentação para aquisição de dados NIR hiperespectral pode ser descrito como o acoplamento de um espectrômetro NIR a uma câmera microscópica por meio de um sistema de *hardware* adicional necessário para aquisição da informação espacial (GELADI et al., 2007). Existem três configurações básicas para obter o cubo de dados hiperespectrais brutos: O mapeamento da imagem por ponto, em linha e a partir de um plano (AMIGO et al., 2013; GELADI et al., 2007).

Os princípios do mapeamento por ponto e em linha são muito similares. Estes se baseiam num modo de aquisição passo a passo (AMIGO et al., 2013). Inicialmente o tamanho da área amostrada e o número de pixels que irá compor a imagem são definidos. Em cada etapa os espectros de reflectância são adquiridos. Se o mapeamento em ponto é usado, apenas um espectro por vez é adquirido em cada pixel. No mapeamento em linha, diversos espectros são coletados a partir de uma linha de pontos adjacentes. Entre cada etapa de aquisição a amostra é deslocada em uma etapa de posicionamento a fim de obter outra sequência de espectros (AMIGO et al., 2013). Desta forma o cubo de dados hiperespectral é construído a partir do registro dos espectros de todos os pixels da área amostral definida.

No mapeamento a partir de um plano o sistema dispensa partes móveis e é capaz de gerar uma imagem em cada comprimento de onda, em vez de registrar os espectros em cada pixel. Esse tipo de mapeamento mede os valores de intensidade de absorção NIR na área definida para a amostra em cada comprimento de onda por vez (AMIGO et al., 2013; GELADI et al., 2007). Para tal, essa técnica usa *arrays* bidimensionais de plano focal (FPA: *Focal Plane Array*), que são câmeras compostas por milhares de detectores individuais em um

plano. Nesse tipo de mapeamento o tamanho e o número de pixels são fixados, portanto, a área amostral é definida pela óptica de ampliação. A terceira dimensão do cubo é obtida alterando o comprimento de onda da luz usando filtros ajustáveis. Como exemplo os filtros acústicos sintonizáveis. Os valores de intensidade de absorção NIR são medidos em cada pixel pelo FPA e em cada comprimento de onda individual, que é alterado sequencialmente por um filtro ajustável (GELADI et al., 2007).

### 3.3.2 Espectroscopia no infravermelho próximo

A fonte de luz usada nos espectrômetros de imagens hiperespectrais NIR é proveniente da radiação no infravermelho próximo. Essa região é localizada entre o infravermelho médio, iniciando em 4000 $cm^{-1}$, e a região do visível, até 12800 $cm^{-1}$ (faixa em comprimento de onda de 2500 nm até 780 nm). As absorções na região do NIR ocorrem principalmente devido à sobretons e combinações de vibrações moleculares fundamentais da região infravermelho médio entre as ligações X-H (X = O, N, C, S). As bandas de absorção são 10-1000 (PASQUINI, 2003) vezes mais fracas em relação às bandas fundamentais no infravermelho médio (MIR: *mid-infrared*) e são caracterizadas por serem muito alargadas e sobrepostas (PASQUINI, 2003).

Ao contrário da espectroscopia no MIR, as bandas de absorção no NIR não são adequadas para fins de identificação de compostos, pois os espectros no NIR, em matrizes com composição química semelhante, podem ser muito similares em estrutura e dessa forma são difíceis de serem resolvidos. Por outro lado, no NIR é possível realizar medidas em amostras com alto teor de água e em amostras sólidas, oferece uma boa penetração da radiação (PASQUINI, 2003).

As medidas de reflectância difusa em amostras sólidas são consideradas a base das medidas no NIR. Esta técnica mede a luz refletida a partir da superfície da amostra, que é composta por uma componente difusa e uma especular. A componente especular é a luz refletida pela superfície da amostra em vez de absorvida e por isso contém pouca ou nenhuma

informação química sobre o sistema em estudo (BLANCO et al., 1998). Já a componente difusa traz informações sobre a composição devido à interação que ocorre entre a luz e a amostra em varias profundidades e diferentes localizações. Dessa forma é o componente difuso, que garante relação com a composição de interesse (BLANCO et al., 1998).

A contribuição especular pode ser minimizada basicamente de duas formas. Primeiro, pelo posicionamento do detector em relação à amostra. Segundo, usando um pré-processamento adequado para remover esta contribuição (BLANCO et al., 1998). No pré-processamento, recorre-se ao uso da refletância relativa (que aqui é chamado de R na **Equação (30)**), que é a razão da luz refletida em cada comprimento de onda, por uma amostra e um padrão não absorvente, estável e com alta reflectância em toda faixa do NIR.

$$R(\lambda) = \frac{S_{amostra}(\lambda)}{S_{padrão}(\lambda)} \quad \textbf{(30)}$$

Para descrever o comportamento entre a concentração e refletância, Kubelka-Munk propuseram um modelo empírico que relaciona estas duas variáveis conforme apresentado na **Equação (31)**.

$$f(c) = \frac{(1-R)^2}{2R} \quad \textbf{(31)}$$

No entanto, esta equação é raramente aplicada sendo substituída, principalmente pelo efeito prático, pela **Equação (32)**.

$$f(c) = \log \frac{1}{R} \quad \textbf{(32)}$$

A **Equação (32)** não se afasta muito da predição de Kubelka-Munk, para pequenas mudanças na refletância (R) (que são comuns em muitas aplicações), pode-se assumir um comportamento linear com a concentração do analito (PASQUINI, 2003).

### 3.3.3 Classificação multivariada

Em quimiometria muitos problemas podem ser reformulados para fins de classificação, que envolve a tentativa de utilizar as medidas sobre um conjunto de amostras e verificar se existe relação entre algumas propriedades das amostras e os dados analíticos. Em seguida, para determinar que relação seja esta, recorre-se a um modelo matemático, que após ser criado, é usado para determinar a origem de uma amostra desconhecida (BRERETON, 2009b).

A classificação envolve determinar se uma amostra pertence a um ou mais grupos predeterminados (VARMUZA; FILZMOSER, 2009). Estes métodos são muitas vezes chamados métodos supervisionados, pois eles exigem algum tipo de informação sobre as origens das amostras utilizadas para construir o modelo, com antecedência, ao contrário dos métodos não supervisionados, tais como a análise de componentes principais (PCA: *Principal Component Analysis*).

Recentemente, problemas de classificação estão ganhando cada vez mais interesse, como por exemplo, em aplicações envolvendo amostras: de vinhos (SERRANO-LOURIDO et al., 2012), de células cancerosas (KARIMI; HEMMATEENEJAD, 2013), de petróleo (BORGES et al., 2010), de espécies de ervas (WONG et al., 2013), de gordura de porco (FOCA et al., 2013), de medicamentos (DA SILVA FERNANDES et al., 2012), são alguns exemplos. Nesse contexto, as técnicas de classificação multivariadas que foram usadas nesta tese serão apresentadas nas seções a seguir.

### 3.3.4 Análise de discriminante por mínimos quadrados parciais

O PLS descrito na **seção 2.3.1.3** não foi inicialmente proposto como uma ferramenta para discriminação. Contudo, foi adaptado para classificar amostras desconhecidas, em classes definidas a *priori (BARKER; RAYENS, 2003)*. Essa adaptação faz com que, o método de análise de discriminante por mínimos quadrados parciais (PLS-DA: *Partial Least Squares for Discriminant Analysis*) seja considerado um método de reconhecimento de padrões supervisionado (DE ALMEIDA et al., 2013).

Cada classe do sistema em estudo, representada em uma coluna da matriz **Y**, é definida usando a codificação 0 ou 1. Os valores iguais a 1 ao longo de uma coluna da matriz **Y** representam as amostras pertencentes à classe codificada, e os valores iguais a 0 representam as amostras de classes distintas. Em casos com mais de duas classes em estudo, é possível construir um modelo global PLS2, ou um modelo PLS1 para cada classe representada em **Y**. A definição de cada classe pode ser realizada com base em informações do sistema em estudo ou usando, por exemplo, uma PCA.

Na classificação de uma amostra desconhecida, o ideal seria que os valores preditos fossem iguais a 1, ou caso contrário, quando uma amostra não pertence a uma classe, a predição deveria ser igual a 0. Como dito, seria o "ideal", na prática isso é um caso raro de acontecer. A fim de corrigir este problema, recorre-se ao uso de um limiar. No *software Unscrambler* (CAMO, 2007) o valor 0,5 é sugerido. No entanto, alguns autores têm utilizado o teorema de Bayes como base para o cálculo deste limiar (DE ALMEIDA et al., 2013; IVORRA et al., 2013).

Em adição ao limiar, um intervalo de confiança para cada amostra de predição com determinada incerteza é obtido de modo a avaliar se o limiar de cada classe está contido (CAMO, 2007; DE ALMEIDA et al., 2013). A obtenção do intervalo de confiança pode ocorrer a partir do método *Bootstrap* (DE ALMEIDA et al., 2013), ou de acordo com o *deviation* do

*Unscrambler* (CAMO, 2007), calculado por meio do *leverage* e da variância em **X** para cada amostra.

Após ser obtido o limiar e o intervalo de confiança, a classificação é dada da seguinte forma: se o limiar da classe for menor que o valor predito e não estiver contido no intervalo de confiança, a amostra é dita pertencente à classe; caso contrário, quando o limiar for maior que o valor predito e não estiver contido no intervalo de confiança, a amostra não pertence à classe; por último, se o intervalo de confiança contiver o limiar, a amostra não poderá ser classificada com segurança nesta classe (CAMO, 2007; DE ALMEIDA et al., 2013).

Devido às dificuldades em calcular o intervalo de confiança, pode-se comparar o valor predito somente com o limiar, ou de uma maneira mais simples, pode-se atribuir a amostra à classe que possuir o maior valor predito pelo modelo (BALLABIO; CONSONNI, 2013).

### 3.3.5 Análise discriminante linear

A análise discriminante linear (LDA: *Linear Discriminante Analysis*) (NÆS, 2002; VARMUZA; FILZMOSER, 2009) proposta por Fisher em 1938, surgiu para a discriminação de dois grupos e em seguida foi estendida para ser aplicada em casos contendo mais de duas classes. Quando a classificação binaria é realizada, uma combinação linear entre as variáveis originais gera a função discriminante que promove a máxima separação entre as classes. A **Equação (33)** descreve a combinação linear:

$$y^d = b^d{}_1 x_1 + b^d{}_2 x_2 + \cdots + b^d{}_m x_m \qquad \textbf{(33)}$$

os coeficientes $b^d{}_1,\ldots, b^d{}_m$ formam o vetor de decisão (também chamado de *loading vector*). Quando os objetos são projetados nos eixos definidos por $\boldsymbol{b^d}$ os escores de Fisher $y^d{}_{1h}$ com $h = 1, \ldots, m_1$ são obtidos para o primeiro grupo de amostras e $y^d{}_{2l}$ com $l = 1, \ldots, m_2$ para o segundo

grupo de amostras. Os valores médios do primeiro e segundo grupo são denotados por $\bar{y}^d{}_1$ e $\bar{y}^d{}_2$. Dessa forma, o critério formulado para obter a melhor separação (NÆS, 2002; VARMUZA; FILZMOSER, 2009) é dado por:

$$\frac{|\bar{y}^d{}_1 - \bar{y}^d{}_2|}{S_{yd}} \rightarrow max \qquad \textbf{(34)}$$

onde, $S_{yd}$ é a raiz quadrada da variância conjunta obtida nos dois grupos. A partir disso, Fisher demonstrou (VARMUZA; FILZMOSER, 2009) que o vetor de decisão $\boldsymbol{b^d}$ é dado pela **Equação (35).**

$$\boldsymbol{b^d} = S_p^{-1}(\bar{x}_1 - \bar{x}_2) \qquad \textbf{(35)}$$

em que, $\bar{x}_1$ e $\bar{x}_2$ são os vetores das médias aritméticas dos grupos 1 e 2 para a matriz **X**, e $S_p$ é a matriz de variância conjunta. Quando mais de dois grupos estão sendo avaliados, o caso binário pode ser estendido (VARMUZA; FILZMOSER, 2009). Para tal, considere a variação entre os grupos definida por $\boldsymbol{B^d}$**.**

$$\boldsymbol{B^d} = \sum_{j=1}^{k} p_j(\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^{\mathrm{T}} \qquad \textbf{(36)}$$

em que $\boldsymbol{\mu}_j$ é a media de cada grupo e $\boldsymbol{\mu} = \sum_{j=1}^{k} p_j \boldsymbol{\mu}_j$ é a média populacional ponderada pela probabilidade a *priori*. Além disso, a matriz de covariância dentro dos grupos $\mathbf{W^d}$ é dada por:

$$\mathbf{W}^{\boldsymbol{d}} = \sum_{j=1}^{k} p_j \boldsymbol{\Sigma}_j \tag{37}$$

A matriz de $\mathbf{W}^{\mathbf{d}}$ pode ser vista como uma versão da matriz de covariância conjunta. Assumindo que as matrizes de covariância dos grupos são iguais, pode-se mostrar que os centros dos grupos são melhores separados pela maximização da **Equação (38)**.

$$\frac{\boldsymbol{b^{d}}^{\mathbf{T}} \boldsymbol{B^{d}}\, \boldsymbol{b^{d}}}{\boldsymbol{b^{d}}^{\mathbf{T}} \boldsymbol{W^{d}}\, \boldsymbol{b^{d}}} \tag{38}$$

em que $\boldsymbol{b^{d}}$, diferente de zero, é um vetor *m*-dimensional. Esta resolução é equivalente a buscar a direção de $\boldsymbol{b^{d}}$ no espaço multivariado em que a diferença entre as médias dos grupos seja maior possível em comparação com a variância dentro do grupo. Dessa maneira, o vetor $\boldsymbol{b^{d}}$ pode ser encontrado como o autovetor da matriz $\boldsymbol{W^{d^{-1}}B^{d}}$ que está associado ao maior autovalor. Este vetor é chamado de primeira variável canônica.

