



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

DISSERTAÇÃO DE MESTRADO

**Algoritmo das Projeções Sucessivas
aplicado à seleção de variáveis em
regressão PLS**

Adriano de Araújo Gomes

João Pessoa – PB - Brasil
Março/2012



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

DISSERTAÇÃO DE MESTRADO

Algoritmo das Projeções Sucessivas aplicado à seleção de variáveis em regressão PLS

Adriano de Araújo Gomes*

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

1º Orientador: **Prof. Dr. Edvan Cirino da Silva**

2º Orientador: **Prof. Dr. Mário César Ugulino de Araújo**

* Bolsista:



João Pessoa – PB – Brasil

Março/2012

G633a Gomes, Adriano de Araújo.

Algoritmo das projeções sucessivas aplicado à seleção de variáveis em regressão PLS / Adriano de Araújo Gomes.--João Pessoa, 2012.

120f. : il.

Orientadores: Edvan Cirino da Silva, Mário César Ugulino de Araújo

Dissertação (Mestrado) – UFPB/CCEN

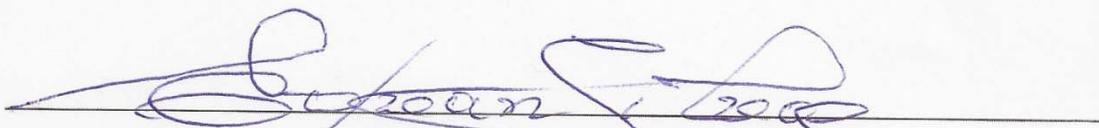
1. Química Analítica. 2. Algoritmo das projeções sucessivas. 3. Mínimos quadrados parciais. 4. Seleção de variáveis. 5. Intervalos.

UFPB/BC

CDU: 543(043)

Algoritmo das Projeções Sucessivas aplicado à seleção de variáveis em regressão PLS

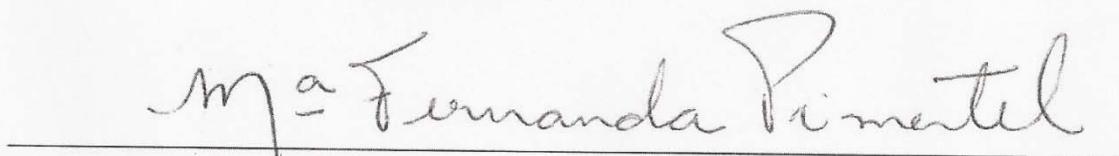
Dissertação de Mestrado de Adriano de Araújo Gomes aprovada pela banca examinadora em 08 de março de 2012:



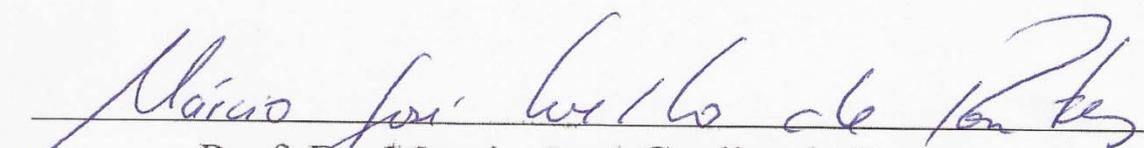
Prof. Dr. Edvan Cirino da Silva
Orientador/Presidente



Prof. Dr. Mário Cesar Ugulino de Araújo
2º. Orientador



Profa. Drª. Maria Fernanda Pimentel Avelar
Examinador



Prof. Dr. Marcio José Coelho de Pontes
Examinador

A todos que me ajudaram chegar ate aqui! Pois nada se faz sozinho.

Com gratidão,

Dedico.

AGRADECIMENTOS

Ao grande DEUS, pelo seu infinito amor e bondade.

Aos meus pais João Batista e Cícera, por todo amor dedicado ao longo da minha vida.

Aos meus irmãos Tânia, Vanderli, Francisca e Fernando, por toda ajuda durante minha graduação.

A toda família, por ter acreditado em mim e ter dado todo apoio e incentivo para prosseguir.

Ao Prof. Dr. Edvan Cirino da Silva, pela orientação e ensinamentos.

Ao Prof. Dr. Mário Cesar Ugulino de Araújo (Coord. do LAQA), por ter me recebido no LAQA e pelas várias oportunidades de crescimento profissional concedidas e co-orientação deste trabalho.

Ao Prof. Dr. Germano Veras (UEPB) orientador de Iniciação Científica e grande amigo.

Aos grandes amigos que fiz durante minha graduação que levarei por toda minha vida: Gildo William, Odilon, Crislane, Elida Medeiros, Jonathan, Ilza, Francisco, Geovana, Rose, Priscila, Gean, David, Marcelo, Marcelo Barbosa, Anna Luiza.

Ao grande e especial amigo José Crispim pelo companheirismo.

Aos Professores do PPGQ-UFPB pelos ensinamentos durante as disciplinas cursadas.

Aos novos amigos que ganhei no LAQA durante o mestrado: Edilene, Wellington, Paulo Henrique, Sofácles, Pablo, Aline, Anabel, Marcelo, Heberty, Chicote, Stefani, Willame, Fátima, Renato, Cleison, Flaviano, Inakã, Berivaldo, Dayse, Dani, Karla, Neto, Daniel, Jerfeson, Jardel, Urijatam, Fátima em fim a todos que fazem o LAQA.

Ao PPGQ-UFPB e a Capes pela bolsa concedida.

SUMÁRIO

LISTA DE FIGURAS.....	xii
LISTA DE TABELAS.....	xv
LISTA DE ABREVIATURAS.....	xvi
RESUMO	xviii
ABSTRACT	xix
1.0 INTRODUÇÃO.....	21
1.1 Caracterização geral da problemática	21
1.2 Objetivo Geral	22
1.2.1 Objetivos específicos	22
2. FUNDAMENTAÇÃO TEÓRICA	25
2.1 Calibração Multivariada	25
2.1.1 Organização dos dados em calibração multivariada	26
2.1.2 Métodos Clássicos de calibração	27
2.1.3 Métodos Inversos de calibração	28
2.1.4 Regressão Linear Múltipla-MLR.....	28
2.1.5 Regressão em componentes principais	29
2.1.6 Regressão em mínimos quadrados parciais	31
2.2 Seleção de Varáveis em Calibração Multivariada	40
2.3 Seleção de Variáveis em Regressão PLS	41
2.3.1 Busca Exaustiva.....	42
2.3.2 Algoritmo Genético.....	42
2.3.3 Método de eliminação de variáveis não informativas (UVE)	46
2.3.4 Jack-Knife.....	48
2.3.5 Colônia de formigas	48
2.3.6 PLS em intervalos - iPLS	50
2.3.8 Backward PLS.....	51

2.3.9 PLS em intervalos sinérgicos - siPLS	52
2.3.10 OPS-PLS	53
2.3.11 Busca de Tabu	54
2.3.11 Ponderação Iterativa dos Preditores	55
2.4 Algoritmo das Projeções Sucessivas	56
3.0 Metodologia	62
3.1 SPA em regressão PLS	62
3.2 Algoritmos usados para comparação	68
3.2.1 GA-PLS.....	68
3.2.2 iPLS e siPLS	68
3.2.3 Jack-Knife-PLS	69
3.2.4 Comparação dos modelos.....	69
3.3 Estudos de caso	69
3.3.1 Determinações de corantes alimentícios	69
3.3.2 Determinação do teor de proteínas em amostras de trigo	72
3.3.3 Determinação da qualidade de amostras de extrato de cerveja	73
4.0 Determinações de corantes alimentícios.....	75
4.1 Quantificação dos corantes	78
5.0 DETERMINAÇÃO DO TEOR DE PROTEÍNA EM AMOSTRAS DE TRIGO	90
5.1 Quantificação do teor de proteína	91
6.0 Determinação da qualidade de amostras de extrato de cerveja	99
7.0 CONCLUSÃO	107
7.1 Propostas futuras	108
REFERÊNCIAS.....	109

LISTA DE FIGURAS

Figura 2.1	Organização matricial dos dados multivariados de primeira ordem.	27
Figura 2.2	Decomposição de uma matriz em componentes principais.	30
Figura 2.3	Interpretação geometria das projeções em regressão PLS (adptado [17]).	35
Figura 2.4	Exemplo de ajuste por OLS (mínimos quadrados ordinários) entre valor de referencia e valor predito.	38
Figura 2.5	Codificação binária usada na seleção de variáveis do GA.	44
Figura 2.6	Esquema de cruzamento no GA.	45
Figura 2.7	Esquema de processo de mutação no GA.	45
Figura 2.8	Exemplo do processo de eliminação de variáveis.	47
Figura 2.9	Esquema do algoritmo BVSPLS. Adaptado [56]	52
Figura 2.10	Esquema do algoritmo OPS. Adaptado de [59].	54
Figura 2.11	Ilustração da seqüência de projeções realizadas pelo SPA. (a): Primeira iteração. (b): Segunda iteração. Nesse exemplo, a cadeia de variáveis que inicia em x_3 deverá ser $\{x_3, x_1, x_5\}$ Adaptado de [62].	59
Figura 3.1	Saída gráfico do modelo global PLS calculado no inicio da execução do algoritmo iSPA-PLS.	62
Figura 3.2	Ilustração da montagem da matriz iXcal usada na etapa de projeção do SPA.	64
Figura 3.3	Ilustração do relatório de saída do algoritmo iSPA-PLS	65
Figura 3.4	Saída gráfica do iSPA-PLS (a) valor predito versus referência, (b) Intervalos selecionados.	66
Figura 3.5	Fluxograma do algoritmo iSPA-PLS.	67
Figura 3.6	Interface gráfica do iToolBox.	68

Figura 3.7	Espectros de absorção UV-VIS e estruturas moleculares dos corantes puros.	70
Figura 4.1	Espectros de absorção no visível das amostras de calibração.	75
Figura 4.2	RMSECV versus número de fatores PLS incluídos no modelo global (a) amarelo crepúsculo, (b) vermelho 40 e (c) tartrazina.	76
Figura 4.3	Resíduos de validação cruzada para (a) AC um fator, (b) AC 2 fatores, (c) AC 3 fatores, (d) V40 1 fator, (e) V40 2 fatores, (f) V40 3 fatores, (g) TAR 1 fator, (h) TAR 2 fatores e (i) TRA 3 fatores.	77
Figura 4.4	Variáveis selecionadas pelos algoritmos (a) Jack-Knife e (b) GA para o corante AC.	84
Figura 4.5	Variáveis selecionadas pelos algoritmos (a) Jack-Knife e (b) GA para o corante TAR.	84
Figura 4.6	Variáveis selecionadas pelos algoritmos (a) Jack-Knife e (b) GA para o corante V40.	85
Figura 4.7	Variáveis selecionadas pelos algoritmos: (a) iPLS-5 intervalos, (b) iPLS-10 intervalos, (c) iPLS-15 intervalos, (d) siPLS-5 intervalos em combinação de 2, (e) siPLS-10 intervalos em combinação de 2, (f) siPLS-15 intervalos em combinação de 2, (g) siPLS-5 intervalos em combinação de 3, (h) siPLS-10 intervalos em combinação de 3, (i) siPLS-15 intervalos em combinação de 3, (j) iAPS-PLS-5 intervalos, (l) iAPS-PLS-10 intervalos e (m) iAPS-PLS-15 intervalos para o corante AC.	86
Figura 4.8	Variáveis selecionadas pelos algoritmos: (a) iPLS-5 intervalos, (b) iPLS-10 intervalos, (c) iPLS-15 intervalos, (d) siPLS-5 intervalos em combinação de 2, (e) siPLS-10 intervalos em combinação de 2, (f) siPLS-15 intervalos em combinação de 2, (g) siPLS-5 intervalos em combinação de 3, (h) siPLS-10 intervalos em combinação de 3, (i) siPLS-15 intervalos em combinação de 3, (j) iAPS-PLS-5 intervalos, (l) iAPS-PLS-10 intervalos e (m) iAPS-PLS-15 intervalos para o corante TAR.	87
Figura 4.9	Variáveis selecionadas pelos algoritmos: (a) iPLS-5 intervalos, (b) iPLS-10 intervalos, (c) iPLS-15 intervalos, (d) siPLS-5 intervalos em combinação de 2, (e) siPLS-10 intervalos em combinação de 2, (f) siPLS-15 intervalos em combinação de 2, (g) siPLS-5 intervalos em combinação de 3, (h) siPLS-10 intervalos em combinação de 3, (i) siPLS-15 intervalos em combinação de 3, (j) iAPS-PLS-5 intervalos, (l) iAPS-PLS-10 intervalos e (m) iAPS-PLS-15 intervalos para o corante V40.	88