Uma desvantagem da LDA se deve ao fato de não ser adequada para dados em que o número de variáveis é maior que o número de amostras. Além disso, a capacidade de generalização dos modelos LDA pode ser comprometida por problemas de multicolinearidade (NÆS; MEVIK, 2001), provocando a instabilidade na classificação e levando os modelos a um desempenho de classificação inadequado. Com intuito de resolver estes problemas, recorre-se com frequência aos métodos de seleção de variáveis.

#### 3.3.5.1 Algoritmo das projeções sucessivas para classificação

O Algoritmo das Projeções Sucessivas (SPA: *Successive Projections Algorithm*) foi proposto inicialmente para calibração multivariada e em seguida estendido para fins de classificação (PONTES et al., 2005).

O SPA para classificação possui o propósito de melhorar o desempenho dos modelos construídos usando a análise discriminante linear quando problemas de multicolinearidade que prejudicam o desempenho desses modelos estão presentes (NÆS; MEVIK, 2001). Dessa forma, o termo SPA-LDA ao longo dessa tese será empregado como referência ao algoritmo usado para classificação.

Em classificação, o algoritmo SPA-LDA é dividido em duas fases. Na fase 1, a matriz de respostas instrumentais para as amostras de treinamento $\mathbf{X_{trein}}$ de dimensões ($N_{trein}$ *x p*) é centrada na média de cada classe. Em $\mathbf{X_{trein}}$ a *k*-ésima variável (*k* = 1, 2, ..., *p*) de $x_k$ é associada a *k*-ésima coluna do vetor $\mathbf{x}_k \in \Re^{N_{trein}}$. Estes vetores colunas são submetidos a uma sequência de operações de projeção que resultam na criação de *p* cadeias de variáveis. A *k*-ésima cadeia é inicializada com a variável $x_k$ e é progressivamente aumentada com variáveis que exibem a menor colinearidade com as anteriores. Devido aos graus de liberdade associados aos cálculos da média, o comprimento das cadeias de variáveis construídas na fase 1 do SPA-LDA é limitado por $N_{trein} - C$, onde $N_{trein}$ é o número de amostras do conjunto de treinamento e *C* é o número de classes envolvidas no problema.

Na fase 2, os diversos subconjuntos de variáveis, que são candidatos a serem escolhidos, são avaliados de acordo com uma função custo relacionada ao risco médio de classificação incorreta sobre o conjunto de validação. Esta função custo é definida como

$$J_{cost} = \frac{1}{N_{val}} \sum_{n=1}^{N_{val}} g_n \qquad \textbf{(39)}$$

onde

$$g_n = \frac{MD^2[\mathbf{x}_{val,n}, \bar{\mathbf{x}}(I_n)]}{\min_{I_j \neq I_n} MD^2[\mathbf{x}_{val,n}, \bar{\mathbf{x}}(I_j)]} \quad \textbf{(40)}$$

Na **Equação (40)**, o numerador $MD^2[\mathbf{x}_{val,n}, \bar{\mathbf{x}}(I_n)]$ é o quadrado da distância de Mahalanobis (DE MAESSCHALCK; JOUAN-RIMBAUD; MASSART, 2000) entre a *n-ésima* amostra de validação $\mathbf{x}_{val,n}$ (de índice de classe $I_n$) e a media $\bar{\mathbf{x}}(I_n)$ da classe verdadeira (ambos vetores linhas) calculadas sobre o conjunto de treinamento. Esta distância é dada por

$$MD^2[\mathbf{x}_{val,n}, \bar{\mathbf{x}}(I_n)] = [\mathbf{x}_{val,n}, \bar{\mathbf{x}}(I_n)]\mathbf{S}^{-1}[\mathbf{x}_{val,n}, \bar{\mathbf{x}}(I_n)]^T \quad \textbf{(41)}$$

onde **S** é a matriz de covariância conjunta sobre o conjunto de treinamento (WU et al., 1996). O denominador na **Equação (40)** corresponde ao quadrado da distância de Mahalanobis entre $\mathbf{x}_{val,n}$ e o centro da classe errada mais próxima. Um menor valor de $g_n$ indica que $\mathbf{x}_{val,n}$ está próximo do centro da classe verdadeira e distante do centro das classes remanescentes. A função custo $J_{\text{cost}}$ é definida como o valor médio de $g_n$ sobre todo o conjunto de validação ($n$ = 1, 2, ..., $N_{val}$). Portanto, a minimização de $J_{cost}$ resulta em uma melhor separação das amostras de acordo com as classes verdadeiras.

Os subconjuntos de variáveis são avaliados, quase sempre, com base em um conjunto de validação externo que não foi empregado na construção do modelo LDA (SOARES et al., 2014). Esse procedimento é adotado a fim de evitar sobreajuste, que deve ocorrer se as amostras de treinamento forem usadas diretamente para o cálculo da função custo (PONTES et al., 2012). No entanto, a escolha de um conjunto de amostras representativas para a validação não é uma tarefa fácil. E se o número total de amostras também é pequeno, a divisão em dois subconjuntos representativos pode se tornar uma tarefa quase que impossível.

Para evitar a necessidade de usar um conjunto de validação, a função custo $J_{cost}$ pode ser avaliada a partir da validação cruzada. Neste caso, o valor de $g_n$ é avaliado removendo a $n$-ésima amostra de treinamento ($\mathbf{x}_{trein,n}$) e a usando como uma amostra de validação.

$$g_n = \frac{MD^2_{-n}[\mathbf{x}_{trein,n}, \bar{\mathbf{x}}(I_n)]}{\min_{I_j \neq I_n} MD^2_{-n}[\mathbf{x}_{trein,n}, \bar{\mathbf{x}}(I_j)]} \quad \textbf{(42)}$$

em que o subscrito $-n$ em $MD^2_{-n}$ indica que a média da classe $\bar{\mathbf{x}}(I_1)$, $\bar{\mathbf{x}}(I_2)$, ..., $\bar{\mathbf{x}}(I_C)$ e a matriz de covariância conjunta $\mathbf{S}$ são calculadas sem usar $\mathbf{x}_{trein,n}$. Após repetir o cálculo de $g_n$ para $n$ = 1, 2, ..., $N_{trein}$, o resultado da função custo é dado por

$$J_{cost} = \frac{1}{N_{trein}} \sum_{n=1}^{N_{trein}} g_n \quad \textbf{(43)}$$

É possível perceber, que $J_{cost}$ calculado desta maneira pode consumir uma grande quantidade de tempo. De fato, o cálculo da distância de Mahalanobis requer a determinação de uma nova matriz inversa $\mathbf{S}^{-1}$ para cada $n$. Portanto, a operação de inversão de matriz necessita ser feita $N_{trein}$ vezes para cada subconjunto de variáveis. Por este motivo, uma alternativa para o SPA-LDA foi apresentada em Soares et al. (2014) em que uma validação interna pode ser feita a partir de $g_n$ calculada por meio de

$$g_n = \frac{MD^2[\mathbf{x}_{trein,n}, \bar{\mathbf{x}}(I_n)]}{\min_{I_j \neq I_n} MD^2[\mathbf{x}_{trein,n}, \bar{\mathbf{x}}(I_j)]} \quad \textbf{(44)}$$

Neste caso, a média das classes $\bar{\mathbf{x}}(I_1)$, $\bar{\mathbf{x}}(I_2)$, ..., $\bar{\mathbf{x}}(I_C)$ e a matriz de covariância conjunta $\mathbf{S}$ são calculadas usando todas as amostras de treinamento, incluindo $\mathbf{x}_{trein,n}$. O

resultado do custo é então calculado como na **Equação (43)**. Como é possível perceber o esforço computacional é muito menor quando comparado ao uso da validação cruzada *leave-one-out*. Isto é causado porque uma única matriz inversa $\mathbf{S}^{-1}$ é empregada para $n$ = 1, 2, ..., $N_{trein}$. No entanto, tal procedimento deve levar a sobreajuste (PONTES et al., 2012), porque o modelo de classificação é construído e validado usando as mesmas amostras. Como resultado, o custo tende a diminuir à medida que mais variáveis são incluídas no modelo.

No trabalho de Soares et al. (2014) é apresentada a seguinte alternativa para o cálculo do custo

$$J_{cost} = \frac{1}{N_{trein} - L - C} \sum_{n=1}^{N_{trein}} g_n \quad , \tag{45}$$

onde $L$ é o número de variáveis no subconjunto candidato sob avaliação. Neste caso, o denominador é o número de graus de liberdade, ao invés do número de amostras de treinamento. Portanto, um subconjunto com um grande número de variáveis deverá somente ser favorecido se o decréscimo nos valores de $g_n$ for grande o suficiente para compensar o decréscimo no denominador da **Equação (45)**.

### 3.3.6 Parâmetros de validação dos modelos de classificação

Ao construir os modelos de classificação existe a necessidade de avaliar a sua habilidade preditiva. Frequentemente, diversos autores recorrem ao uso de parâmetros como: a Taxa de Classificação Correta (VILAR et al., 2014), Sensibilidade e Especificidade (BALLABIO; CONSONNI, 2013; SANTOS et al., 2014). Todos esses parâmetros podem ser determinados usando a matriz de confusão (LAVINE, 2009).

**Tabela 3.1.** Exemplo de uma matriz de confusão.

| | Classe A (Atribuída) | Classe B (Atribuída) |
|---|---|---|
| Classe A (Experimental) | *a* | *b* |
| Classe B (Experimental) | *c* | *d* |

As linhas da matriz de confusão correspondem à classe a qual as amostras foram rotuladas, e as colunas correspondem à classe a qual as amostras foram atribuídas pelo modelo. Para entender melhor, considere um modelo de classificação com duas classes A e B (A é classe positiva e B é a classe negativa). Assim sendo, quatro possibilidades podem ser consideredas: I) As amostras da classe A são atribuídas a classe A (verdadeiro positivo); II) as amostras da classe B são atribuídas a classe B (verdadeiro negativo); III) as amostras da classe A são atribuídas a classe B (falso negativo) e IV) as amostras da classe B são atribuídas a classe A (falso positivo).

Com uso da matriz de confusão, a taxa de classificação correta é igual a:

$$\text{TCC (\%)} = \frac{a+d}{a+b+c+d} \times 100 \qquad \textbf{(46)}$$

em que *a* é o número de amostras que obedecem o caso I), *d* é a classificação realizada no caso II), b no caso III) e c em IV).

A sensibilidade mede a habilidade do modelo em detectar corretamente as amostras que pertencem à condição positiva. Ao passo que, a Especificidade quantifica a detecção correta da condição negativa. As **Equações (47) e (48)** descrevem matematicamente estes conceitos.

$$\text{Sensibilid ade} = \frac{a}{a+b} \qquad \textbf{(47)}$$

$$\text{Especifici dade} = \frac{d}{d+c} \qquad \textbf{(48)}$$

## 3.4 Experimental

### 3.4.1 Amostras de sementes de algodão

As quatro variedades de sementes de alta qualidade genética usadas foram cedidas pela Embrapa Algodão, localizada na cidade de Campina Grande, Paraíba. As sementes foram condicionadas em uma sala com temperatura (20º C) e umidade relativa (65 %) por no mínimo uma hora antes das medidas. A imagem digital de uma semente de cada variedade é apresentada na **Figura 3.3.**

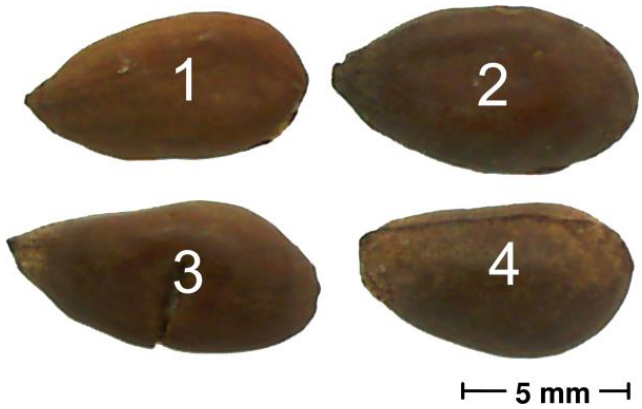



**Figura 3.3.** Imagem representativa de uma semente de cada classe em estudo.

Embora as sementes sejam aparentemente similares nas imagens da **Figura 3.3**, estas sementes são de variedades que se diferenciam de acordo com o rendimento médio da fibra, duração do ciclo (curto: 120 a 140 dias; longo: 150 a 180 dias), porte (baixo ou alto), resistência a doenças, entre outras características. Um resumo das principais características destas sementes é mostrado na**Tabela 3.2**.

**Tabela 3.2** Características das cultivares usadas nessa tese.

| Características das Cultivares | Classe 1 | Classe 2 | Classe 3 | Classe 4 |
|---|---|---|---|---|
| Porte | Médio-Alto | Médio | Alto | Médio |
| Ciclo | Médio-Tardio | Médio-Tardio | Tardio | Médio |
| Red. Fibra | 42,3 % | 40,9 % | 40,5 % | 40,9 % |
| Ramulária | Resistente | Susceptível | Resistente | Susceptível |
| Doença Azul | Resistente | Resistente | Resistente | Resistente |
| Bacteriose | Resistente | Susceptível | Resistente | Resistente |
| Arquitetura | Compacta | Semi-Aberta | Aberta | Semi-Aberta |

Para o registro dos espectros, 99 amostras da classe 1, 110 da classe 2, 101 da classe 3 e 99 da classe 4 foram selecionadas aleatoriamente para serem medidas usando uma estação de imagens hiperespectral NIR. Ao final do processo as amostras de cada classe foram devolvidas ao recipiente de armazenagem. Semelhantemente ao processo de amostragem anterior, 100 sementes das classes 1 e 4, e 99 das classes 2 e 3 foram selecionadas aleatoriamente e medidas no NIR convencional.