Figura 5.1	Espectros brutos das amostras de trigo.	90
Figura 5.2	Espectros derivativos das amostras de trigo.	
Figura 5.3	RMSECV versus número de fatores PLS incluídos no modelo global	91
Figura 5.4	Variáveis selecionadas pelos algoritmos (a) Jack-Knife e (b) GA.	
Figura 5.5	Figura 5.5: Variáveis selecionadas pelos algoritmos: (a) iPLS-10 intervalos, (b) iPLS-15 intervalos, (c) iPLS-20 intervalos, (d) siPLS-10 intervalos em combinação de 2, (e) siPLS-15 intervalos em combinação de 2, (f) siPLS-20 intervalos em combinação de 2, (g) siPLS-10 intervalos em combinação de 3, (h) siPLS-15 intervalos em combinação de 3, (i) siPLS-20 intervalos em combinação de 3, (j) iAPS-PLS-10 intervalos, (l) iAPS-PLS-15 intervalos e (m) iAPS-PLS-20 intervalos.	96
Figura 6.1	Espectros brutos das amostras de extrato de cervejas.	
Figura 6.2	coeficiente de correlação entre ycal e Xcal .	100
Figura 6.3	RMSECV versus número de fatores PLS incluídos no modelo global	100
Figura 6.4	Variáveis selecionadas (a) pelo Jack-Knife e (b) GA.	103
Figura 6.5	Variáveis selecionadas pelos algoritmos: (a) iPLS-10 intervalos, (b) iPLS-15 intervalos, (c) iPLS-20 intervalos, (d) siPLS-10 intervalos em combinação de 2, (e) siPLS-15 intervalos em combinação de 2, (f) siPLS-20 intervalos em combinação de 2, (g) siPLS-10 intervalos em combinação de 3, (h) siPLS-15 intervalos em combinação de 3, (i) siPLS-20 intervalos em combinação de 3, (j) iAPS-PLS-10 intervalos, (l) iAPS-PLS-15 intervalos e (m) iAPS-PLS-20 intervalos.	105

LISTA DE TABELAS

Tabela 3.1	Configurações do GA.	68
Tabela 3.2	Concentração dos corantes do conjunto de calibração [mg L ⁻¹].	71
Tabela 3.3	Matriz completa do planejamento Brereton com 03 níveis e 03 fatores das misturas para conjunto de calibração [mg L ⁻¹].	71
Tabela 3.3	Matriz completa do planejamento Brereton com 03 níveis e 03 fatores das misturas para conjunto de validação externa [mg L ⁻¹].	72
Tabela 4.1	Resumo dos resultados de calibração e validação externa para o corante tartrazina.	79
Tabela 4.2	Resumo dos resultados de calibração e validação externa para o corante amarelo crepúsculo.	80
Tabela 4.3	Resumo dos resultados de calibração e validação externa para o corante vermelho-40.	81
Tabela 4.4	Valores de F calculado para o corante tartrazina	82
Tabela 4.5	Valores de F calculado para o corante amarelo crepúsculo.	82
Tabela 4.6	Valores de F calculado para o corante vermelho 40.	83
Tabela 5.1	Quantificação de proteína em trigo: parâmetros estatísticos	92
Tabela 5.2	Valores de F calculado para comparação dos modelos iSPA-PLS com os demais algoritmos..	93
Tabela 6.1	Parâmetros estatísticos para calibração e predição da qualidade em amostras de cervejas.	101
Tabela 6.2	Valores de F calculado para comparação dos modelos iSPA-PLS com os demais algoritmos	102

LISTA DE ABREVIATURAS

- ACF**- Algoritmo Colônia de Formigas
- ANOVA**- Análise de Variância
- APS**- Algoritmo das projeções Sucessivas
- ASTM**- Associação Americana para Teste de materiais
- BT**- Busca de Tabu
- BVSPLS**- Seleção de Variáveis backward em mínimos quadrados parciais
- GA**- Algoritmo Genético
- iPLS**- Regressão em mínimos quadrados parciais por intervalos
- iSPA-PLS** - Algoritmo das projeções Sucessivas por intervalos
- LDA**- Análise discriminante linear
- MLR**- Regressão Linear Múltipla
- MMQ**- Método dos Mínimos Quadrados
- NAS**- Sinal analítico Líquido
- NIPALS**- Mínimos Quadrados Parciais Iterativos Não Linear
- NIR**- Infravermelho próximo
- OLS**- Mínimos quadrados Ordinários
- PCR**- Regressão em Componentes Principais
- PLS**- Mínimos Quadrados Parciais
- QSAR**- Relação Quantitativa Atividade Estrutura
- QSPR**- Relação Quantitativa Propriedade Estrutura
- QVS**- quantidade de variáveis selecionadas
- RMSEC**- Raiz quadrada do erro médio quadrático de calibração
- RMSECV**- Raiz quadrada do erro médio quadrático de validação cruzada
- RMSEV**- Raiz quadrada do erro médio quadrático de validação

RMSEP- Raiz quadrada do erro médio quadrático de previsão

siPLS- Mínimos Quadrados Parciais em intervalos sinérgicos

SPXY – partição de amostras usando X e Y

UV-Vis- Ultravioleta visível

UVE- Eliminação de variáveis não-informativas

RESUMO

A combinação de técnicas espectroscópicas com calibração multivariada tem permitido o desenvolvimento de métodos para determinação de analitos (ou outras propriedades) em matrizes complexas. Nesse contexto, destacam-se as determinações usando modelos baseados na regressão PLS (*Partial Least Square*), bem difundida e consolidada na literatura. Apesar da eficácia dos modelos PLS obtidos a partir de espectros completos, alguns trabalhos da literatura têm mostrado que a seleção de variáveis pode melhorar a capacidade preditiva dos modelos PLS. No presente trabalho, desenvolve-se um algoritmo, em *MatLab*[®], que utiliza o Algoritmo das Projeções Sucessivas-APS, proposto originalmente para MLR (*Multiple Linear Regression*), a fim de melhorar a capacidade preditiva de modelos PLS obtidos por intervalos. O algoritmo proposto, denominado Algoritmo das projeções sucessivas em intervalos para regressão PLS (iSPA-PLS), foi avaliado em três estudos de caso, a saber: (i) determinação simultânea de três corantes alimentícios em amostras sintéticas usando espectrometria UV-Vis, (ii) quantificação do teor de proteínas em trigo por espectrometria NIR e (iii) determinação da qualidade de amostras de extrato de cervejas usando também espectrometria NIR. O desempenho do iSPA-PLS foi comparado ao dos seguintes algoritmos e modelos bem estabelecidos na literatura: GA-PLS, PLS-Jack-Knife, iPLS e siPLS. Os resultados das três aplicações atestam as vantagens do iSPA-PLS frente aos demais algoritmos. Entre elas, destacam-se os menores erros de predição e a capacidade de selecionar um número menor de fatores PLS.

Palavras-chaves: Algoritmo das projeções Sucessivas, Mínimos Quadrados Parciais, seleção de variáveis, Intervalos.

ABSTRACT

Spectroscopy techniques combined with multivariate calibration have allowed the development of methods for analyte determinations (or other properties) in complex matrices. In this context, it can be mentioned the determinations that uses models based on PLS (Partial Least Square) regression, which is well established and consolidated in literature. In spite of efficiency of PLS models obtained from full spectrum, some papers reported in literature show that a variable selection may improve the predictive ability of the PLS models. In the present work, it was developed an algorithm, in Matlab®, that employs the SPA (Successive Projection Algorithm), originally proposed for MLR (*Multiple Linear Regression*), in order to improve the predictive ability of interval PLS models. The proposed algorithm, termed iSPA-PLS, was evaluated in three case studies, namely: (i) simultaneous determination of three artificial colorants by UV-VIS spectrometry, (ii) quantification of protein contents in wheat using NIR spectrometry, and (iii) quality determination of samples of beer extract using NIR spectrometry too. The performance of iSPA-PLS was compared to the following well-established algorithms and methods: GA-PLS, PLS-Jack-Knife, iPLS e siPLS. In all applications, the results show that the iSPA-PLS presented some advantageous when compared to other algorithms used for comparison. The main advantageous include the smallest errors of prediction and the capacity of selecting a smaller number of PLS factors.

Keywords: Successive Projection Algorithm, Partial Least Square, Variable Selection, Intervals.

Capítulo 1

Introdução

Sempre é preciso saber quando uma etapa
chega ao final..

Se insistirmos em permanecer nela mais do que o
tempo necessário, perdemos a alegria e o sentido das
outras etapas que precisamos viver.

1.0 INTRODUÇÃO

1.1 Caracterização geral da problemática

Com o avanço da instrumentação analítica e a capacidade de processamento de dados pelos microcomputadores, é possível gerar com rapidez uma grande quantidade de informação sobre a(s) propriedade(s) de interesse na amostra. Nesse contexto, destacam-se a espectroscopia no ultravioleta e visível, espectroscopia no infravermelho, cromatografia, espectrofluorimetria, etc^[1]. Tais informações podem ser modeladas, via análise de regressão, com intuito de construir modelos matemáticos capazes de prever o parâmetro (concentração de analito, umidade, etc) objeto da análise.

Em modelagem multivariada usando dados espectroscópicos, nem todas as variáveis apresentam correlação com os parâmetros de interesse ou, quando se correlacionam, possuem um ruído excessivo. Assim, a seleção de um subconjunto de variáveis espectrais constitui uma ferramenta de pré-processamento de dados capaz de melhorar a capacidade preditiva e aumentar a robustez dos modelos, bem como facilitar a interpretação dos resultados. Neste contexto, vários algoritmos de seleção de variáveis têm sido relatados na literatura ^[39-40] e aplicados à calibração multivariada baseada em regressão linear múltipla (*Multiple Linear Regression, MLR*), uma vez que os modelos MLR apresentam algumas limitações matemáticas frente a dados com forte multicolinearidade entre as variáveis independentes.