### 3.4.2 Aquisição das imagens hiperespectrais e pré-processamentos

As imagens hiperespectrais das amostras de sementes foram registradas em uma estação de trabalho de imagens SisuCHEMA SWIR com lentes de 30 mm para um campo de visão de 50-100 mm e sistema de mapeamento em linha operando ao longo da faixa de comprimento de onda de 928,37- 2524,01 nm e um intervalo de 6,31- 6,21 nm. A região do NIR é formada por 256 segmentados (Uma imagem para cada comprimento de onda) com a primeira imagem iniciando em 928,37 nm. As intensidades dos sinais foram registradas em cada pixel da imagem e em cada comprimento de onda. As imagens para os comprimentos de onda fora da faixa de 1100:2500 não foram usadas, devido à baixa intensidade do sinal no detector. Todas as imagens hiperspectrais foram registradas no laboratório do grupo de automação e instrumentação em química analítica do instituto de química da UNICAMP.

Os dados das imagens foram transformados em pseudo-absorbância (GAO et al., 2013) usando o sinal padrão do material de referência e o sinal no escuro, que é obtido quando a fonte de luz é desligada e a lente da câmera está completamente coberta com uma capa opaca. A transformação é realizada com os espectros da amostra, corrigidos pela corrente do detector no escuro (referência no escuro subtraído do espectro da amostra), dividindo por um espectro de reflectância do padrão também corrigido pela corrente no escuro. Após isso, a função logarítmica é aplicada ao inverso da razão obtida**.** Nas imagens registradas não foram encontrados *dead pixels* e *spikes*. Uma placa de teflon com dimensões de 10 cm de largura e 20 cm de comprimento foi usada como base para a acomodação das amostras.

As imagens hiperspectrais foram desdobradas e os espectros de reflectância foram pré-processados, usando a segunda derivada e um filtro Savitzky-Golay com polinômio de segunda ordem e janela de 21 pontos, a fim de corrigir os efeitos da linha de base e de espalhamento provenientes da medida (RINNAN et al., 2009). Esse pré-processamento foi usado para auxiliar na etapa de remoção de *background*. Na etapa de classificação, outros pré-processamentos como derivação, MSC (*Multiplicative Scatter Correction*), SNV (*Standard Normal Variate*) e suavização foram avaliados a partir do número de erros de classificação obtidos no conjunto de validação. Contudo, os melhores resultados são apresentados na seção de resultados e discussão.

#### 3.4.2.1 Remoção do background

Com objetivo de selecionar os pixels correspondentes a cada semente nas imagens NIR-HI, recorreu-se a uma análise de componentes principais. Em cada imagem, os escores em PC1 foram usados para separar os pixels de fundo dos pixels das sementes. Para isso, os escores de PC1 maiores que zero não foram usados para fins de classificação. Vale a pena esclarecer, que a escolha do limiar igual à zero será apresentada na **seção 3.5.1**.

Com objetivo de realizar a separação entre as sementes vizinhas, a seleção manual da região de interesse (VIDAL; AMIGO, 2012) foi realizada usando a *interface* "Seleção Manual da Região de Interesse" desenvolvida em Matlab R2010b. Para fins de classificação o espectro médio de todos os pixels da região de interesse (os espectros de cada semente) foi calculado.

### 3.4.3 Aquisição dos espectros no NIR convencional e pré-processamentos

Os espectros de reflectância difusa foram obtidos por meio de um espectrômetro da Foss Analytical, modelo XDS Rapid Content™ analyzer VIS-NIR, ajustado para usar uma célula circular de quartzo de diâmetro de 3 cm. Cada espectro foi adquirido como resultado de 32 varreduras na faixa de 1100-2500 com resolução de 0,5 nm. A fim de comparar os resultados com os obtidos pelo HSI-NIR foi usado um *boxcar* com uma janela de 12 pontos ao longo das variáveis. Todos os espectros foram registrados no laboratório avançado de tecnologia de química da EMBRAPA algodão.

Com intuito de diminuir os efeitos da linha de base e do espalhamento, pré-processamentos como MSC, SNV, suavização e derivação com filtro *Savitzky-Golay* foram avaliados. O melhor resultado é apresentado na seção de resultados e discussão.

### 3.4.4 Remoção de outlier

As amostras com espectros anômalos foram removidas do conjunto total dos dados registrados nos dois equipamentos. Sete amostras foram removidas como *outlier* no HSI-NIR e duas amostras no NIR convencional. Para realizar essa tarefa, o *leverage* e a soma quadrática dos resíduos (*Q*) obtidos por meio de uma PCA forneceram as informações necessárias para embasar essa remoção. O limiar adotado foi de três vezes o valor médio para o *leverage* (NÆS, 2002). Para a soma quadrática dos resíduos, a equação de Jackson & Mudholkar (1979) foi usada como critério (BRERETON, 2009a).

### 3.4.5 Escolha das amostras de treinamento, validação e teste

Os conjuntos de treinamento, validação e teste foram selecionadas usando o algoritmo Kennard-Stone para que as amostras mais externas sejam dispostas no conjunto de treinamento. Esse algoritmo foi aplicado a cada classe por vez. Na **Tabela 3.3** é apresentado o número de amostras usadas em cada conjunto.

**Tabela 3.3.** Número de amostras de treinamento, validação e teste medidas nos dois equipamentos.

| Dados | $N_{trein}$ | $N_{val}$ | $N_{test}$ |
|---|---|---|---|
| HSI-NIR | 204 | 99 | 99 |
| Convencional NIR | 202 | 97 | 97 |

### 3.4.6 Procedimento de modelagem e software utilizado

Os modelos foram construídos com base no conjunto de treinamento. Ao passo que as amostras de validação foram usadas para guiar a escolha das variáveis no SPA-LDA e variáveis latentes no PLS-DA. As amostras de teste foram usadas para realizar uma avaliação final dos modelos. Todos os cálculos foram realizados usando rotinas *Lab-made* implementadas no software Matlab R2010b.

No PLS-DA o número de variáveis latentes foi determinado de forma a minimizar o número total de erros tipo I (amostras não incluídas na sua classe) e tipos II (amostras incluídas em uma classe errada), no conjunto de validação. Para construir o modelo de classificação, o PLS2 foi usado, e em seguida o modelo resultante foi usado para classificar as amostras de validação e teste em cada classe. A decisão de atribuir uma amostra a uma determinada classe é realizada com base no valor estimado para cada classe (BALLABIO; CONSONNI, 2013).

### 3.4.7 Avaliação dos modelos de classificação

Na seção seguinte, os modelos de classificação SPA-LDA e PLS-DA serão avaliados em termos da taxa de classificação correta (TCC) e dos parâmetros de desempenho sensibilidade e especificidade.

## 3.5 Resultados e discussão

### 3.5.1 Seleção da região de interesse

Nesta seção uma abordagem para a seleção dos pixels de cada semente nos dados HSI-NIR é demonstrada. Para realizar essa tarefa, na **Figura 3.4a** é apresentado um conjunto de amostras de sementes de algodão da classe 1 representada em uma imagem em pseudocores sem nenhum pré-processamento. Com a finalidade de corrigir os efeitos de espalhamento, e consequentemente facilitar a seleção da região de interesse, a segunda derivada com filtro Savitzky-Golay, polinômio de segunda ordem e janela de 21 pontos foi usado.

Na **Figura 3.4b** é apresentado à imagem após o pré-processamento ser aplicado. Como é possível perceber, com esse tratamento os pixels correspondentes às sementes são realçados em relação aos pixels de fundo.

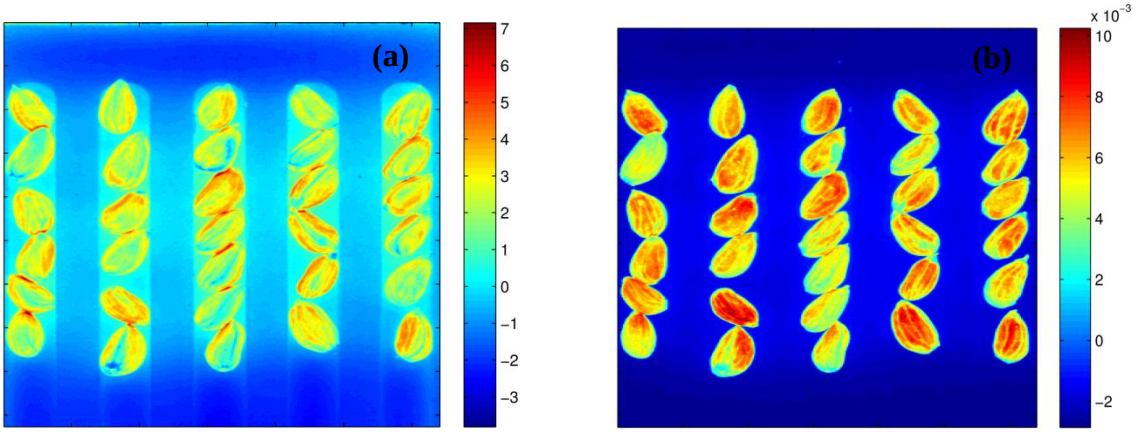



**Figura 3.4.** Imagem em pseudocores das sementes de algodão. (a) antes do pré-processamento e (b) após o pré-processamento dos espectros em cada pixel. Cores azuis e vermelhas referem-se a baixos e altos valores para a pseudo-absorbância média em cada pixel, respectivamente.

Na **Figura 3.5a** é apresentado o gráfico de escores de PC1 x PC2 após a imagem acima apresentada ser desdobrada em uma matriz de espectros, em que cada objeto corresponde a um determinado pixel. Como é possível perceber, dois agrupamentos destacadas pelas elipses

preta e verde são observados. Os pontos dentro da elipse verde estão associados aos pixels que formam a imagem da semente, enquanto que os pontos azuis dentro da elipse preta são relacionados com os espectros da base de teflon® usado como suporte para as amostras. Os escores com valores entre -0,7 $_X$ $10^{-3}$ e 1,5 $_X$ $10^{-3}$ em PC1 (destacados em vermelho) correspondem aos pixels das bordas das sementes, como apresentados na **Figura 3.5b**. A partir da análise realizada na **Figura 3.5**, um valor igual à zero em PC1 foi adotado como limiar para a remoção dos pixels de fundo.

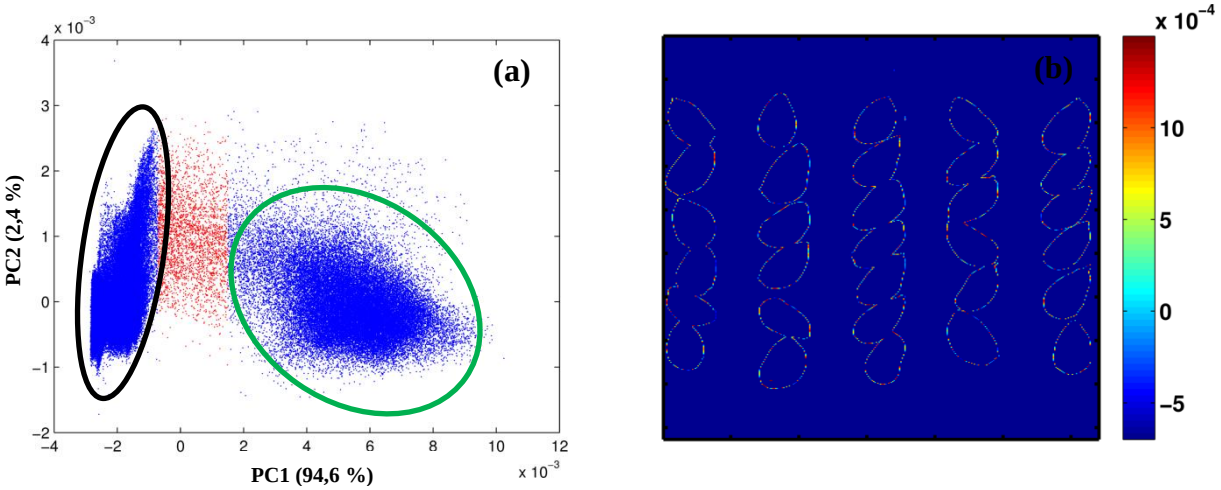



**Figura 3.5** (a) Escores em PC1 versus PC2. Os pontos em vermelho correspondem à faixa entre -0,7 X $10^{-3}$ e 1,5 X $10^{-3}$ em PC1. (b) Imagem após selecionar apenas os pontos da região em vermelho. Em (b) as cores azuis e vermelhas referem-se a baixos e altos valores de escores em PC1, respectivamente

Na **Figura 3.6a** os escores dos dois grupos separados pelo limiar em PC1 é mostrado. Os escores maiores que o limiar são mostrados na imagem em pseudocores representada na **Figura 3.6b**. Como é possível perceber, estes pontos correspondem à localização das sementes na imagem.