Por muito tempo, defendeu-se a idéia do uso do sinal analítico completo (conhecidos modelos "*full spectrum*", quando se emprega dados espectroscópicos, comum em química analítica) em conjunto com ferramentas quimiométricas, tais como Regressão por Componentes Principais (*Principal Component Regression, PCR*) e regressão por Mínimos Quadrados Parciais (*Partial Least Square Regression, PLS*). Isto porque esses métodos de calibração

multivariada promovem uma transformação nos dados originais para reduzir a dimensionalidade da matriz e estabelecer ortogonalidades entre as novas variáveis. No caso do PLS a ortogonalidade é quebrada por meio de uma rotação que visa aumentar a correlação das novas variáveis com o parâmetro de interesse.

A despeito da eficiência dos modelos PLS baseados em espectros completos (*full spectrum*), alguns trabalhos [47-56] têm apontado que a seleção de variáveis pode melhorar a capacidade preditiva dos mesmos. Uma das razões principais é que a seleção de variáveis remove canais analíticos que não guardam correlação com o parâmetro de interesse e não linearidade dos dados.

Em modelos de regressão PLS as técnicas de seleção de variáveis pode ser categorizadas em métodos de seleção de variáveis individuais e métodos de seleção de intervalos [39-40]. Segundo Höskuldsson [43] os métodos de seleção de intervalos são superiores quando comparados aos que fazem seleção de variáveis individuais em termos de estabilidade numérica dos coeficientes de regressão, contudo existem poucos métodos para tal modalidade de seleção de variáveis.

1.2 Objetivo Geral

O objetivo central deste trabalho é desenvolver um algoritmo em ambiente *MatLab* de modo a usar o Algoritmo das Projeções Sucessivas-SPA, proposto originalmente em conjunto com regressão MLR, para melhorar a capacidade preditiva de modelos PLS atuando com ferramenta de seleção de variáveis em intervalo.

1.2.1 Objetivos específicos

- ✓ Desenvolver o Algoritmo denominado iSPA-PLS para fazer seleção de variáveis em intervalos para construção de modelos PLS;

- ✓ Avaliar a capacidade preditiva do algoritmo propostos em três estudos de caso: (i) determinação simultânea de três corantes alimentícios em amostras sintéticas usando espectrometria UV-Vis, (ii) quantificação do teor de proteínas em trigo por espectrometria NIR e (iii) determinação da qualidade de amostras de extrato de cervejas usando também espectrometria NIR.
- ✓ Comparar os resultados obtidos com os métodos já consolidados na literatura.

Capítulo 2

Fundamentação

Teórica

2. FUNDAMENTAÇÃO TEÓRICA

2.1 Calibração Multivariada

A determinação da concentração de uma dada espécie química (analito) em uma amostra é na maioria dos casos, o alvo ou objetivo dos métodos analíticos. Contudo, a concentração não é uma grandeza mensurável diretamente, e deve ser estimada de forma indireta por meio da medida, em geral, de uma propriedade física (absorção de luz por moléculas, por exemplo) das amostras. No contexto da análise quantitativa instrumental, a propriedade mensurada guarda, geralmente, uma relação linear com a concentração do analito ^[1].

O procedimento matemático e estatístico usado para relacionar medidas de uma propriedade física das amostras com um dado constituinte ou parâmetro físico-químico é chamado de calibração. O caso mais simples é a calibração univariada ou calibração de ordem zero ^[2]. A calibração univariada é bem estabelecida na literatura e amplamente difundida em métodos analíticos de referência.

Na calibração univariada, o sinal medido em um único canal analítico (comprimento de onda no espectro, potencial fixado num voltamograma, etc) é relacionado com a concentração do analito nas soluções (ou amostras) de calibração. Frequentemente utiliza-se o método dos mínimos quadrados ordinários ^[3] (*ordinary least squares*, OLS) pra estimar os coeficientes de regressão do modelo de calibração.

A relação linear estabelecida entre a variável aleatória (y , sinal analítico) e a variável assumida como não aleatória (x , concentração das amostras de calibração) é expressa pela **Equação 1**.

$$y_i = b_0 + b_1 x_i \quad (1)$$

As estimativas b_0 e b_1 são obtidas minimizando a soma dos quadrados das diferenças entre valor medido y_i e o valor predito \hat{y}_i , também chamado de minimizar a soma dos quadrados dos resíduos. A qualidade do modelo construído é avaliada por meio de uma análise de variância (Analysis of variance, ANOVA), no qual se avalia a significância da correlação linear entre x e y , e uma possível existência de falta de ajuste no modelo [4].

A principal vantagem da calibração univariada é a simplicidade da matemática envolvida. Contudo, com base na **Equação 1**, percebe-se que é necessária a seletividade do sinal medido. Em muitas situações, o uso de medidas em um único canal não é suficiente para descrever quantitativamente o sistema, a exemplo da calibração baseada em espectros NIR [5], devido às sobreposições promovidas pelos interferentes sobre o sinal analítico em matrizes complexas.

A calibração multivariada emprega medidas realizadas em múltiplos canais na construção do modelo para relacionar concentração e sinal analítico. Esse processo surge como alternativa capaz de superar as limitações da calibração univariada, permitindo a determinações simultâneas de analitos com maior sensibilidade e confiabilidade na presença de interferentes. Para isso, os interferentes devem também estar presentes na etapa de calibração, o que é conhecido como vantagem de primeira ordem [6-7].

2.1.1 Organização dos dados em calibração multivariada

Na construção de modelos de calibração multivariada (calibração de primeira ordem) a informação analítica associada a cada amostra corresponde a um vetor. Os dados a serem modelados são dispostos em uma matriz, como ilustrado na **Figura 2.1**, que contém os dados de espectros para um conjunto de amostras. A

matriz de dados é composta de tal modo que cada linha corresponde a uma amostra e cada coluna contém a informação referente a um canal analítico (neste caso, cada comprimento de onda no espectro).

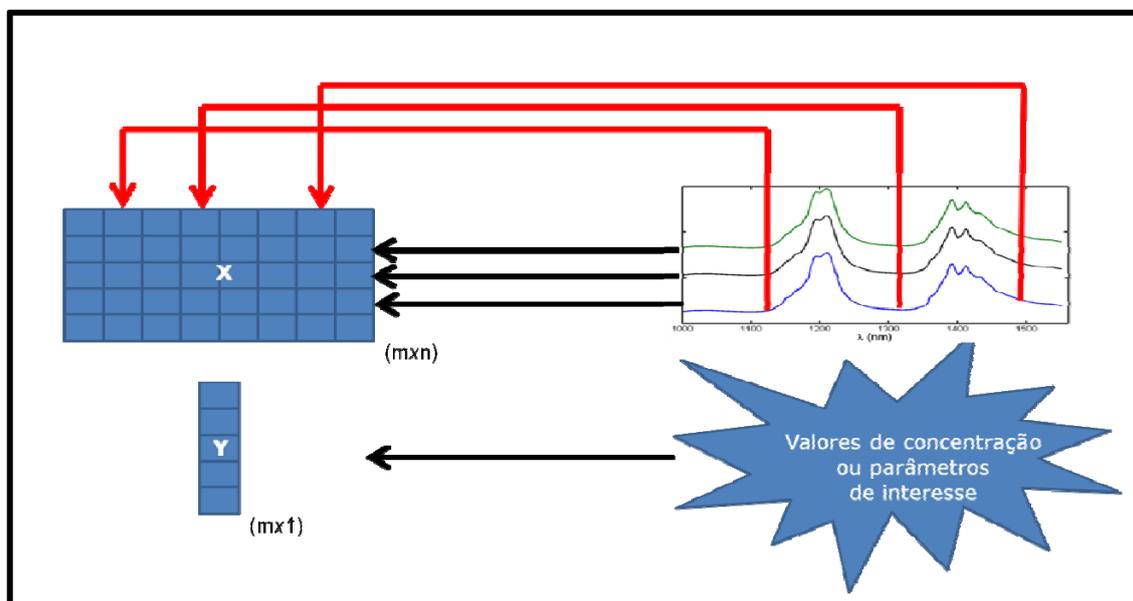


Figura 2.1- Organização matricial dos dados multivariados de primeira ordem.

Convencionalmente, a matriz que contém as variáveis independentes, comumente chamadas de matriz de respostas instrumentais, é denominada matriz **X**. O vetor contendo a variável dependente ou parâmetro de referência é denotado por **y**.

Para maior clareza, a seguinte convenção será adotada neste trabalho: letras maiúsculas e negritas representam matrizes, letras minúsculas em negritos representam vetores linha, letras minúsculas itálico representam escalares e o sobrescrito " ` " indica vetor ou matriz transposta.

2.1.2 Métodos Clássicos de calibração

Historicamente, os modelos de calibração onde o sinal analítico é função da concentração são ditos métodos clássicos, por

apresentarem uma relação de proporcionalidade tal qual a prevista na Lei de Lambert-Beer ^[8]. Matematicamente, um modelo clássico pode ser expresso por:

$$\mathbf{X} = \mathbf{K} * \mathbf{Y} \quad (2)$$

Em que \mathbf{X} é a matriz de respostas instrumentais, \mathbf{Y} é a matriz das concentrações e \mathbf{K} é a matriz que contém o sinal puro de cada componente da mistura. Se \mathbf{K} é obtido a partir de medidas experimentais o método é denominado clássico direto (*Direct Classical Least Square- DCLS*). Por outro lado, se \mathbf{K} é estimado empregando \mathbf{X} e \mathbf{Y} o método é dito clássico indireto (Indirect Classical Least Square-ICLS) ^[9].

A principal vantagem dos métodos clássicos é a simplicidade dos cálculos envolvidos. Todavia, a necessidade de dispor ou estimar a matriz \mathbf{K} significa ter que conhecer todas as substâncias que geram sinal analítico. Este fato restringe as aplicações dos métodos clássicos.

2.1.3 Métodos Inversos de calibração

Os métodos inversos assumem que a concentração é uma função do sinal analítico medido. Embora contrarie a lei de Lambert-Beer para caso de medidas de absorvância, contorna uma série de restrições dos métodos clássicos por empregar a estrutura de variância/covariância da matriz \mathbf{X} na modelagem dos dados.

2.1.4 Regressão Linear Múltipla-MLR

Na regressão linear múltipla, assume-se que existe uma relação linear entre as matrizes $\mathbf{X}_{(m \times n)}$ e $\mathbf{y}_{(m \times 1)}$, como mostrado na **Equação 3**, em que $\mathbf{E}_{(m \times 1)}$ representa o resíduo não modelado em \mathbf{y} ^[10].

$$\mathbf{y} = \mathbf{X}\mathbf{b}_{\text{MLR}} + \mathbf{E} \quad (3)$$

Na etapa de calibração o vetor de regressão \mathbf{b} é estimado empregando método mínimos quadrados ordinários (com base na **Equação 4**).

$$\mathbf{b}_{\text{MLR}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (4)$$

A resolução da **Equação 4**, para obter o vetor dos coeficientes de regressão (\mathbf{b}), requer a inversão da matriz $(\mathbf{X}'\mathbf{X})$ e esta operação algébrica envolve algumas suposições acerca dos dados:

- ✓ O número de amostras de calibração deve ser maior ou igual ao número de variáveis ($m > n$), caso contrário o sistema de equações será indeterminado.
- ✓ As variáveis (colunas de \mathbf{X}) devem ser idealmente vetores linearmente independentes. A violação desta suposição pode levar a uma matriz singular.