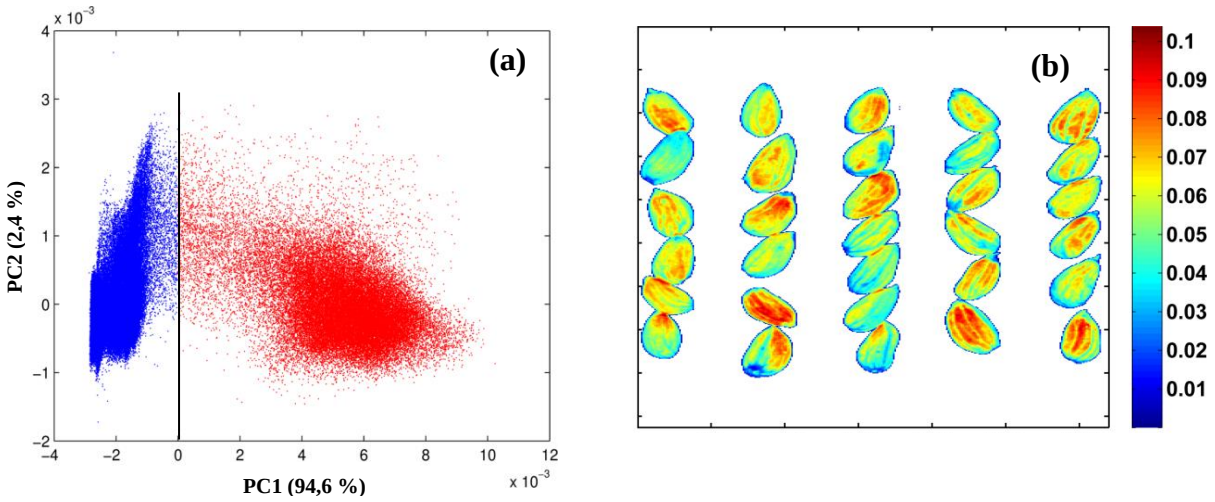



**Figura 3.6.** (a) Escores em PC1 versus PC2. (b) Imagem dos escores de PC1 após aplicar o limiar. Em (b) as cores azuis e vermelhas referem-se a baixos e altos valores de escores em PC1, respectivamente

Com a finalidade de separar cada semente da sua vizinha, o modulo de "seleção manual da região de interesse" apresentado na **Figura 3.7** foi usado na imagem em pseudocores dos escores de PC1. Os polígonos envoltos de cada semente foram desenhados manualmente e os pixels de fundo dentro de cada polígono foram eliminados usando o limiar obtido pela análise de componentes principais. Ao final, as posições dos pixels que corresponde a cada semente foram armazenadas. Essa informação é muito importante na recuperação dos espectros das imagens sem tratamento, o que possibilitou realizar outros pré-processamentos e tratar cada semente individualmente. Todo este processo foi repetido em outras 11 imagens adquiridas para todas as amostras restantes usadas neste estudo.

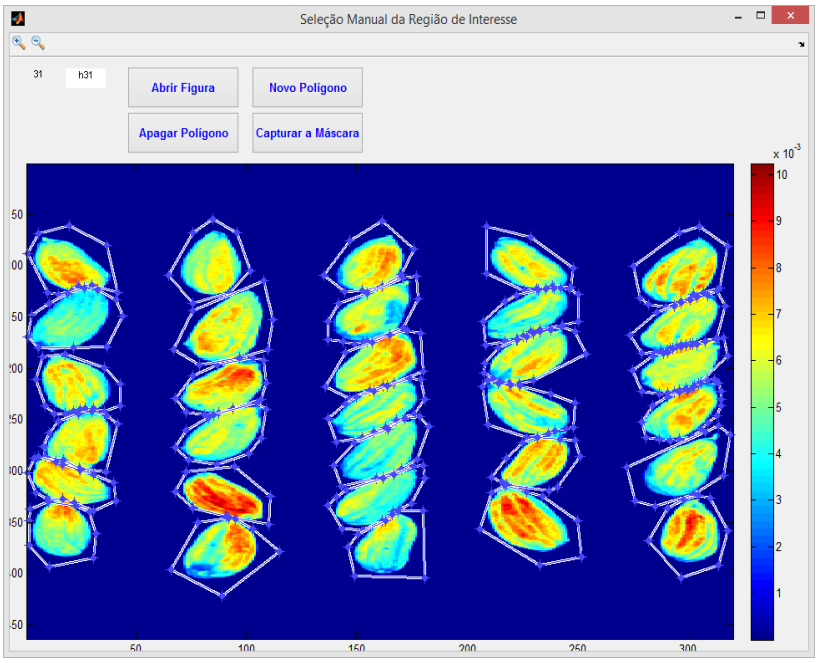



**Figura 3.7.** Modulo de seleção manual da região de interesse usado na separação das sementes vizinhas.

### 3.5.2 Pré-processamento dos espectros de cada semente

Na **Figura 3.8a** são apresentados os espectros médios da região de interesse para as 402 amostras de sementes que foram medidos usando a estação de imagens HSI-NIR. Com objetivo de remover o perfil de linha base e os efeitos de espalhamento, a segunda derivada com filtro Savitzky-Golay e um polinômio de segunda ordem ajustado a uma janela de 17 pontos foram usados. Os espectros pré-processados são mostrados na **Figura 3.8b**.

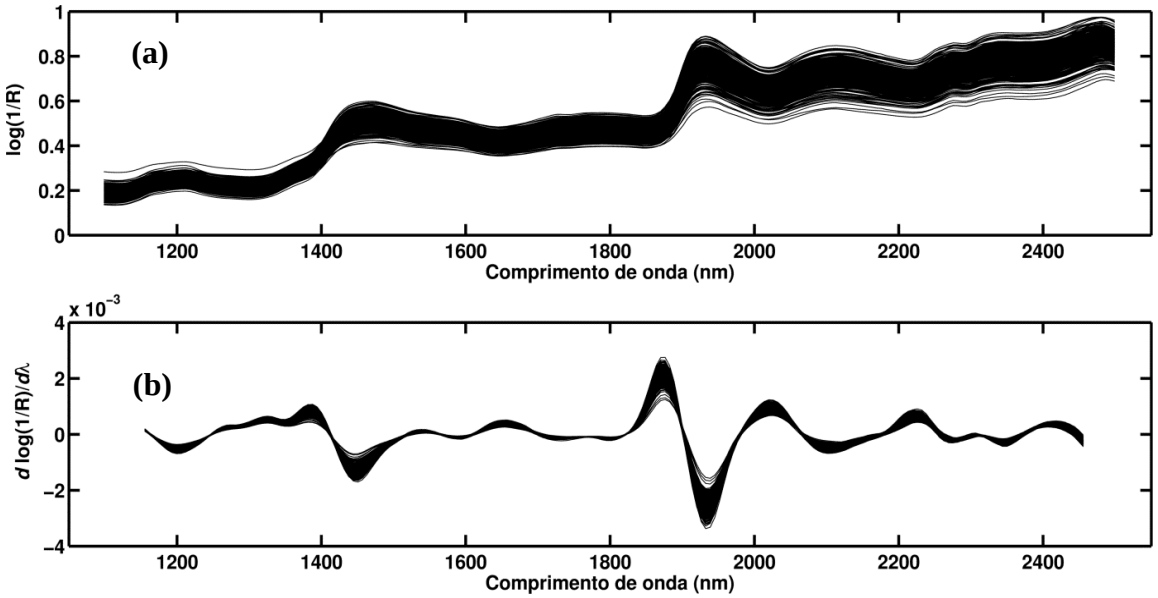



**Figura 3.8** Espectros médios das sementes registradas no HSI (a) antes e (b) após o pré-processamento.

Nas **Figura 3.9a e Figura 3.9b** os espectros de 392 amostras de sementes obtidos usando a tecnologia NIR convencional são apresentados. O pré-processamento também foi usado aqui com o mesmo propósito do caso anterior, a única diferença entre os dois tratamentos de dados deve-se à largura da janela de 19 pontos utilizada.

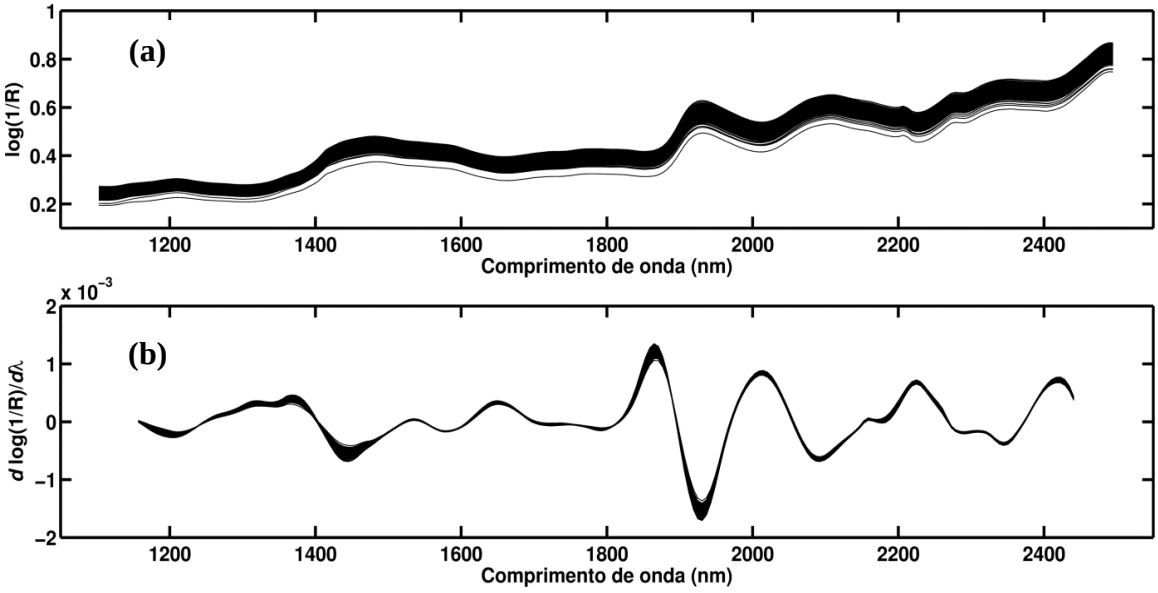



**Figura 3.9.** Espectros individuais registrados no NIR convencional (a) antes e (b) após o pré-processamento.

### 3.5.3 Análise exploratória dos dois conjuntos de dados

Uma análise de componentes principais foi realizada nos espectros dos quatro grupos de sementes registradas no equipamento HSI-NIR após serem pré-processados. Por contraste, a PCA é também aplicada ao NIR convencional. Os gráficos de escores obtidos pelas três primeiras PC's são apresentados nas **Figura 3.10a e Figura 3.10b**. Como se pode perceber, nenhuma tendência de separação entre as classes é observada para os dois conjuntos de dados em estudo.

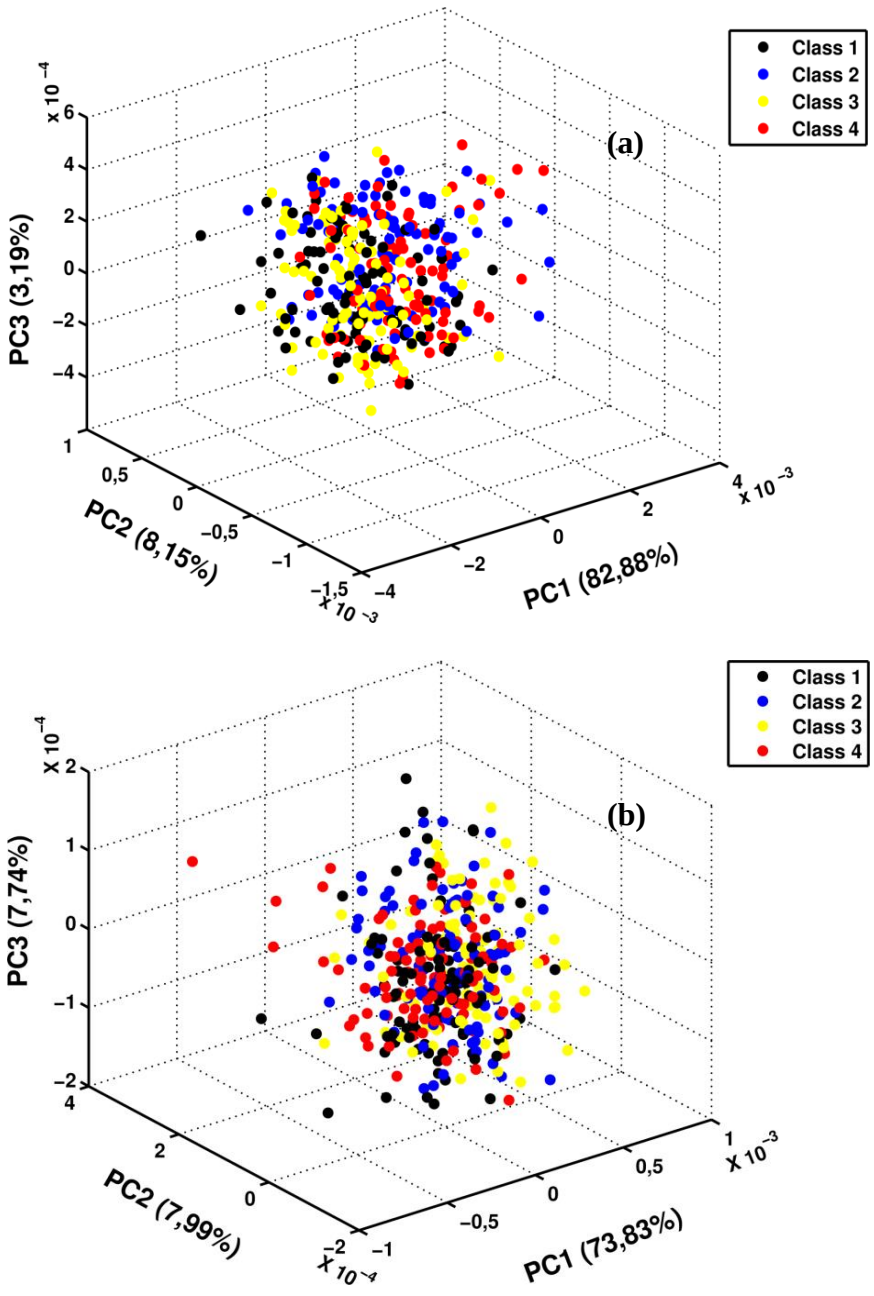



**Figura 3.10.** Gráfico de escores para os espectros registrados no (a) HSI-NIR e (b) NIR convencional.

### 3.5.4 Modelos SPA-LDA e PLS-DA para o HSI-NIR

Na **Figura 3.11a** são apresentados os valores da função custo do SPA-LDA, quando aplicado ao conjunto de validação, que foram usados para guiar a escolha das variáveis. Como pode ser visto, um total de 22 variáveis foram selecionadas. Na **Figura 3.11b** são mostrados os valores da taxa de erro de classificação que foram usados na escolha do número de variáveis latentes no PLS-DA. Como é possível perceber, a partir de 16 variáveis latentes a taxa de erro obtida no conjunto de validação não é alterado significativamente. A variância explicada em X representa 99,84% e em Y 72,68% quando estas variáveis latentes são usadas.