Em dados espectroscópicos estas suposições impossibilitam o uso da regressão MLR sem a realização de uma seleção previa de variáveis. Isso impede o uso do MLR aos problemas envolvendo matrizes de dados com alta dimensionalidade.

2.1.5 Regressão em componentes principais

A regressão em componentes principais é um método de calibração de primeira ordem que, ao contrário do MLR, não necessita de uma seleção de variáveis prévia para contornar o problema de multicolinearidade dos dados. Em vez disso, faz uso de uma transformação ortogonal da matriz \mathbf{X} , de modo a obter um novo conjunto de variáveis linearmente independentes ^[11].

A decomposição da matriz \mathbf{X} realizada em PCR é uma análise por componentes principais (*principal component analysis*, PCA). Em PCA uma matriz de alta dimensão é decomposta em duas pequenas matrizes escores (\mathbf{T}) e pesos (\mathbf{P}), de acordo com a **Equação 5**, em que \mathbf{E} representa os resíduos de \mathbf{X} [12].

$$\mathbf{X} = \mathbf{T} * \mathbf{P}' + \mathbf{E} \quad (5)$$

Portanto cada componente principal fica caracterizada por um vetor de escores (\mathbf{t}) que corresponde a t -ésima coluna de \mathbf{T} , um vetor de pesos (\mathbf{p}) correspondente a p -ésima coluna de \mathbf{P} . Geometricamente, a decomposição de \mathbf{X} ocorre conforme mostrado na **Figura 2.2**:

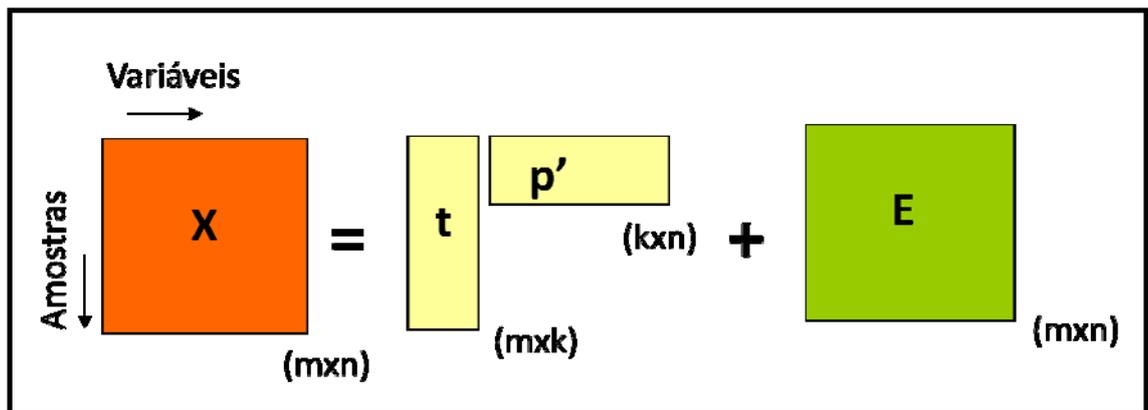


Figura 2.2- Decomposição de uma matriz em componentes principais.

Diversas metodologias são conhecidas para decompor matrizes com alta dimensão em suas componentes principais, a exemplo da decomposição em valores singulares (Singular Value Decomposition, SVD) [13], o qual usa a matriz de dados \mathbf{X} e decomposição em autovalores (Eigenvalue Decomposition, EVD) [14] que trabalha com a matriz de produto cruzado $\mathbf{X}'\mathbf{X}$. Além destes, são utilizados métodos iterativos como o mínimos quadrados parciais iterativo não linear (non linear iterative partial least squares, NIPALS) [15].

Portanto a regressão PCR faz uso da matriz \mathbf{T} , que é ortogonal, para obter o vetor dos coeficientes de regressão \mathbf{b}_{PCR} empregando o método dos mínimos quadrados (OLS) similar ao MLR, de acordo com a **Equação 6**.

$$\mathbf{y} = \mathbf{T}_{(m \times k)} \mathbf{b}_{\text{PCR}} + \mathbf{F} \quad (6)$$

Onde k corresponde ao número de componentes principais empregados na obtenção dos coeficientes de regressão e \mathbf{F} aos resíduos não modelados.

2.1.6 Regressão em mínimos quadrados parciais

A regressão em mínimos quadrados parciais foi desenvolvido por Herman Wold e colaboradores no período 1975 a 1982. Assim como o PCR, na modelagem PLS a matriz \mathbf{X} sofre uma decomposição. Entretanto, ao contrário do PCR, no PLS usa-se a informação contida em \mathbf{y} na obtenção dos fatores ^[16-17].

Na PCR os pesos são computados de modo a descreverem a maior fração possível de variância. Contudo, isto não necessariamente garante uma boa correlação com \mathbf{y} . Calculando os pesos de forma que o produto da variância de \mathbf{X} com a correlação de $\mathbf{X}\mathbf{P}$ com \mathbf{y} seja maximizada, estamos otimizando a decomposição de \mathbf{X} para uma melhor predição de \mathbf{y} . Essa otimização ocasiona pequenas distorções nas direções dos *pesos*, de modo que, rigorosamente eles perdem a ortogonalidade, levando a pequenas redundâncias de informação ^[18].

Atualmente na literatura se encontram disponíveis diversas formas de obter os parâmetros de um modelo PLS ^[19]. Como exemplos, destacam-se o algoritmo de escores não ortogonalizado desenvolvido por Martens ^[13], e o mais conhecido e utilizado NIPALS proposto por Wold ^[7].

Teoricamente o uso de qualquer um dos algoritmos mencionados acima para obter vetor de regressão PLS deve levar ao mesmo resultado. Contudo, como investigado por Andersson [19], do ponto de vista numérico existem diferenças. Em outras palavras, o mesmo conjunto de dados, sendo modelado por algoritmos PLS diferentes, leva a resultados diferentes. A magnitude dessa diferença está associada à natureza dos dados, ao número de fatores PLS empregado e precisão usada nos cálculos.

No presente trabalho, o algoritmo NIPALS [7,17] foi empregado na obtenção dos modelos PLS. Nesta abordagem os parâmetros do modelo PLS são obtidos de forma iterativa um por vez. Na modelagem PLS as matrizes $\mathbf{X}_{(m \times n)}$ e $\mathbf{Y}_{(m \times z)}$, são decompostas conforma as equações abaixo:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E}_X = \sum t_k \mathbf{p}'_k + \mathbf{E}_x \quad (7)$$

$$\mathbf{Y} = \mathbf{UQ}' + \mathbf{E}_y = \sum u_k \mathbf{q}'_k + \mathbf{E}_y \quad (8)$$

Entre os *escores* de \mathbf{X} e os *escores* de \mathbf{Y} , uma relação linear é, então, estabelecida.

$$\mathbf{u}_k = \mathbf{b}_k \mathbf{t}_k \quad (9)$$

Em que \mathbf{b} é o vetor dos coeficientes de regressão para k fatores, que é obtido por meio da expressão.

$$\mathbf{b}_k = \frac{\mathbf{u}'_k \mathbf{t}_k}{\mathbf{t}'_k \mathbf{t}_k} \quad (10)$$

Para obtenção dos fatores PLS os seguintes passos são seguidos de modo a levar em consideração a informação de \mathbf{y} na decomposição da matriz \mathbf{X} .

1: faça $\mathbf{y} = \mathbf{u}_k$,

Enquanto o critério de convergência não for atingido.

2: Faça

Estimam-se os pesos de \mathbf{X} :

$$\mathbf{W}'_k = \frac{\mathbf{u}'_k * \mathbf{X}}{\mathbf{u}'_k * \mathbf{u}_k} \quad (11)$$

Os pesos de \mathbf{X} obtidos na **Equação 11** são normalizados para comprimento 1.

$$\mathbf{W}'_{k,norm} = \frac{\mathbf{W}'_k}{norm(\mathbf{W}'_k)} \quad (12)$$

Os escores de \mathbf{X} baseada nos pesos estimados em 11 é dados por:

$$\mathbf{t}_k = \frac{\mathbf{X}\mathbf{W}'_{k,norm}}{\mathbf{W}'_k\mathbf{W}_k} \quad (13)$$

Estima se o conjunto de pesos de \mathbf{y} empregando a **Equação 14** que posteriormente são normalizados para comprimento 1 de acordo com a **Equação 15**.

$$\mathbf{q}'_k = \frac{\mathbf{t}'_k * \mathbf{Y}}{\mathbf{t}'_k * \mathbf{t}_k} \quad (14)$$

$$\mathbf{q}'_{k,norm} = \frac{\mathbf{q}'_k}{norm(\mathbf{q}'_k)} \quad (15)$$

Estima se os escores de \mathbf{Y}

$$\mathbf{u}_k = \frac{Y\mathbf{q}_k}{\mathbf{q}_k' \mathbf{q}_k} \quad (16)$$

3: Checa a convergência

Se atingido o critério de convergência, segue para o passo 4, caso contrário o passo 2 é repetido.

4: Como o algoritmo não fornece os valores de \mathbf{t} ortogonais, os valores de \mathbf{p}' são substituídos por \mathbf{w}' e um passo extra é incluído depois da convergência tornando os valores de \mathbf{t} ortogonais.

$$\mathbf{p}_k = \frac{\mathbf{t}_k' * X}{\mathbf{t}_k' * \mathbf{t}_k} \quad (17)$$

5: Os efeitos do k -ésimo fator é removido pela subtração do produto dos escores e pesos da matriz original:

$$\mathbf{X}_k = \mathbf{X}_{k-1} - \mathbf{t}_k \mathbf{p}_k \quad (18)$$

$$\mathbf{Y}_k = \mathbf{Y}_{k-1} - \mathbf{u}_k \mathbf{q}_k \quad (19)$$

Os passos 2, 3, 4 e 5 são repetidos até que os k fatores tenham sido calculados. Na **Figura 2.3** abaixo é apresentada uma interpretação geométrica das projeções realizadas na decomposição PLS ^[17].

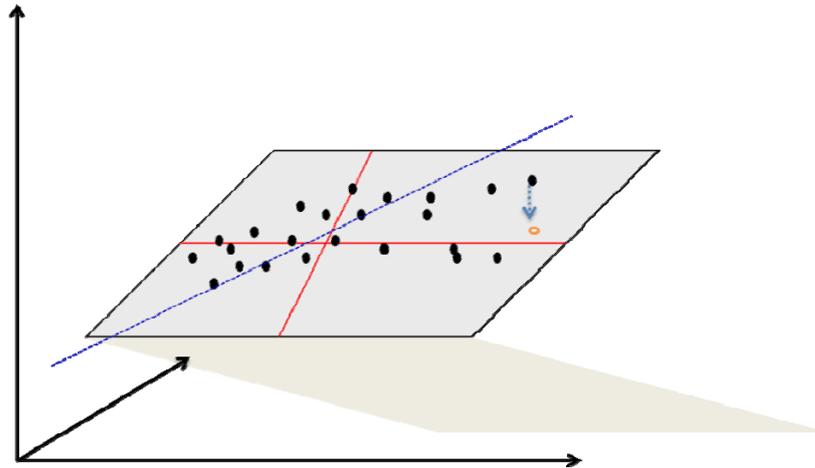


Figura 2.3- Interpretação geometria das projeções em regressão PLS (adptado [17]).

As linhas pretas representam os preditores originais de \mathbf{y} e os pontos pretos a representação de cada objeto no espaço n dimensional. O plano em cinza representa o espaço definido pelos dois primeiros fatores (linhas vermelhas) que apontam na direção de máxima variância e a linha pontilhada em azul mostra a rotação que deve ser promovida nos fatores para obter projeções de \mathbf{X} que são bons preditores de \mathbf{y} [22].

2.1.6.1 Ferramentas de diagnóstico

Após a obtenção dos parâmetros do modelo de regressão é necessário avaliar a qualidade deste antes de empregá-lo na predição de amostras desconhecidas, que é o objetivo final de qualquer modelo de regressão. As métricas utilizadas para estes propósitos são conhecidas na literatura como figuras de mérito [20], que atestam que se o método proposto é confiável e atende às exigências dos órgãos de fiscalização.

Algumas métricas como, por exemplo, exatidão, precisão, robustez, ajuste e erro sistemático (*bias*) não apresentam maiores dificuldades e são estimadas de maneira bastante similar aos

métodos univariados. Contudo, figuras de mérito como linearidade, sensibilidade, razão sinal/ruído, seletividade e intervalos de confiança ou incerteza são bastante distintos daqueles estimados para métodos univariados, e em sua grande maioria, fazem uso do conceito de sinal analítico líquido (NAS – Net Analyte Signal) [21].

Para as métricas que não fazem uso do NAS, os desvios entre o valor real do parâmetro de interesse e o valor estimado pelo modelo é a base das equações utilizadas. A raiz quadrada do erro médio quadrático de calibração (RMSEC – Root Mean Squares Error of Calibration), dado pela **Equação 20**, é um parâmetro que incorpora erros aleatórios e sistemáticos do modelo [9].

$$RMSEC = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{(m-k-1)}} \quad (20)$$

Onde y_i é o valor de referência e \hat{y}_i é o valor predito pelo modelo, m é a quantidade de objetos de calibração, k o número de fatores empregados e 1 é o grau de liberdade perdido pela centralização dos dados na média das colunas.

Contudo o RMSEC por ser uma função que depende do número de fatores, de modo que um aumento no valor de k (ver **Equação 20**) pode levar a valores de erro enganosos. Este efeito é conhecido como sobrejuste do modelo, portanto é adequado usar métricas como raiz quadrada do erro médio quadrático de validação cruzada (Root Mean Squares Error of Cross Validation, RMSECV) e raiz quadrada do erro médio quadrático de predição (Root Mean Squares Error of Prediction, RMSEP) para se ter uma idéia mais realista da qualidade do modelo [8].

No calculo do RMSECV e RMSEP o número de fatores não é levado em consideração como mostrado na **Equação 21**.

$$RMSECV \text{ ou } RMSEP = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{m}} \quad (21)$$

Em geral, para modelos que não apresentam sobreajuste, o valor de RMSECV é ligeiramente maior que o valor o valor de RMSEC, uma vez que no processo de validação cruzada à remoção dos pontos extremos acarreta extrapolação na predição da amostra removida. Já para o caso de um conjunto de amostras externas, escolhidas criteriosamente empregando algoritmo de seleção de amostra como Kernard-Stone [22], partição de amostras empregando as matrizes **X** e **Y** (samples partition using the matrix **X** and **Y**, SPXY) [23], Duplex [24] ou mesmo planejamento experimental [25] de modo a não extrapolar a região calibrada, o RMSEP deve ser menor que o RMSEC para modelos inversos, embora isto nem sempre ocorra [8].

Contudo, como mencionado acima o RMSECV ou RMSEP contém uma fração que é devida ao erro sistemático e é expresso pela razão do somatório dos desvios dos valores real e predito pelo número de objetos de calibração ou predição (**Equação 22**) [31].

$$bias = \frac{\sum(y_i - \hat{y}_i)}{m} \quad (22)$$

Em uma situação ideal o somatório no numerador da **Equação 22** deve ser zero, mas como os valores preditos também são afetados por flutuações aleatórias este valor deve está próximo de zero. A norma E1655-00 da ASTM [26] sugere a investigação desse parâmetro por meio de um teste-*t* para as amostras de validação externas no nível de 95% de confiança. Calcula-se o desvio padrão dos erros de validação (Standard Deviation of Validation, SDV) empregando a **Equação 23**. O SVD nada mais é que o RMSE diminuído da fração sistemática do erro.

$$SDV = \sqrt{\frac{\sum[(y_i - \hat{y}_i) - bias]^2}{(m_v - 1)}} \quad (23)$$

Finalmente, o valor de t é calculado de acordo com a **Equação 24**, e comparado ao valor de t crítico para os devidos graus de liberdade. Caso o valor de t seja maior que t crítico é um indicativo que o erro sistemático cometido pelo modelo é significativo.

$$t_{bias} = \frac{bias \sqrt{m_v}}{SDV} \quad (24)$$

Outra forma complementar de avaliar a qualidade de um modelo consiste em ajustar por meio do método OLS uma reta entre o valor previsto e o valor real ou de referência (**Figura 2.4**). Com isso, podem-se estimar alguns parâmetros como coeficiente de correlação (r), coeficiente de determinação (R^2) coeficiente angular (*slope*) e o intercepto ou coeficiente linear (*offset*) [27].

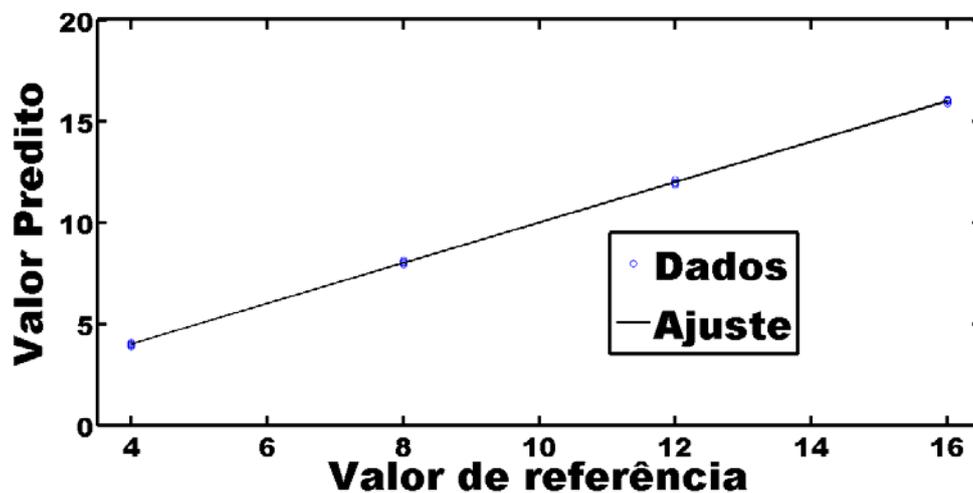


Figura 2.4- Exemplo de ajuste por OLS entre valor de referencia e valor predito.

Idealmente, o valor predito pelo modelo deve ser igual ao valor de referência, neste caso o *coeficiente angular* dever ser igual a um e o intercepto igual à zero. Todavia, flutuações aleatórias distorcem estes valores, mas os seus respectivos intervalos de confiança devem

conter o um e o zero. Esta abordagem é considerada mais adequada para avaliar a precisão de um modelo com relação ao RMSECV ou RMSEP.

Outro aspecto importante da qualidade de um modelo de calibração é o ajuste do mesmo. Embora não seja uma regra, mas segundo o princípio da parcimônia [28], um modelo deve conter apenas o necessário para predição e nada mais. Por exemplo, se dois preditores são suficientes para predizer satisfatoriamente \mathbf{y} , apenas dois preditores devem ser usados. Em outra situação, se a correlação entre \mathbf{x} e \mathbf{y} é descrita por um modelo linear, o uso de uma relação quadrática viola a parcimônia do problema. Portanto, usar mais termos que o necessário ou procedimentos de modelagem desnecessários leva a modelos sobreajustados, violando o princípio da parcimônia [8].

A problemática associada ao sobreajuste é que a adição de preditores irrelevantes pode acarretar piores predições para um conjunto externo de amostras, uma vez que os coeficientes de regressão ajustados a estes preditores estão descrevendo em sua maior parte flutuações aleatórias. Em outras palavras, o sobreajuste compromete a capacidade de generalização de um modelo [28].

Em métodos que fazem uso de compressão de dados (como no caso do PLS), a quantidade de fatores utilizada é um critério que pode levar ao sobreajuste do modelo. O uso de um número excessivo de fatores certamente incluirá informação desnecessária e redundante ao modelo acarretando o sobreajuste. A situação contrária, ou seja, o uso de um número reduzido de fatores gera modelos subajustados, entretanto tal situação é facilmente detectada, pois predições muito pobres são obtidas, já o caso contrário não [36].

Segundo Martens e Naes [8], o sobreajuste de um modelo depende fortemente do número de amostras usadas, à medida que o

número de amostras de calibração aumenta menos significativos são os efeitos de um possível sobreajuste.

2.2 Seleção de Variáveis em Calibração Multivariada

É muito comum na literatura propostas de metodologias para determinação de substâncias e/ou parâmetros físico-químicos em matrizes complexas empregando técnicas espectroscópicas, em especial espectroscopia vibracional (infravermelho, infravermelho próximo e RAMAN), e quimiometria [29-38].

O avanço da eletrônica e da informática permite o desenvolvimento de equipamentos como espectrômetros capazes de gerar uma grande quantidade de informação em um curto intervalo de tempo por amostra. Contudo, nem toda a informação é útil na hora de se construir um modelo de calibração que relacione o sinal analítico medido com o parâmetro de interesse.

As técnicas de seleção de variáveis têm como objetivo encontrar um subconjunto de preditores capaz de melhorar os resultados, ou em último caso mantê-los constante em termos de erro. Os métodos de seleção de variáveis buscam ainda produzir modelos mais simples ou parcimoniosos. A busca por esse subconjunto de variáveis consiste de um problema de otimização combinatorial guiado por uma função objetivo, geralmente o erro de validação cruzada ou o erro para um conjunto externo de amostras. As restrições impostas às combinações e as funções de custo definem a estratégia do algoritmo de seleção. Apesar de diversas propostas de algoritmos de seleção de variáveis terem sido reportadas na literatura [39-40], ainda é um tópico em discussão em quimiometria e áreas afins.

2.3 Seleção de Variáveis em Regressão PLS

Em regressão MLR a principal justificativa para se fazer seleção de variáveis, em especial em dados espectrais, é a forte multicolinearidade entre os preditores, que pode impossibilitar a inversão da matriz de covariância, como descrito na **Seção 2.1.4**.