Os resultados obtidos no conjunto de validação produziram um total de 3 erros para o modelo SPA-LDA e 2 erros para o PLS-DA. Em termos de taxas de classificação correta os resultados foram 96,97% e 97,98%, respectivamente.

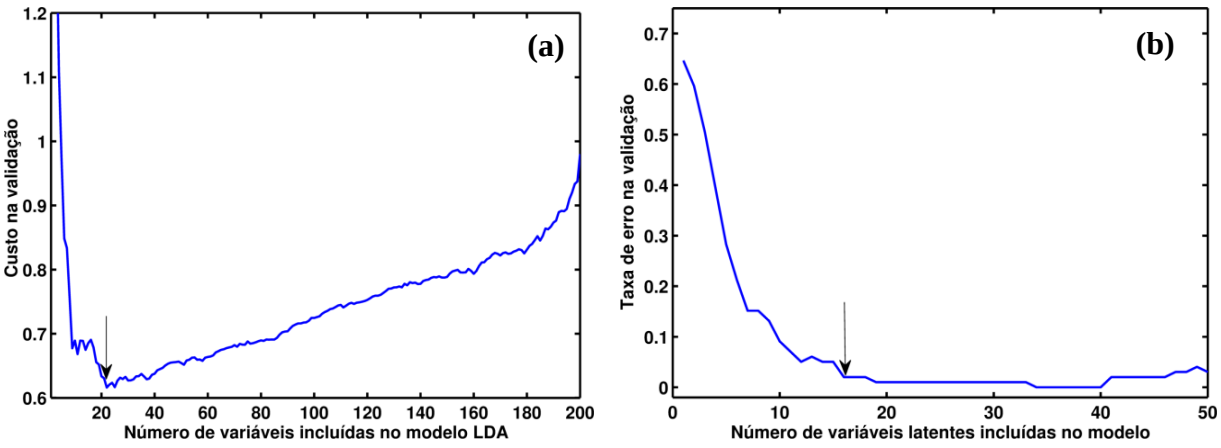



**Figura 3.11.** Dados HSI: (a) Gráficos do custo na validação *versus* número de variáveis incluídas no modelo LDA; (b) taxa de erro de classificação obtida no conjunto de validação *versus* número de variáveis latentes incluídas no modelo.

A fim de realizar uma avaliação final dos modelos SPA-LDA e PLS-DA, recorreu-se ao uso do conjunto de teste. As matrizes de confusão contendo o total de amostras classificadas em cada classe são apresentadas na **Tabela 3.4**.

**Tabela 3.4.** Matriz de confusão obtida para os dois modelos SPA-LDA e PLS-DA na classificação das amostras do conjunto de teste registradas no HSI-NIR.

| | | SPA-LDA | | | | PLS-DA | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Classe calculada | | | | Classe calculada | | | |
| Classe Experimental | *N* | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 1 | 24 | 24 | - | - | - | 23 | 1 | - | - |
| 2 | 26 | - | 26 | - | - | - | 26 | - | |
| 3 | 25 | 1 | - | 24 | - | - | 1 | 24 | - |
| 4 | 24 | - | - | - | 24 | - | - | - | 24 |

Como é possível perceber na **Tabela 3.4**, apenas uma amostra da classe 3 foi classificada na classe 1 pelo SPA-LDA. A passo que no PLS-DA, uma amostra da classe 1 e uma da classe 3 foram erroneamente classificadas na classe 2. Estes resultados proporcionaram uma taxa de classificação correta igual a 98,99% e 97,98% para os modelos SPA-LDA e PLS-DA, respectivamente.

Na **Tabela 3.5** são apresentados os valores de especificidade e sensibilidade para as quatro classes e para os dois modelos SPA-LDA e PLS-DA estudados. Como podem ser observados, os valores de sensibilidade não foram iguais apenas para a classe 1, mostrando o SPA-LDA um pouco superior. No parâmetro especificidade, o desempenho foi diferente para as classes 1 e 2, com o SPA-LDA um pouco superior para a classe 1 e o PLS-DA para a classe 2.

**Tabela 3.5.** Parâmetros de classificação obtidos para os modelos SPA-LDA e PLS-DA na classificação das amostras do conjunto de teste registradas no HSI-NIR.

| Modelos | Sensibilidade | | | | Especificidade | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| SPA-LDA | 1 | 1 | 0,960 | 1 | 0,987 | 1 | 1 | 1 |
| PLS-DA | 0,958 | 1 | 0,960 | 1 | 1 | 0,973 | 1 | 1 |

Na **Figura 3.12** são apresentados os valores dos escores de Fisher para o modelo LDA com as variáveis selecionadas pelo SPA.

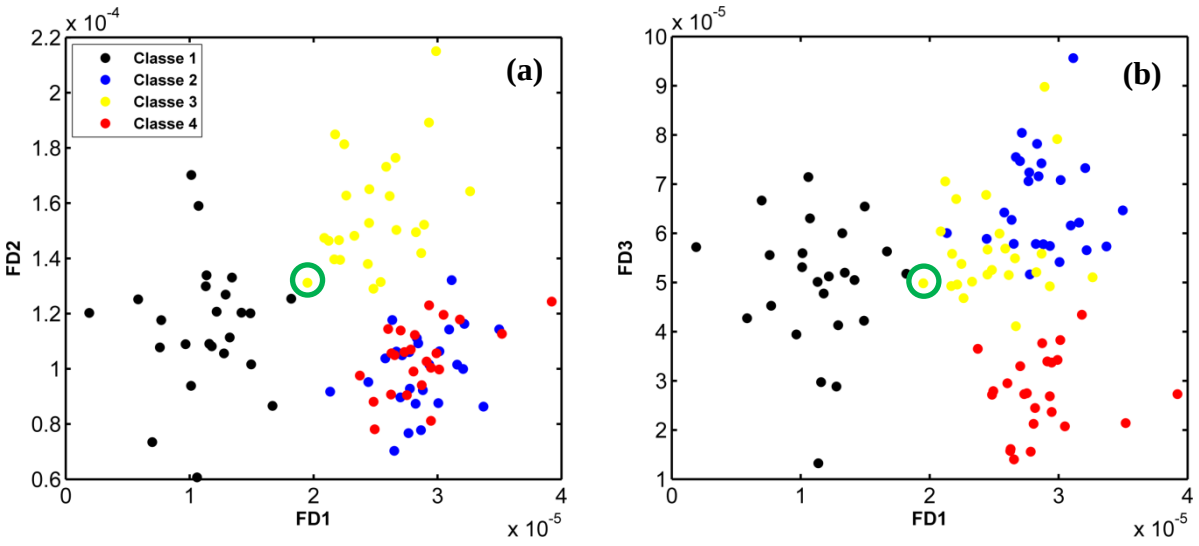



**Figura 3.12** Gráfico dos escores (a) da função discriminante 1 (FD1) *versus* função discriminante 2 (FD2) e (b) da função discriminante 1 *versus* função discriminante 3 (FD3) para as amostras do conjunto de teste medidas no HSI-NIR.

Na **Figura 3.12** é possível perceber que a LDA proporcionou a separação entre as classes estudadas. De fato, percebe-se na **Figura 3.12a** que a FD1 junto com a FD2 foi capaz de separar as amostras em três grupos formados pelas classes 1 e 3 e um outro pelas classes 2 e 4. Ao passo que na **Figura 3.12b**, observou-se que função discriminante FD3 foi responsável por promover a separação entre as amostras das classes 2 e 4. A amostra marcada pelo circulo verde foi classificada erradamente na classe 1.

### 3.5.5 Modelos SPA-LDA e PLS-DA para o NIR convencional

Na **Figura 3.13a** são apresentados os valores da função custo para o conjunto de validação usado na escolha das variáveis no SPA-LDA. Como pode ser visto, um total de 26 variáveis foram selecionadas. Na **Figura 3.13b** as taxas de erros usadas na escolha do número de variáveis latentes no PLS-DA são apresentadas. Nesse caso, 26 variáveis latentes foram

selecionadas a partir do melhor compromisso entre a menor taxa de erro e a variância explicada (que alcançou 99,97% em X e 70,95% em Y).

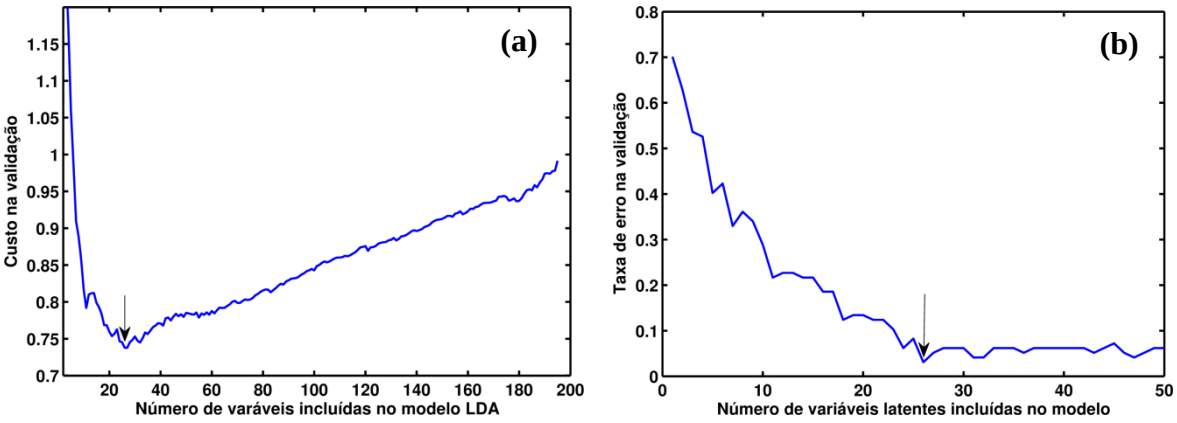



**Figura 3.13.** Dados NIR convencional: (a) Gráficos do custo na validação *versus* número de variáveis incluídas no modelo LDA; (b) taxa de erro de classificação obtida no conjunto de validação *versus* número de variáveis latentes incluídas no modelo.

O modelo SPA-LDA resultante com as variáveis selecionadas produziu um total de 10 erros na predição do conjunto de validação. Enquanto que o PLS-DA quando aplicado a este mesmo conjunto, proporcionou um total de 3 erros. Esses resultados em termos de taxas de classificação correta foram 89,69% e 96,91%, respectivamente.

A fim de realizar uma avaliação final dos modelos SPA-LDA e PLS-DA nos dados NIR convencional, recorreu-se ao uso do conjunto de teste. Os resultados são apresentados na **Tabela 3.6**.

**Tabela 3.6.** Matriz de confusão obtida para os modelos SPA-LDA e PLS-DA na classificação das amostras do conjunto de teste registradas no NIR convencional.

| | | SPA-LDA | | | | PLS-DA | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Classe calculada | | | | Classe calculada | | | |
| Classe Experimental | *N* | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 1 | 25 | 23 | 2 | - | - | 24 | 1 | - | - |
| 2 | 24 | 1 | 17 | 3 | 3 | 2 | 18 | - | 4 |
| 3 | 24 | - | 3 | 21 | - | - | 1 | 23 | - |
| 4 | 24 | 1 | 4 | - | 19 | - | 2 | 1 | 21 |

Como é possível observar na **Tabela 3.6**, o modelo SPA-LDA quando aplicado ao conjunto de teste produziu um total de 17 erros, enquanto o que o PLS-DA errou 11 vezes. Em termos de taxa de classificação correta o SPA-LDA e o PLS-DA alcançaram 82,47 % e 88,66 %, respectivamente.

Na **Tabela 3.7** são apresentados os valores de especificidade e sensibilidade para as quatro classes e para os dois modelos SPA-LDA e PLS-DA estudados. Como podem ser observados, com exceção a classe 4, o PLS-DA obteve maiores valores de sensibilidade e especificidade.

**Tabela 3.7.** Parâmetros de classificação obtidos para os modelos SPA-LDA e PLS-DA na classificação das amostras do conjunto de teste registradas no NIR convencional.

| Modelos | Sensibilidade | | | | Especificidade | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| SPA-LDA | 0,920 | 0,708 | 0,875 | 0,792 | 0,972 | 0,877 | 0,959 | 0,959 |
| PLS-DA | 0,960 | 0,750 | 0,958 | 0,875 | 0,972 | 0,945 | 0,986 | 0,945 |

Na **Figura 3.14** são apresentados os valores dos escores de Fisher para o conjunto amostras de teste registradas no equipamento NIR-Convencional.