Esse problema não ocorre em modelos como PCR e PLS, uma vez que as novas variáveis são ortogonais como no caso do PCR ou apresentam uma leve correlação como no caso do PLS [12]. Por muito tempo, acreditou-se ser desnecessário implementar procedimentos de seleção de variáveis acoplados a modelos de compressão de dados, inclusive acreditava-se que o PLS era insensível ao ruído [41].

No entanto, os benefícios da seleção de variáveis vão além de obter um subconjunto de variáveis minimamente correlacionados que possibilita o cálculo da matriz inversa de covariância. De fato, a remoção de variáveis sem correlação com o parâmetro de interesse, bem como produzir modelos mais simples e robustos, são exemplos de benefícios que favorecem qualquer tipo de modelo de regressão.

Spiegelman e colaboradores [42] publicaram, em 1988, o trabalho intitulado "*Theoretical Justification of Wavelength Selection in PLS Calibration: Development of a New Algorithm*", no qual se discute o efeito da inclusão de variáveis não informativas no modelo. A principal conclusão do estudo é que, independente da distribuição do ruído dos dados, a inclusão de variáveis com baixa correlação com o parâmetro de interesse leva a um aumento do *erro sistemático*.

Para modelos PLS, os métodos de seleção de variáveis podem ser categorizados em os que selecionam intervalos e que selecionam variáveis independentes. Segundo Höskuldsson, os escores obtidos de intervalos apresentam maior estabilidade na etapa de previsão quando comparado a escores obtidos de um conjunto de variáveis independentes [43].

A seguir são apresentadas as principais estratégias de seleção de variáveis associadas aos modelos de regressão PLS, tanto na forma de intervalos como na forma de preditores independentes.

2.3.1 Busca Exaustiva

O procedimento de busca exaustiva apresenta, como maior vantagem, a garantia de localizar o mínimo global da função de custo associado ao processo de seleção de variáveis, dado que todas as possíveis combinações são testadas. Entretanto, a essa busca pode compreender um tempo demasiadamente grande ^[44].

Para o caso de uma matriz com n variáveis em que se deseja saber qual a melhor combinação destas com k variáveis ($k=1$ até n), a quantidade de modelos de regressão que deve ser avaliado é dado por:

$$\sum_{k=1}^m \frac{n!}{[k!(n-k)!]} \quad (25)$$

Em uma situação em que **Xcal** tem dimensões 15x25, a quantidade de modelos que devem ser avaliados para $k \leq 6$ é 26 milhões. No entanto, para $k \leq 10$ este número chega 25 bilhões de modelos inviabilizando, assim tal procedimento. Esse fato motiva a busca de novas estratégias que conduzam a solução do problema na direção do mínimo global com menos esforço computacional ^[44].

2.3.2 Algoritmo Genético

Proposto na década de 60 pelo pesquisador John H. Holland da Universidade de Michigan, a idéia central do Algoritmo Genético (Genetic Algorithm, GA) é simular matematicamente os mecanismo da "Teoria da Evolução das Espécies" de Charles R. Darwin para

otimizar sistemas complexos [45]. Os algoritmos genéticos, GAs, buscam reproduzir o mecanismo biológico da evolução explorando todas as suas vantagens. Apesar do potencial demonstrado desde o início, só na década de oitenta é que os GAs ganharam força com a popularização dos microcomputadores. Atualmente, o GA encontra aplicações nas mais diversas áreas da ciência e engenharia como, por exemplo, processamento de imagem, filtros para cancelamento de ruídos, robótica, e seleção de parâmetros de redes neurais [46].

Em química, o primeiro registro de aplicação do GA foi o trabalho de Lucasius e Kateman [47], no qual a técnica foi usada para selecionar comprimentos de onda na região do ultravioleta aplicado à determinação de nucleotídeos. Atualmente, encontram-se muitos trabalhos que aplicam o GA a dados químicos devido a vantagens como:

- ✓ Não requer informação *a priori* sobre o gradiente de resposta;
- ✓ A existência de descontinuidade não afeta o desempenho do algoritmo;
- ✓ Existem operadores que evitam que o algoritmo fique preso a mínimos locais;

Em problemas de seleção de variáveis, o AG apresenta basicamente cinco etapas. A primeira consiste em codificar as variáveis de modo que estas se assimilem a cromossomos biológicos. Neste contexto, o procedimento mais comum é usar codificação binária em que cada variável representa um gene e pode receber valor 0 ou 1, o que significa respectivamente não incluída e incluída no modelo. Na **Figura 2.5** é apresentado um esquema da codificação binária.

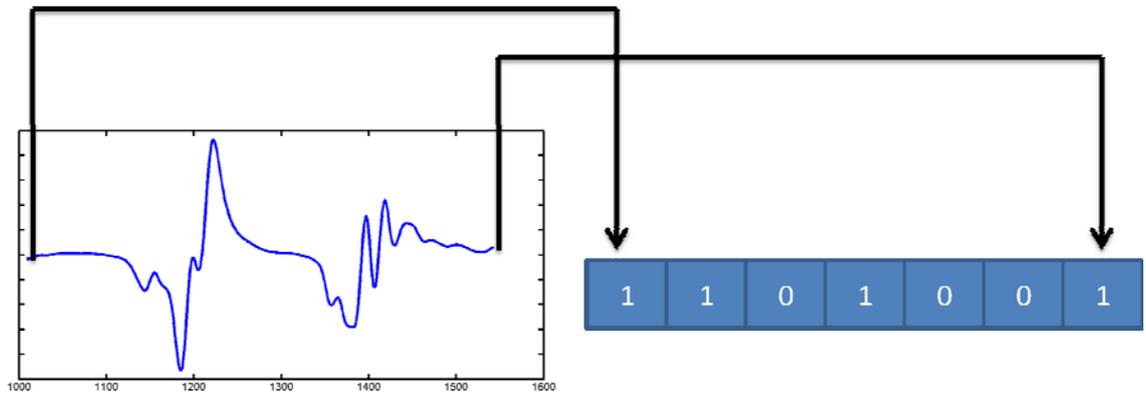


Figura 2.5- Codificação binária usada na seleção de variáveis do GA.

Como observado na **Figura 2.5** cada coluna em azul representa um indivíduo da população inicial, criada com base em um gerador randômico, para evitar influências tendenciosas. A quantidade de indivíduos na população inicial depende do problema, e quanto maior esse número maior o tempo de convergência do algoritmo.

A etapa seguinte consiste em avaliar a aptidão de cada indivíduo, o que na teoria da evolução corresponderia à capacidade de sobreviver. Matematicamente, representa a capacidade de gerar melhores respostas, sendo que, no contexto de seleção de variáveis, quando maior a aptidão menor o erro de predição obtido. A aptidão é obtida calculando um modelo de regressão para cada indivíduo e estimando se o RMSE para um conjunto de amostras externas ou por validação cruzada. Os que apresentarem boa aptidão são selecionados para próxima fase do algoritmo, a reprodução. O critério de escolha é conhecido como método da roleta ^[45], sendo que a probabilidade de cada indivíduo gerar descendentes é proporcional a sua aptidão em sobreviver.

A nova população é formada a partir do cruzamento aleatório de pares de cromossomos, gerando filhos que guardam a informação genética dos seus progenitores. Durante o processo de permuta do material genético a tendência natural é que, após certo número de

gerações, as características dominantes começam a prevalecer na população, levando assim a convergência do algoritmo. A reprodução pode ser por ruptura ou por recombinação. A **Figura 2.6** mostra de forma esquemática o processo de reprodução em uma AG.

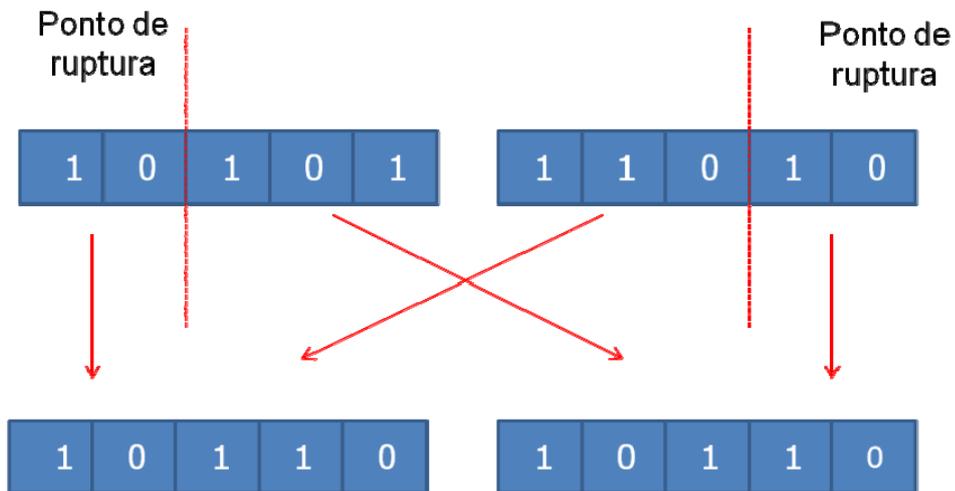


Figura 2.6 - Esquema de cruzamento no GA.

Em conjunto com a reprodução, ocorre também o processo de mutação, que é um operador que evita que o algoritmo fique preso a mínimos locais. Em termos práticos, a mutação consiste em substituir 0 ou 1 e vice e versa, como ilustrado na **Figura 2.7**.

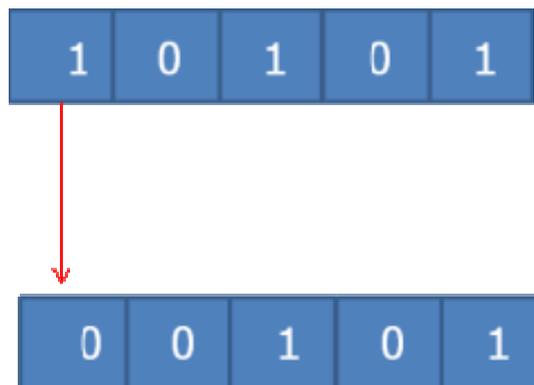


Figura 2.7- Esquema de processo de mutação no GA.

Ao final das n gerações escolhidas pelo usuário, o subconjunto de variáveis que produzir o melhor resultado com base no valor da função objetivo é a solução apresentada pelo GA.

Como apresentado acima, o GA já é conhecido para seleção de variáveis em calibração MLR. Em 1998, Leardi apresentou uma versão do GA associado a modelos de calibração PLS [48] no trabalho intitulado "*Genetic algorithms applied to feature selection in PLS regression: how and when to use them*". Neste trabalho, são discutidas as vantagens de se fazer seleção de variáveis individuais em oposição a intervalo. Segundo o autor, selecionar variáveis espalhadas pelo espectro representa um indício das regiões importantes do espectro para o modelo [48-49].

Como principais desvantagens dos métodos baseados em GA é o fato do mesmo ser estocástico e apresentar maior demanda computacional.

2.3.3 Método de eliminação de variáveis não informativas (UVE)

O método de eliminação de variáveis não informativas (Elimination of Uninformative Variables- UVE) [50], foi proposto por Center e co-autores no ano de 1996. O princípio básico do UVE é avaliar a estabilidade dos coeficientes de regressão PLS de forma comparativa com coeficientes obtidos de variáveis simuladas.