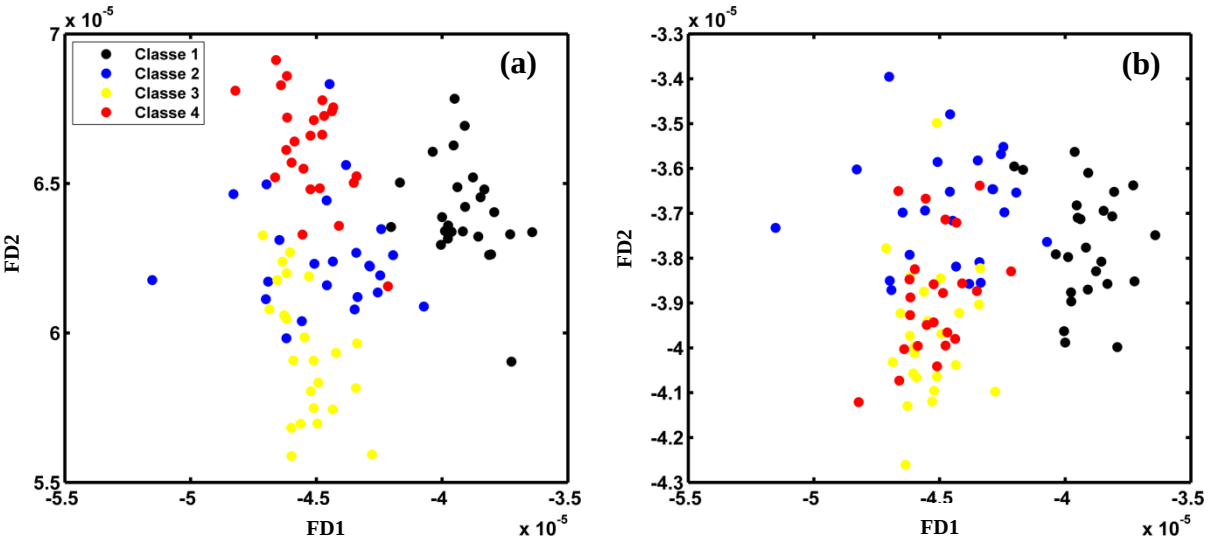



**Figura 3.14** Gráfico dos escores (a) da função discriminante 1 *versus* função discriminante 2 e (b) da função discriminante 1 *versus* função discriminante 3 para as amostras do conjunto de teste medidas no NIR convencional.

Ao observar a **Figura 3.14a** é possível perceber uma tendência de separação semelhante à observada no HSI-NIR. Como no caso anterior, a FD1 foi responsável por separar as amostras da classe 1 das demais classes. Por outro lado, em FD2 uma separação tão efetiva não foi observada e como consequência uma grande sobreposição entre as amostras das classes 2, 3 e 4. Na **Figura 3.14b** a função discriminante FD3 não foi capaz de promover a separação entre as amostras sobrepostas.

## 3.6 Conclusões

Como visto ao longo deste capítulo, neste estudo foi apresentada uma nova estratégia de classificação de sementes de algodão de forma não destrutiva usando metodologias quimiométricas aliadas a técnicas de imagens e/ou espectroscópicas.

Os resultados obtidos para o conjunto de teste demonstraram um excelente desempenho de classificação dos modelos SPA-LDA e PLS-DA. Quando aplicados aos dados HSI-NIR, os modelos conseguiram alcançar taxas de classificação correta iguais a 98,99% e 97,98% para o SPA-LDA e PLS-DA, respectivamente.

Nos espectros registrados no equipamento NIR-convencional, as taxas de classificação correta para os modelos SPA-LDA e PLS-DA foram iguais a 82,47 % e 88,66 %, respectivamente.

Por fim, percebe-se que os espectros registrados na estação de imagens proporcionaram resultados mais satisfatórios quando comparados ao NIR convencional. Isso foi alcançado porque o sistema de imagens HSI permite realizar uma varredura de uma área maior da semente e consequentemente obter um espectro médio mais representativo. Além disso, a aquisição dos espectros usando a tecnologia HSI é mais rápida em relação ao NIR convencional.

# CAPÍTULO 4: PROPOSTAS FUTURAS

Ao longo do desenvolvimento desse trabalho observou-se que alguns possíveis estudos podem ser vislumbrados como propostas futuras. Diante disso destacaremos as seguintes:

1. No desenvolvimento do novo método de transferência de calibração:
    - O procedimento de regressão robusta poderá ser estendido para o uso no PDS, substituindo a correção univariada. Nesse caso, o método de regressão PLS também poderá ser substituído por uma "regressão robusta PLS", e assim não será necessário usar métodos de seleção de variáveis;
    - Acoplar o novo método com a regressão via LASSO (*least absolute shrinkage and selection operator*) com objetivo de integrar o processo de seleção de variáveis ao procedimento de regressão. Para esse fim, a função custo definida na regressão robusta será minimizada e sujeita a uma restrição na 1-norma do vetor de coeficientes de regressão **b**. Neste caso, as propriedades estatísticas da perturbação $\Delta\mathbf{X}$ deverá ainda ser determinada a partir dos resíduos do procedimento de correção univariada, que poderá ser realizada independentemente da regressão robusta

2. Na nova estratégia de classificação de sementes:
    - Estudar a composição das sementes de algodão usando as técnicas de tratamentos de imagens e suas distribuição em mapas de pixels;
    - Desenvolver um novo método de transferência de classificação usando métodos de regularização.

# CAPÍTULO 5: REFERÊNCIAS BIBLIOGRÁFICAS

ABRAPA. Associação Brasileira dos produtores de Algodão. **Relatório de Gestão, Biênio 2011/2012**, 2016. Disponível em: < http://www.abrapa.com.br/biblioteca/Paginas/biblioteca-institucional.aspx >. Acesso em: 15 de Fevereiro de 2016.

AMIGO, J. M.; MARTÍ, I.; GOWEN, A. Chapter 9 - Hyperspectral Imaging and Chemometrics: A Perfect Combination for the Analysis of Food Structure, Composition and Quality. In: FEDERICO, M. (Ed.). **Data Handling in Science and Technology**: Elsevier, v.Volume 28, 2013. p.343-370. ISBN 0922-3487.

ANDREW, A.; FEARN, T. Transfer by orthogonal projection: making near-infrared calibrations robust to between-instrument variation. **Chemometrics and Intelligent Laboratory Systems,** v. 72, n. 1, p. 51-56, 2004.

ARAÚJO, F. H. **Previsão de Propriedades de Gasolinas do Nordeste Empregando Espectroscopia NIR/MIR e Transferência de Calibração**. 2006. 91 f. Tese (Doutorado) - CCEN Química, Universidade Federal de Pernambuco, Recife.

BAKEEV, K.; KURTYKA, B. Sources of measurement variability and their effect on the transfer of near infrared spectral libraries. **Journal of Near Infrared Spectroscopy,** v. 13, n. 6, p. 339-348, 2005.

BALLABIO, D.; CONSONNI, V. Classification tools in chemistry. Part 1: linear models. PLS-DA. **Analytical Methods,** v. 5, n. 16, p. 3790-3798, 2013.

BARKER, M.; RAYENS, W. Partial least squares for discrimination. **Journal of Chemometrics,** v. 17, n. 3, p. 166-173, 2003.

BAYLEY, C. et al. Engineering 2,4-D resistance into cotton. **Theoretical and Applied Genetics,** v. 83, n. 5, p. 645-649, 1992.

BEEBE, K. R.; PELL, R. J.; SEASHOLTZ, M. B. **Chemometrics: a practical guide**. Wiley, 1998. ISBN 9780471124511.

BELTRÃO, N. E. M.; AZEVEDO, D. M. P. **O agronegócio do algodão no Brasil**. Brasília: Embrapa Informação Tecnológica, 2008.

BINFENG, Y.; HAIBO, J. Near-infrared calibration transfer via support vector machine and transfer learning. **Anal. Methods,** v. 7, n. 6, p. 2714-2725, 2015.

BLANCO, M. et al. Near-infrared spectroscopy in the pharmaceutical industry . Critical Review. **Analyst,** v. 123, n. 8, p. 135R-150R, 1998.

BLANK, T. B. et al. Transfer of Near-Infrared Multivariate Calibrations without Standards. **Analytical Chemistry,** v. 68, n. 17, p. 2987-2995, 1996.

BOHNERT, H. J.; JENSEN, R. G. Strategies for engineering water-stress tolerance in plants. **Trends in Biotechnology,** v. 14, n. 3, p. 89-97, 1996.

BORGES, C. et al. Geographical classification of weathered crude oil samples with unsupervised self-organizing maps and a consensus criterion. **Chemometrics and Intelligent Laboratory Systems,** v. 101, n. 1, p. 43-55, 2010.

BOUVERESSE, E.; MASSART, D. L. Standardisation of near-infrared spectrometric instruments: A review. **Vibrational Spectroscopy,** v. 11, n. 1, p. 3-15, 1996.

BOUVERESSE, E.; MASSART, D. L.; DARDENNE, P. Modified Algorithm for Standardization of Near-Infrared Spectrometric Instruments. **Analytical Chemistry,** v. 67, n. 8, p. 1381-1389, 1995.

BRERETON, R. G. One Class Classifiers. In: (Ed.). **Chemometrics for Pattern Recognition**: John Wiley & Sons, Ltd, 2009a. p.233-287. ISBN 9780470746462.

_____. Two Class Classifiers. In: (Ed.). **Chemometrics for Pattern Recognition**: John Wiley & Sons, Ltd, 2009b. p.177-231. ISBN 9780470746462.

BRITO, G. G. et al. Leaf-level carbon isotope discrimination and its relationship with yield components as a tool for cotton phenotyping in unfavorable conditions. **Acta Scientiarum. Agronomy,** v. 36, p. 335-345, 2014.

BRO, R. Multivariate calibration: What is in chemometrics for the analytical chemist? **Analytica Chimica Acta,** v. 500, n. 1–2, p. 185-194, 2003.

BROWN, S. D. 3.08 - Transfer of Multivariate Calibration Models. In: WALCZAK, S. D. B. T. (Ed.). **Comprehensive Chemometrics**. Oxford: Elsevier, 2009. p.345-378. ISBN 978-0-444-52701-1.

CAMO. **The Unscrambler v9.7**. Nedre Vollgate 8, N-0158 OSLO, Norway 2007.

CAPITÁN-VALLVEY, L. F.; PALMA, A. J. Recent developments in handheld and portable optosensing—A review. **Analytica Chimica Acta,** v. 696, n. 1–2, p. 27-46, 2011.

CARVALHO, S. I. C. D.; BIANCHETTI, L. D. B.; REIFSCHNEIDER, F. J. B. Registro e proteção de cultivares pelo setor público: a experiência do programa de melhoramento de Capsicum da Embrapa Hortaliças. **Horticultura Brasileira,** v. 27, p. 135-138, 2009.

CHEN, J. et al. Population, water, food, energy and dams. **Renewable and Sustainable Energy Reviews,** v. 56, p. 18-28, 2016.

CHEN, Z. P. et al. Systematic prediction error correction: a novel strategy for maintaining the predictive abilities of multivariate calibration models. **Analyst,** v. 136, n. 1, p. 98-106, 2011.

CHENG, J.-H.; SUN, D.-W. Hyperspectral imaging as an effective tool for quality analysis and control of fish and other seafoods: Current research and potential applications. **Trends in Food Science & Technology,** v. 37, n. 2, p. 78-91, 2014.

COOPER, J. B.; LARKIN, C. M.; ABDELKADER, M. F. Calibration transfer of near-IR partial least squares property models of fuels using virtual standards. **Journal of Chemometrics,** v. 25, n. 9, p. 496-505, 2011.

DA SILVA FERNANDES, R. et al. Non-destructive detection of adulterated tablets of glibenclamide using NIR and solid-phase fluorescence spectroscopy and chemometric methods. **Journal of Pharmaceutical and Biomedical Analysis,** v. 66, p. 85-90, 2012.

DE ALMEIDA, M. R. et al. Discrimination between authentic and counterfeit banknotes using Raman spectroscopy and PLS-DA with uncertainty estimation. **Microchemical Journal,** v. 109, p. 170-177, 2013.

DE LIMA, K. M. G. A portable photometer based on LED for the determination of aromatic hydrocarbons in water. **Microchemical Journal,** v. 103, p. 62-67, 2012.

DE MAESSCHALCK, R.; JOUAN-RIMBAUD, D.; MASSART, D. L. The Mahalanobis distance. **Chemometrics and Intelligent Laboratory Systems,** v. 50, n. 1, p. 1-18, 2000.

DONG, W. et al. Discriminating plants using the DNA barcode rbcLb: an appraisal based on a large data set. **Molecular Ecology Resources,** v. 14, n. 2, p. 336-343, 2014.

DRAPER, N. R.; SMITH, H. **Applied regression analysis**. Wiley, 1998. ISBN 9780471170822.

DU, W. et al. Maintaining the predictive abilities of multivariate calibration models by spectral space transformation. **Anal Chim Acta,** v. 690, n. 1, p. 64-70, 2011.

ESTEVE AGELET, L.; HURBURGH, C. R., JR. Limitations and current applications of Near Infrared Spectroscopy for single seed analysis. **Talanta,** v. 121, p. 288-99, 2014.

FAN, W. et al. Calibration model transfer for near-infrared spectra based on canonical correlation analysis. **Analytica Chimica Acta,** v. 623, n. 1, p. 22-29, 2008.

FERNÁNDEZ PIERNA, J. A. et al. Direct Orthogonalization: some case studies. **Chemometrics and Intelligent Laboratory Systems,** v. 55, n. 1–2, p. 101-108, 2001.

FEUDALE, R. N. et al. Transfer of multivariate calibration models: a review. **Chemometrics and Intelligent Laboratory Systems,** v. 64, n. 2, p. 181-192, 2002.

FLATEN, G. R.; WALMSLEY, A. D. Using design of experiments to select optimum calibration model parameters. **Analyst,** v. 128, n. 7, p. 935-943, 2003.

FOCA, G. et al. Classification of pig fat samples from different subcutaneous layers by means of fast and non-destructive analytical techniques. **Food Research International,** v. 52, n. 1, p. 185-197, 2013.

FORINA, M.; LANTERI, S.; CASALE, M. Multivariate calibration. **Journal of Chromatography A,** v. 1158, n. 1–2, p. 61-93, 2007.

FORRESTER, J. B.; KALIVAS, J. H. Ridge regression optimization using a harmonious approach. **Journal of Chemometrics,** v. 18, n. 7-8, p. 372-384, 2004.

GAIÃO, E. D. N. et al. An inexpensive, portable and microcontrolled near infrared LED-photometer for screening analysis of gasoline. **Talanta,** v. 75, n. 3, p. 792-796, 2008.

GALVÃO, R. K. H.; ARAÚJO, M. C. U. 3.05 - Variable Selection. In: WALCZAK, S. D. B. T. (Ed.). **Comprehensive Chemometrics**. Oxford: Elsevier, 2009. p.233-283. ISBN 978-0-444-52701-1.

GALVÃO, R. K. H. et al. Calibration transfer employing univariate correction and robust regression. **Analytica Chimica Acta,** v. 864, p. 1-8, 2015.

GAO, J. et al. Application of hyperspectral imaging technology to discriminate different geographical origins of Jatropha curcas L. seeds. **Computers and Electronics in Agriculture,** v. 99, p. 186-193, 2013.

GELADI, P. L. M.; GRAHN, H. F.; BURGER, J. E. Multivariate Images, Hyperspectral Imaging: Background and Equipment. In: (Ed.). **Techniques and Applications of Hyperspectral Image Analysis**: John Wiley & Sons, Ltd, 2007. p.1-15. ISBN 9780470010884.

GIOVENZANA, V. et al. Testing of a simplified LED based vis/NIR system for rapid ripeness evaluation of white grape (Vitis vinifera L.) for Franciacorta wine. **Talanta,** v. 144, p. 584-591, 2015.

GOLUB, G. H.; HEATH, M.; WAHBA, G. Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. **Technometrics,** v. 21, n. 2, p. 215-223, 1979.

GOODARZI, M. et al. Multivariate calibration of NIR spectroscopic sensors for continuous glucose monitoring. **TrAC Trends in Analytical Chemistry,** v. 67, n. 0, p. 147-158, 2015.

HAALAND, D. M.; THOMAS, E. V. Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. **Analytical Chemistry,** v. 60, n. 11, p. 1193-1202, 1988.

HAN, J. et al. A peptide hormone gene, GhPSK promotes fibre elongation and contributes to longer and finer cotton fibre. **Plant Biotechnology Journal,** v. 12, n. 7, p. 861-871, 2014.

HANSEN, P. Analysis of Discrete Ill-Posed Problems by Means of the L-Curve. **SIAM Review,** v. 34, n. 4, p. 561-580, 1992.

HAUSER, P. C.; RUPASINGHE, T. W. T.; CATES, N. E. A multi-wavelength photometer based on light-emitting diodes. **Talanta,** v. 42, n. 4, p. 605-612, 1995.

HINDI, H. A.; BOYD, S. P. Robust solutions to l<sub>1</sub>, l<sub>2</sub>, and l<sub>∞</sub> uncertain linear approximation problems using convex optimization. American Control Conference, 1998. Proceedings of the 1998, 1998. 21-26 Jun 1998. p.3487-3491 vol.6.

HOERL, A. E.; KENNARD, R. W. Ridge Regression: Applications to Nonorthogonal Problems. **Technometrics,** v. 12, n. 1, p. 69-82, 1970a.

_____. Ridge Regression: Biased Estimation for Nonorthogonal Problems. **Technometrics,** v. 12, n. 1, p. 55-67, 1970b.

HONORATO, F. A. et al. Transferência de calibração em métodos multivariados. **Química Nova,** v. 30, p. 1301-1312, 2007.

HONORATO, F. A. et al. Robust modeling for multivariate calibration transfer by the successive projections algorithm. **Chemometrics and Intelligent Laboratory Systems,** v. 76, n. 1, p. 65-72, 2005.

HOPKE, P. K. The evolution of chemometrics. **Analytica Chimica Acta,** v. 500, n. 1–2, p. 365-377, 2003.

IVORRA, E. et al. Detection of expired vacuum-packed smoked salmon based on PLS-DA method using hyperspectral images. **Journal of Food Engineering,** v. 117, n. 3, p. 342-349, 2013.

JACKSON, J. E.; MUDHOLKAR, G. S. Control Procedures for Residuals Associated With Principal Component Analysis. **Technometrics,** v. 21, n. 3, p. 341-349, 1979.

KALIVAS, J. H. Interrelationships of multivariate regression methods using eigenvector basis sets. **Journal of Chemometrics,** v. 13, n. 2, p. 111-132, 1999.

KALIVAS, J. H. 3.01 - Calibration Methodologies. In: WALCZAK, S. D. B. T. (Ed.). **Comprehensive Chemometrics**. Oxford: Elsevier, 2009. p.1-32. ISBN 978-0-444-52701-1.

KALIVAS, J. H.; HEBERGER, K.; ANDRIES, E. Sum of ranking differences (SRD) to ensemble multivariate calibration model merits for tuning parameter selection and comparing calibration methods. **Anal Chim Acta,** v. 869, p. 21-33, 2015.

KARIMI, S.; HEMMATEENEJAD, B. Identification of discriminatory variables in proteomics data analysis by clustering of variables. **Analytica Chimica Acta,** v. 767, p. 35-43, 2013.

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

KONG, W. et al. Rice seed cultivar identification using near-infrared hyperspectral imaging and multivariate data analysis. **Sensors (Basel),** v. 13, n. 7, p. 8916-27, 2013.

KUMAR, N. et al. Chemometrics tools used in analytical chemistry: An overview. **Talanta,** v. 123, p. 186-199, 2014.

LAVINE, B. K. 3.19 - Validation of Classifiers A2 - Walczak, Steven D. BrownRomá TaulerBeata. In: (Ed.). **Comprehensive Chemometrics**. Oxford: Elsevier, 2009. p.587-599. ISBN 978-0-444-52701-1.

LEARDI, R. Genetic algorithms in chemometrics and chemistry: a review. **Journal of Chemometrics,** v. 15, n. 7, p. 559-569, 2001.

LEARDI, R.; SEASHOLTZ, M. B.; PELL, R. J. Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data. **Analytica Chimica Acta,** v. 461, n. 2, p. 189-200, 2002.

LEE, J. H.; CHOUNG, M.-G. Nondestructive determination of herbicide-resistant genetically modified soybean seeds using near-infrared reflectance spectroscopy. **Food Chemistry,** v. 126, n. 1, p. 368-373, 2011.

LEVI, A. et al. Field evaluation of cotton near-isogenic lines introgressed with QTLs for productivity and drought related traits. **Molecular Breeding,** v. 23, n. 2, p. 179-195, 2008.

MAHESH, S. et al. Feasibility of near-infrared hyperspectral imaging to differentiate Canadian wheat classes. **Biosystems Engineering,** v. 101, n. 1, p. 50-57, 2008.

MALINEN, J. et al. LED-based NIR spectrometer module for hand-held and process analyser applications. **Sensors and Actuators B: Chemical,** v. 51, n. 1–3, p. 220-226, 1998.

MARTINS, M. N.; GALVÃO, R. K. H.; PIMENTEL, M. F. Multivariate calibration transfer employing variable selection and subagging. **Journal of the Brazilian Chemical Society,** v. 21, p. 127-134, 2010.

MYLONAS, I. G. et al. Barley cultivar discrimination and hybrid purity control using RAPD markers. **Romanian Biotechnological Letters,** v. 19, n. 3, p. 9421-9428, 2014.

NÆS, T. **A User-friendly Guide to Multivariate Calibration and Classification**. NIR Publications, 2002. ISBN 9780952866626.

NÆS, T.; MEVIK, B.-H. Understanding the collinearity problem in regression and discriminant analysis. **Journal of Chemometrics,** v. 15, n. 4, p. 413-426, 2001.

NGO, S. H.; KEMÉNY, S.; DEÁK, A. Performance of the ridge regression method as applied to complex linear and nonlinear models. **Chemometrics and Intelligent Laboratory Systems,** v. 67, n. 1, p. 69-78, 2003.

NØRGAARD, L. Direct standardisation in multi wavelength fluorescence spectroscopy. **Chemometrics and Intelligent Laboratory Systems,** v. 29, n. 2, p. 283-293, 1995.

OLIVEIRA, E. J.; DIAS, N. L. P.; DANTAS, J. L. L. Selection of morpho-agronomic descriptors for characterization of papaya cultivars. **Euphytica,** v. 185, n. 2, p. 253-265, 2011.

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

PASQUINI, C. et al. Laser Induced Breakdown Spectroscopy. **Journal of the Brazilian Chemical Society,** v. 18, p. 463-512, 2007.

PENG, J. et al. Near-infrared calibration transfer based on spectral regression. **Spectrochim Acta A Mol Biomol Spectrosc,** v. 78, n. 4, p. 1315-20, 2011.

PEREIRA, C. F. et al. A comparative study of calibration transfer methods for determination of gasoline quality parameters in three different near infrared spectrometers. **Anal Chim Acta,** v. 611, n. 1, p. 41-7, 2008.

PONTES, M. J. C. et al. The successive projections algorithm for spectral variable selection in classification problems. **Chemometrics and Intelligent Laboratory Systems,** v. 78, n. 1–2, p. 11-18, 2005.

PONTES, M. J. C. et al. Internal and external validation in SPA-LDA: A comparative study involving diesel/biodiesel blends. **NIR news,** v. 23, n. 5, p. 6, 2012.

RASMUSSEN, M. A.; BRO, R. A tutorial on the Lasso approach to sparse modeling. **Chemometrics and Intelligent Laboratory Systems,** v. 119, n. 0, p. 21-31, 2012.

RIDGWAY, C.; CHAMBERS, J. Detection of insects inside wheat kernels by NIR imaging. **Journal of Near Infrared Spectroscopy,** v. 6, n. 1, p. 115-119, 1998.

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

RODRÍGUEZ-PULIDO, F. J. et al. Grape seed characterization by NIR hyperspectral imaging. **Postharvest Biology and Technology,** v. 76, p. 74-82, 2013.

SANTOS, M. B. H. et al. Non-destructive NIR spectrometric cultivar discrimination of castor seeds resulting from breeding programs. **Journal of the Brazilian Chemical Society,** v. 25, p. 969-974, 2014.

SERRANO-LOURIDO, D. et al. Classification and characterisation of Spanish red wines according to their appellation of origin based on chromatographic profiles and chemometric data analysis. **Food Chemistry,** v. 135, n. 3, p. 1425-1431, 2012.

SERRANTI, S. et al. Classification of oat and groat kernels using NIR hyperspectral imaging. **Talanta,** v. 103, p. 276-84, 2013.

SHENK, J. S.; WESTERHAUS, M. O. **Optical instrument calibration system**: U.S. Patent 4.866.644, 1989.

SHENK, J. S.; WESTERHAUS, M. O.; TEMPLETON, W. C. Calibration Transfer Between near Infrared Reflectance Spectrophotometers. **Crop Science,** v. 25, p. 159-161, 1985.

SHOWALTER, A. M. et al. A Primer for Using Transgenic Insecticidal Cotton in Developing Countries. **Journal of Insect Science,** v. 9, n. 22, p. 1-39, 2009.

SILVA, C. S. et al. Near infrared hyperspectral imaging for forensic analysis of document forgery. **Analyst,** v. 139, n. 20, p. 5176-5184, 2014.

SIMON, M. et al. DNA fingerprinting and new tools for fine-scale discrimination of Arabidopsis thaliana accessions. **The Plant Journal,** v. 69, n. 6, p. 1094-1101, 2012.

SINGH, C. B. et al. Detection of insect-damaged wheat kernels using near-infrared hyperspectral imaging. **Journal of Stored Products Research,** v. 45, n. 3, p. 151-158, 2009.

SINGH, C. B. et al. Detection of midge-damaged wheat kernels using short-wave near-infrared hyperspectral and digital colour imaging. **Biosystems Engineering,** v. 105, n. 3, p. 380-387, 2010.

SOARES, S. F. C. et al. A new validation criterion for guiding the selection of variables by the successive projections algorithm in classification problems. **Journal of the Brazilian Chemical Society,** v. 25, p. 176-181, 2014.

SOLOMON, C.; BRECKON, T. Representation. In: (Ed.). **Fundamentals of Digital Image Processing**: John Wiley & Sons, Ltd, 2010. p.1-19. ISBN 9780470689776.

SWIERENGA, H. et al. Comparison of Two Different Approaches toward Model Transferability in NIR Spectroscopy. **Applied Spectroscopy,** v. 52, n. 1, p. 7-16, 1998.

TORRES, J. B.; RUBERSON, J. R.; WHITEHOUSE, M. Transgenic Cotton for Sustainable Pest Management: A Review. In: LICHTFOUSE, E. (Ed.). **Organic Farming, Pest Control and Remediation of Soil Pollutants: Organic farming, pest control and remediation of soil pollutants**. Dordrecht: Springer Netherlands, 2010. p.15-53. ISBN 978-1-4020-9654-9.

VARMUZA, K.; FILZMOSER, P. **Introduction to Multivariate Statistical Analysis in Chemometrics**. CRC Press, 2009. ISBN 9781420059496.

VIDAL, M.; AMIGO, J. M. Pre-processing of hyperspectral images. Essential steps before image analysis. **Chemometrics and Intelligent Laboratory Systems,** v. 117, p. 138-148, 2012.

VILAR, W. T. S. et al. Classification of Individual Castor Seeds Using Digital Imaging and Multivariate Analysis. **Journal of the Brazilian Chemical Society**, 2014.

VITALE, R. et al. A rapid and non-invasive method for authenticating the origin of pistachio samples by NIR spectroscopy and chemometrics. **Chemometrics and Intelligent Laboratory Systems,** v. 121, p. 90-99, 2013.