Inicialmente, um modelo PLS é computado e determina-se o número ótimo de fatores como sendo igual a k . A etapa seguinte consiste em gerar variáveis simuladas randomicamente, multiplicando-se por uma pequena constante. Isso produz a matriz $\mathbf{R}_{n \times p}$ com o número de variáveis p iguais ao número de variáveis em \mathbf{X} . A probabilidade, a priori, para cometer um erro na seleção, ou seja, para eliminar uma variável informativa ou de manter uma não-informativa é a mesma em \mathbf{X} e \mathbf{R} .

A matriz $\mathbf{X}_{(n \times p)}$ e a matriz $\mathbf{R}_{(n \times p)}$ são justapostas resultando na matriz $\mathbf{XR}_{(n \times 2p)}$, onde as primeiras p colunas sendo de \mathbf{X} e as p últimas sendo de \mathbf{R} . Um novo modelo *PLS* é calculado, mediante validação cruzada completa, o que resulta em n modelos e o número de fatores usando é o mesmo determinado para \mathbf{X} que é igual a k . Cada um dos n modelos obtidos terá com $2p$ coeficientes de regressão que representa o vetor \mathbf{b} .

A estabilidade de cada coeficiente é dada por:

$$c_j = \bar{b}_j / s(b_j) \quad (26)$$

Onde c_j é a estabilidade do coeficiente \mathbf{b}_j , \bar{b}_j é a média aritmética dos n b_j e $s(b_j)$ é o respectivo desvio padrão. A etapa seguinte consiste em determinar o maior valor absoluto de c_j para valores de j maiores que p , chamado de C_{max} . As variáveis em \mathbf{X} para qual o valor de j é $\leq p$ são mantidas se somente se c_j for maior que C_{max} . A **Figura 2.8** mostra uma interpretação gráfica do UVE.

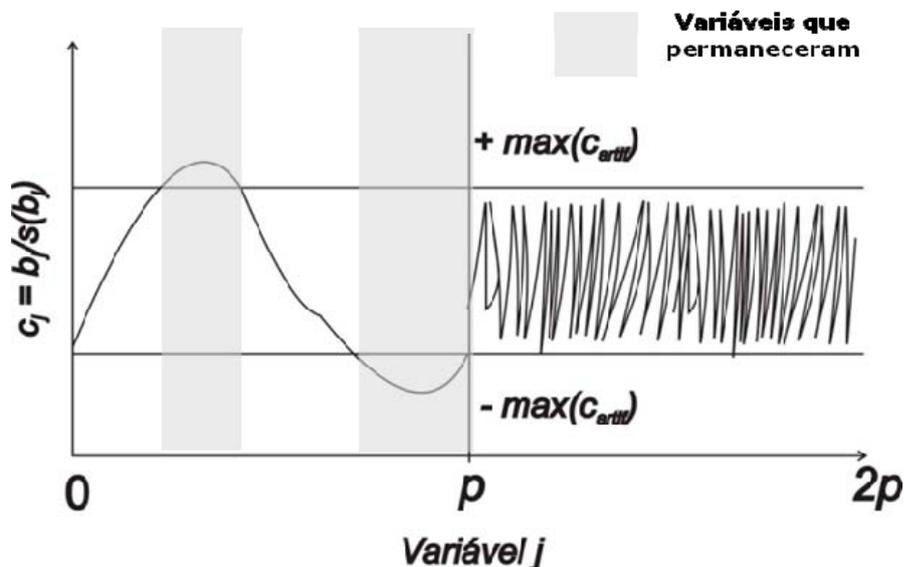


Figura 2.8 - Exemplo do processo de eliminação de variáveis.

Com as variáveis mantidas um novo modelo PLS é construído para prever as amostras externas e avalia-se o valor k de modo a evitar problemas de sobre ajuste do novo modelo.

2.3.4 Jack-Knife

O método de eliminação de variáveis Jack-knife, também é conhecido por teste de incerteza, foi proposto por Efrom [51] e adaptado por Martens e colaboradores [52] para o contexto de regressão bilinear.

Inicialmente, calcula-se um modelo PLS com validação cruzada completa e posteriormente estima a variância associado a cada elemento do vetor de coeficientes de regressão de acordo com a **Equação 27**.

$$S_j^2 = \sum_{j=1}^J ((\mathbf{b} - \mathbf{b}_j) \mathbf{g})^2 \quad (27)$$

onde g é um fator de escala definido como $g = \sqrt{\frac{n-1}{n}}$. Ao contrário de UVE que usa o \mathbf{b} médio dos n modelos PLS obtidos na validação cruzada, no método Jack-Knife o \mathbf{b} é o coeficiente para o modelo global com todos os objetos de calibração [52]. Depois de estimados os valores de S^2 , aplica-se um teste t à raiz quadrada de S^2 a 99% de confiança estatística. As variáveis que não passarem no teste, são ditas não informativas e passam a ter coeficientes de regressão igual à zero.

2.3.5 Colônia de formigas

Embora bastante conhecido em problemas de otimização, o Algoritmo de Colônia de Formigas (ACF) apenas recentemente foi introduzida em química no contexto de seleção de variáveis, mais

precisamente em regressão PLS [53]. As colônias de formigas representam uma estrutura organizacional capaz de realizar tarefas complexas, o ACF é um método estocástico resultante de observações reais de formigas na busca do melhor caminho até o alimento. No processo real as formigas depositam feromônio no chão formando trilhas do ninho até a fonte de comida, e que outras formigas podem usá-las para otimizar o caminho ninho-fonte de alimento em termos de menor distância.

No ACF são usadas formigas artificiais e que diferem das formigas reais em:

- ✓ As formigas artificiais têm memória;
- ✓ Não são completamente cegas;
- ✓ Elas vivem em um ambiente onde o tempo é discreto.

Por outro lado, diversas características são idênticas aos formigueiros reais:

- ✓ Formigas artificiais têm uma maior preferência probabilística por trilhas com maior quantidade de feromônio;
- ✓ Caminhos mais curtos tendem a terem maiores taxa de crescimento em termos de depósito de feromônio;
- ✓ As formigas artificiais usam um tipo de comunicação indireta, baseada na quantidade de feromônio depositada em uma trilha.

Basicamente, o algoritmo executa um laço contendo dois procedimentos básicos: Como o problema a ser resolvido deve ser estruturado e/ou modificado e a atualização das trilhas de feromônio.

Inicialmente todas as variáveis recebem a mesma quantidade de feromônio, o que representa inicialmente a mesma probabilidade de ocorrência (ou seja, $p_j=1$). O vetor \mathbf{t} (vetor feromônio) com dimensões $1 \times j$, onde j é a quantidade de variáveis em \mathbf{Xcal} , recebe

em todos os elementos o valor 1. A probabilidade de uma variável ser selecionada será dada por:

$$P_j = \frac{t_j}{\sum_{i=1}^j t_j} \quad (27)$$

A quantidade de feromônio em cada preditor é variada através de procedimento iterativo, quanto maior a quantidade de feromônio depositado em uma variável maior as chances de ser selecionada por uma formiga. Então a cada iteração do algoritmo a probabilidade cumulativa de cada variável é armazenada no vetor **cp** para uma colônia com Φ formigas, é calculada pela **Equação 28**:

$$cp_j = \sum_{i=1}^j P \quad (28)$$

O número de variáveis selecionadas por cada formiga é a solução do problema proposta por esta, e sua aptidão é avaliada por uma função custo ou objetivo (RMSECV). Após o procedimento de avaliação a melhor formiga deposita uma quantidade de feromônio sobre suas variáveis, que são possíveis candidatas a ser a resposta final do processo de otimização. O algoritmo é encerrado após um número pré definido de ciclos [53].

Uma versão mais recente de ACF foi proposta por Allegrini e Olivieri [54] usando na validação o procedimento Monte Carlos e paralelização dos cálculos de regressão.

2.3.6 PLS em intervalos - iPLS

A idéia do PLS por intervalos consiste em encontrar uma região do espectro (faixa) que produza melhores resultados que o espectro completo. Para atingir este objetivo, o espectro é dividido em iguais intervalos, a quantidade de intervalos é definida pelo usuário. É

importante ressaltar que esse parâmetro influencia no resultado final, e dividir os dados em um número pequeno de intervalos podem fazer com que uma faixa com potencial para gerar um bom modelo seja “contaminada” por informação desnecessária. Em caso contrário, dividir os dados em intervalos muito estreitos pode levar a faixas pobres em informações [55].

Após a divisão dos dados em intervalos o Algoritmo iPLS calcula um modelo global com validação cruzada, para auxiliar o usuário a decidir quantos fatores são necessários para descrever o sistema. Posteriormente, para cada intervalo é calculado um modelo PLS por validação cruzada e o que produzir menor RMSECV é o intervalo selecionado. A principal desvantagem da técnica é que se dois intervalos não sequenciados for à solução que leva ao mínimo global, esta nunca será obtida.

2.3.8 Backward PLS

Métodos de seleção *backward* são metodologias iterativas em que, ao contrário dos métodos *forward*, a cada iteração uma variável ou conjunto de variável é eliminada do modelo de regressão. Em 2009, Pierna e co-autores propuseram um método *backward* acoplado a modelos PLS [56], denominado *backward* seleção de variáveis PLS (Backward Variable Selection for PLS, BVSPLS).

Nessa abordagem, inicialmente um modelo PLS com validação cruzada no domínio do espaço coluna é efetuado. Para prevenir problemas de sobreajuste, o algoritmo emprega um conjunto de validação externa (denominado conjunto de parada “stop set”). O modelo associado ao menor valor de RMSEP é selecionado, isto significa que a variável excluída neste modelo é prejudicial e de fato dever permanecer fora do modelo. O procedimento descrito acima é repetido enquanto houver modelos capazes de minimizar o RMSEP,

na iteração em que ocorre aumento do valor de RMSEP o algoritmo é encerrado. Na **Figura 2.9** é apresentado um fluxograma do BVSPLS.

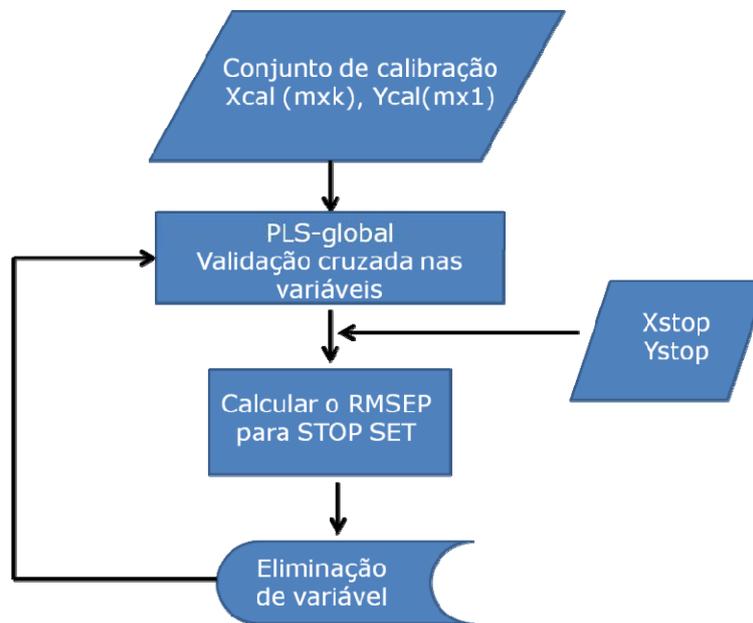


Figura 2.9 – Esquema do algoritmo BVSPLS. Adaptado [66].