WALCZAK, B.; BOUVERESSE, E.; MASSART, D. L. Standardization of near-infrared spectra in the wavelet domain. **Chemometrics and Intelligent Laboratory Systems,** v. 36, n. 1, p. 41-51, 1997.

WANG, J. et al. Development and implementation of a multiplexed single nucleotide polymorphism genotyping tool for differentiation of ryegrass species and cultivars. **Molecular Breeding,** v. 33, n. 2, p. 435-451, 2014.

WANG, Y.; VELTKAMP, D. J.; KOWALSKI, B. R. Multivariate instrument standardization. **Analytical Chemistry,** v. 63, n. 23, p. 2750-2756, 1991.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. **Chemometrics and Intelligent Laboratory Systems,** v. 58, n. 2, p. 109-130, 2001.

WONG, K. H. et al. Differentiation of Pueraria lobata and Pueraria thomsonii using partial least square discriminant analysis (PLS-DA). **Journal of Pharmaceutical and Biomedical Analysis,** v. 84, p. 5-13, 2013.

WU, W. et al. Comparison of regularized discriminant analysis linear discriminant analysis and quadratic discriminant analysis applied to NIR data. **Analytica Chimica Acta,** v. 329, n. 3, p. 257-265, 1996.

YANG, L.; SUN, Q. Recognition of the hardness of licorice seeds using a semi-supervised learning method and near-infrared spectral data. **Chemometrics and Intelligent Laboratory Systems,** v. 114, p. 109-115, 2012.

YOON, J.; LEE, B.; HAN, C. Calibration transfer of near-infrared spectra based on compression of wavelet coefficients. **Chemometrics and Intelligent Laboratory Systems,** v. 64, n. 1, p. 1-14, 2002.

ZHANG, X. et al. Application of hyperspectral imaging and chemometric calibrations for variety discrimination of maize seeds. **Sensors (Basel),** v. 12, n. 12, p. 17234-46, 2012.

ZHENG, K. et al. Calibration transfer of near-infrared spectra for extraction of informative components from spectra with canonical correlation analysis. **Journal of Chemometrics,** v. 28, n. 10, p. 773-784, 2014.

# ANEXO: PRODUÇÃO CIENTÍFICA

Ao longo do curso de doutorado em Química Analítica, foram publicados artigos científicos relacionados com o presente trabalho e que permitiram o desenvolvimento efetivo da presente tese. Os principais estão elencados a seguir. Outras produções científicas realizadas neste período podem ser visualizadas na Plataforma Lattes: http://lattes.cnpq.br/7653207334385666.

Galvão, R. K. H.; Soares, S. F. C.; Martins, M. N.; Pimentel, M. F.; Araújo, M. C. U.; Calibration transfer employing univariate correction and robust regression. Analytica Chim. Acta (Print), v. 864, p. 1-8, 2015.

Soares, S. F. C.; Galvão, R. K. H.; Pontes, M. J. C.; Araújo, M. C. U.; A New Validation Criterion for Guiding the Selection of Variables by the Successive Projections Algorithm in Classification Problems. Journal of the Brazilian Chemical Society (Impresso), v. 25, p. 176, 2014.

Soares, S. F. C.; Gomes, A. A.; Araujo, M. C. U.; Galvão Filho, A. R.; Galvão, R. K. H.; The successive projections algorithm. TrAC. Trends in Analytical Chemistry (Regular ed.), v. 42, p. 84-98, 2012.

Analytica Chimica Acta 864 (2015) 1–8

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

# Calibration transfer employing univariate correction and robust regression

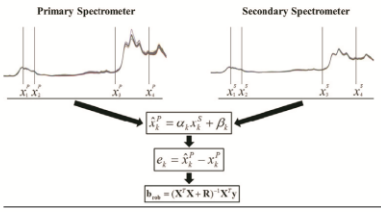
Roberto Kawakami Harrop Galvão [a], Sófacles Figueredo Carreiro Soares [b,c], Marcelo Nascimento Martins [a], Maria Fernanda Pimentel [d], Mário César Ugulino Araújo [b,*]

[a] Instituto Tecnológico de Aeronáutica, Divisão de Engenharia Eletrônica, São José dos Campos,São Paulo 12228-900, Brazil
[b] Universidade Federal da Paraíba, CCEN, Departamento de Química, Caixa Postal 5093, João Pessoa, Paraíba CEP 58051-970, Brazil
[c] Universidade Federal da Paraíba, CT, Departamento de Engenharia Química, João Pessoa, Paraíba CEP 58051-900, Brazil
[d] Universidade Federal de Pernambuco, Departamento de Engenharia Química, Recife, Pernambuco 50740-521, Brazil

## HIGHLIGHTS

- Calibration transfer involving individual wavelengths.
- Suitable for dedicated instruments.
- Examples involving near infrared spectrometric analysis of gasoline and corn.
- Better results compared to piecewise direct standardization (PDS).

## GRAPHICAL ABSTRACT




## ARTICLE INFO

*Article history:*
Received 17 June 2014
Received in revised form 23 September 2014
Accepted 2 October 2014
Available online 8 October 2014

*Keywords:*
Multivariate calibration
Calibration transfer
Univariate correction
Robust regression
Variable selection

## ABSTRACT

This paper proposes a new method for calibration transfer, which was specifically designed to work with isolated variables, rather than the full spectrum or spectral windows. For this purpose, a univariate procedure is initially employed to correct the spectral measurements of the secondary instrument, given a set of transfer samples. A robust regression technique is then used to obtain a model with low sensitivity with respect to the univariate correction residuals. The proposed method is employed in two case studies involving near infrared spectrometric determination of specific mass, research octane number and naphthenes in gasoline, and moisture and oil in corn. In both cases, better calibration transfer results were obtained in comparison with piecewise direct standardization (PDS). The proposed method should be of a particular value for use with application-targeted instruments that monitor only a small set of spectral variables.



## 1. Introduction

The development of multivariate calibration models involves several stages, typically including the collection of samples and recording of analytical signals, followed by the actual construction and validation of the model. All these stages are important to achieve good predictions when the resulting model is employed in the analysis of new samples. In particular, it would be desirable to eliminate or minimize the sources of data variability that are not related to the analytical property of interest. However, there are cases in which changes in the analytical conditions occur after the calibration has been carried out, with adverse effects on the prediction ability of the model [1,2]. Such changes may refer to the physical/chemical characteristics of the sample (such as viscosity, granularity, surface texture, and presence of interferent species),

* Corresponding author. Tel.: +55 83 3216 7438; fax: +55 83 3216 7437.
*E-mail addresses:* laqa@quimica.ufpb.br, mariougulino@gmail.com (M.C.U. Araújo).

http://dx.doi.org/10.1016/j.aca.2014.10.001
0003-2670/

http://dx.doi.org/10.5935/0103-5053.20130262
*J. Braz. Chem. Soc.*, Vol. 25, No. 1, 176-181, 2014.
Printed in Brazil - ©2014 Sociedade Brasileira de Química
0103 - 5053 $6.00+0.00

**Short Report**

# A New Validation Criterion for Guiding the Selection of Variables by the Successive Projections Algorithm in Classification Problems

*Sófacles F. C. Soares,[a] Roberto K. H. Galvão,[b] Márcio J. C. Pontes[a] and Mário C. U. Araújo*[,a]*

[a]*Departamento de Química, Centro de Ciências Exatas e da Natureza, Universidade Federal da Paraíba, Caixa Postal 5093, 58051-970 João Pessoa-PB, Brazil*

[b]*Divisão de Engenharia Eletrônica, Instituto Tecnológico de Aeronáutica, 12228-900 São José dos Campos-SP, Brazil*

Uma simplificação no SPA-LDA é proposta para contornar a necessidade de conjuntos de treinamento e validação separados. O número de graus de liberdade é empregado na função de custo para evitar sobreajuste do modelo. Três exemplos são apresentados: classificação de cafés, diesel e óleos vegetais empregando espectrometria UV-Vis, NIR e voltametria, respectivamente.

A simplification in SPA-LDA is proposed to circumvent the need for separate training and validation sets. The number of degrees of freedom is employed in the cost function to avoid model overfitting. Three examples are presented: classification of coffee, diesel and vegetable oils by using UV-Vis spectrometry, NIR spectrometry and voltammetry, respectively.

**Keywords:** variable selection, successive projections algorithm, classification, model validation

## Introduction

The successive projections algorithm (SPA) is a variable selection method originally proposed for the construction of multivariate calibration models[1] and subsequently extended to address classification problems.[2] Applications of SPA have involved different instrumental techniques and samples as summarized in a recent review paper.[3]

The SPA formulation for classification problems involves two phases. In the first phase, a sequence of projection operations involving the columns of the instrumental response matrix is employed to form subsets of variables with small collinearity. In the second phase, the best subset is selected on the basis of a cost function associated to the risk of incorrect classification by linear discriminant analysis (LDA). In Pontes *et al.*,[2] and all subsequent papers,[4-11] this cost function was evaluated by using an external set of validation samples, which were not employed in the construction of the LDA model. This procedure was adopted to avoid model overfitting, which might result if the training set itself was used in the evaluation of the cost function.

*e-mail: laqa@quimica.ufpb.br

Within this scope, two inconveniences related to the use of a separate validation set could be pointed out. Firstly, the analyst is faced with the problem of splitting the available samples into representative training and validation sets, which may not be a straightforward task. Secondly, if the number of samples is too small, it may not be possible to split them into two representative sets. Cross-validation could be an alternative, but the computational effort involved can be substantial, due to the need of constructing an LDA model for each sample (or group of samples) that is removed from the training set in the course of the cross-validation procedure. Another possibility would be the use of the training set itself for validation purposes. However, such an internal validation approach may lead to overfitting as discussed elsewhere.[12]

In this context, the present paper proposes a new criterion for internal validation in SPA-LDA in which the number of degrees of freedom is employed in the cost function calculation. As a result, model overfitting is avoided without the need to divide the available data into separate training and validation sets. The utility of the proposed criterion is investigated in a comparative study involving external validation and cross-validation. For this purpose, three analytical problems are considered, namely UV-Vis spectrometric classification of coffee,[7]

Trends Trends in Analytical Chemistry, Vol. 42, 2013

# The successive projections algorithm

Sófacles Figueredo Carreiro Soares, Adriano A. Gomes, Arlindo Rodrigues Galvão Filho, Mario Cesar Ugulino Araujo, Roberto Kawakami Harrop Galvão

**The successive projections algorithm (SPA) is a variable-selection technique that has attracted increasing interest in the analytical-chemistry community in the past 10 years. The present review presents the basic features of SPA for Multiple Linear Regression (MLR) and Linear Discriminant Analysis (LDA) and reports some variants that have been proposed for sample selection, calibration transfer and Quantitative Structure-Activity Relationship (QSAR) and Quantitative Structure-Property Relationship (QSPR) studies. We also discuss computational and pre-processing issues. By way of illustration we present two case studies involving near-infrared determination of protein in wheat and voltammetric classification of vegetable oils. The code employed in this article is freely available from us upon request.**


*Keywords:* Calibration transfer; Classification; Linear Discriminant Analysis (LDA); Multiple Linear Regression (MLR); Multivariate calibration; Quantitative Structure-Activity Relationship (QSAR); Quantitative Structure-Property Relationship (QSPR); Sample selection; Successive projections algorithm (SPA); Variable selection

**Sófacles Figueredo Carreiro Soares, Adriano A. Gomes, Mario Cesar Ugulino Araujo***
Universidade Federal da Paraíba, CCEN, Departamento de Química, Caixa Postal 5093, CEP 58051-970, João Pessoa, PB, Brazil

**Arlindo Rodrigues Galvão Filho, Roberto Kawakami Harrop Galvão**
Instituto Tecnológico de Aeronáutica, Divisão de Engenharia Eletrônica, 12228-900, São José dos Campos, SP, Brazil

*Corresponding author.
Tel.: +55 83 3216 7438;
Fax: +55 83 3216 7437;
E-mail: laqa@quimica.ufpb.br

## 1. Introduction

Modern analytical methods typically employ instrumental techniques to analyze solid, liquid or gaseous samples with fewer chemical treatments and reduced waste generation. The instruments employed for this purpose usually involve many analytical channels, so they generate data sets with a considerable number of variables. Examples include laser-induced breakdown spectroscopy (LIBS) [1] and near-infrared spectroscopy (NIR) [2], which deliver measurements over a large number of wavelengths for each sample. However, in many cases, the instrumental response exhibits strong correlation over different analytical channels, which leads to redundancy in the acquired data. Moreover, some channels may not provide relevant information for the problem under consideration and their use may even compromise the precision and the accuracy of the result [3]. For these reasons, the analytical method may benefit from the use of a reduced subset of channels, rather than the entire set of instrumental measurements obtained for each sample. In addition, the identification of an appropriate subset of channels facilitates the interpretation of the results and may be useful to guide the design of less costly instruments that are dedicated to the analytical application at hand [4]. In the chemometrics literature, this procedure is termed variable selection.

Variable selection generally benefits from *a priori* knowledge about the physical and chemical properties of the system under analysis and the technical features of the measurement instrument (e.g., in spectroscopy, the analyst should exclude wavelength regions in which the measured signal saturates, the signal-to-noise ratio of the detector is too small, or the analyte response is strongly overlapped by interferents). However in some cases, the decision is not so clear cut, which motivates the use of chemometrics techniques.

A pragmatic approach to variable selection involves using a computational method to search for the combination of variables that optimizes some performance index related to the analytical result [5]. This index is usually termed cost function when the optimization involves the search for a minimum value. Examples include minimization of the root-mean-square error of prediction (RMSEP) or cross-validation (RMSECV) in


0165-9936/$ - see front matter © 2012 Elsevier Ltd. All rights reserved. doi:http://dx.doi.org/10.1016/j.trac.2012.09.006