Uma versão backward PLS por intervalos (biPLS) foi proposta por Leardi e Norgaard [57]. A idéia básica é parecida com o procedimento descrito na referência [56], contudo intervalos são usados no lugar de variáveis individuais.

2.3.9 PLS em intervalos sinérgicos - siPLS

Como mencionado na **Seção 2.5.6**, o *iPLS* fornece como solução um único intervalo, o que pode não conduzir ao ótimo desejado. O *siPLS* é uma variação do *iPLS* que combina intervalos. O usuário define em quantos intervalos deseja particionar a matriz **Xcal** e o algoritmo *siPLS* faz combinações de dois, três ou quatro intervalos [58].

O *siPLS* oferece a vantagem de poder selecionar mais de um intervalo, contudo esta característica também se torna uma

desvantagem. Uma vez que se o usuário particiona os dados em vinte intervalos, e somente um dos vinte intervalos contém informação relacionada com parâmetro de interesse o algoritmo, ainda assim, selecionará mais de um intervalo.

Outra característica do siPLS é que dentro dos limites predefinidos ele se torna um tipo de busca exaustiva. À medida que o número de intervalos e a quantidade de combinações aumentam, aproxima-se cada vez mais de uma busca exaustiva com variáveis individuais. Isso levará a problemas de demanda computacional, já mencionados na **Seção 2.5.1** e que desmotivam o uso de tais procedimentos.

2.3.10 OPS-PLS

O método OPS (do inglês "ordered predictors selection") foi proposto por Teófilo e colaboradores no ano de 2008 ^[59]. A essência do OPS consiste em ordenar os preditores (colunas de X) em função de sua relação com o parâmetro a ser calibrado, os seguintes critérios são usados para ordenar as variáveis:

- ✓ Vetor dos coeficientes de regressão;
- ✓ Correlação com \mathbf{y} ;
- ✓ Resíduos de validação cruzada;
- ✓ Influência de variáveis na projeção (IVP);
- ✓ Sinal analítico líquido (NAS)
- ✓ Covariância
- ✓ Relação sinal ruído.

Um modelo PLS inicial é construído com os n primeiros preditores, e seqüencialmente um incremento de preditores é adicionado até um número limite. Os modelos construídos são avaliados com base no valor de RMSECV, e os preditores que

gerarem o melhor modelo correspondem às variáveis selecionadas pelo algoritmo. Especificamente, o algoritmo OPS-PLS consiste das seguintes etapas:

- ✓ Obtenção de vetores informativos e suas combinações de \mathbf{X} e \mathbf{Y} ;
- ✓ Construção dos modelos PLS;
- ✓ Cálculos dos parâmetros de qualidade do modelo por validação cruzada;
- ✓ Comparação entre os modelos e escolha do modelo ótimo.

Na **Figura 2.10** é apresentado um esquema de funcionamento do OPS-PLS.

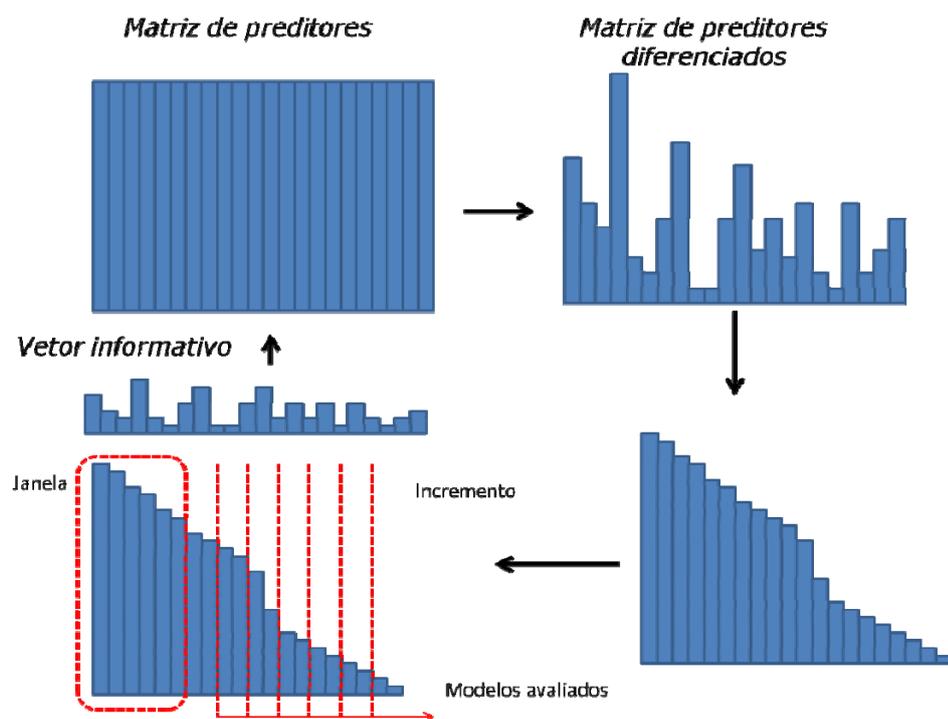


Figura 2.10 – Esquema do algoritmo OPS. Adaptado de [59].

2.3.11 Busca de Tabu

Busca de Tabu (TB) é uma técnica de otimização relativamente nova para o contexto de seleção de variáveis. O algoritmo TB é um método iterativo de otimização determinística global. Ele examina o

espaço de busca de uma forma altamente ordenada usando memória para manter o controle de partes já visitadas [60]. Esse algoritmo objetiva otimizar uma função do tipo $G(x)$ procurando no domínio de x uma solução ótima para G . As etapas básicas de BT são mostradas a seguir:

- ✓ **Inicialização:** uma solução s para G é escolhida aleatoriamente. Esta solução é avaliada pela função e armazenada na memória do algoritmo, que compõe a lista de tabu.
- ✓ **Exploração da vizinhança:** Todas as possíveis soluções envolvendo vizinhos de s são avaliadas. Soluções vizinhas são as que podem ser alcançadas a partir da atual por uma transformação simples e básica da solução atual.
- ✓ **Atualização:** Uma nova solução é escolhida com base nos vizinhos avaliados. Esta nova solução não pode estar na lista de tabu e ter valor de G menor que os demais vizinhos. Contudo, o algoritmo permite que a solução a ser adotada como nova seja inferior a atual. Este procedimento evita que o BT fique preso em mínimos locais, mas esta nova solução é adicionada à lista de tabu.
- ✓ **Parada:** Se todos os vizinhos estão na lista de tabu ou um número limite de iterações for atingido o algoritmo pára.

Com a solução final é computado o modelo PLS definitivo empregado para estimar a propriedade de interesse nas amostras externas.

2.3.11 Ponderação Iterativa dos Preditores

Ponderação iterativa de preditores ("ITERATIVE PREDICTOR WEIGHTING -IPW") [61] é um método de eliminação de variáveis não

informativas similar ao UVE [50] e ao Jack-knife [52], proposto por Forina e colaboradores. Este método é baseado em um ciclo de modelos PLS. Em cada ciclo, calcula-se a importância do preditor, definida como o produto entre o valor absoluto do coeficiente de regressão e o seu respectivo desvio padrão. No ciclo seguinte os preditores são ponderados por suas importâncias. Após 10 ou 20 ciclos, o algoritmo converge levando a um pequeno número de variáveis no modelo PLS final.

2.4 Algoritmo das Projeções Sucessivas

O Algoritmo das Projeções Sucessivas foi proposto, em 2001, por Araújo e colaboradores [62] no contexto de regressão linear múltipla e aplicado a dados espectroscópicos. O SPA é uma técnica do tipo *forward* com a restrição de que a variável incorporada em cada iteração deve ser a menos multicolinear possível com as variáveis previamente selecionadas. Em versões mais recentes [76], como disponível em www.ele.ita.br/~kawakami/spa/ o SPA é composto por três fases.

Na primeira fase são geradas as cadeias de variáveis minimamente redundantes, empregando somente a matriz \mathbf{Xcal} , geralmente centrada na média das colunas.

FASE-1: Geração das cadeias

Dada a matriz de calibração $\mathbf{Xcal}_{(n \times k)}$ os seguintes passos são seguidos:

Passo 1: (inicialização) faça:

De $k=1$ ate k

$$z^1 = x_k$$

$$x_{j=1}^1 = x_j, j=1, \dots, K$$

$$\mathbf{SEL}(1,k) = k$$

$$i = 1$$

Passo 2: Calcular a matriz de projeção P^i no subespaço ortogonal a z^i :

$$P^i = I - \frac{z^i(z^i)'}{(z^i)'z^i}$$

onde I é a matriz identidade de dimensões $n \times n$.

Passo 3: Calcular os vetores projetados x_j^{i+1} a partir de:

$$x_j^{i+1} = P^i x_j^i$$

para todos os $j = 1, \dots, K$.

Passo 4: Determinar o índice de j^* do vetor de maior projeção e armazená-lo na matriz **SEL**:

$$j^* = \arg(\max \|x_j^{i+1}\|) \text{ e } \mathbf{SEL}(i+1, k) = j^*.$$

Passo 5: Fazer $z^{i+1} = x_{j^*}^{i+1}$ (vetor que define a próxima operação de projeção)

Passo 6: Fazer $i = i + 1$. Se $i < M$, retorne para o **Passo 2**.

FASE-2: Avaliação das cadeias

A etapa seguinte (fase 2 do APS), consiste em avaliar a correlação das cadeias com o parâmetro de interesse. De forma esquemática o seguinte procedimento é usado:

Dados os valores de m_{\min} e m_{\max} (número máximo e o mínimo de variáveis a selecionar);

De $k = 1$ até K faça

De $m = m_{\min}$ até m_{\max} faça

- ✓ Use as variáveis com índices **SEL**(1,k), **SEL**(2,k), ..., **SEL**(m,k) para construir um modelo MLR. Aplique o modelo para o conjunto de validação e calcule o RMSE(m,k)

Próximo m

Próximo k

A obtenção do RMSE pode ser realizada de duas maneiras, utilizando validação cruzada ou um conjunto de teste independente.

A terceira e última fase, proposta por Galvão e colaboradores [76], consiste em eliminar as variáveis que não apresentam melhoria em termos de valor PRESS (*Predicted Residual Error Sum of Squares*), com base em um teste F. Para isso, a cada variável é associado um "fator de relevância" dado pelo produto do desvio padrão amostral e módulo do coeficiente de regressão desta variável. Posteriormente, os modelos MLR são construídos incluindo progressivamente as variáveis em ordem decrescente de importância e a cada nova variável adicionada calcula-se o valor de PRESS para um conjunto de validação a cada variável incluída. O menor número de variáveis para qual o valor de PRESS não difere do mínimo global empregando um teste *F* a 75% de confiança é empregado no modelo MLR final. De modo a facilitar o entendimento das projeções, na **Figura 2.11** uma interpretação geométrica para o caso em que $n = 3$ e $k = 5$.

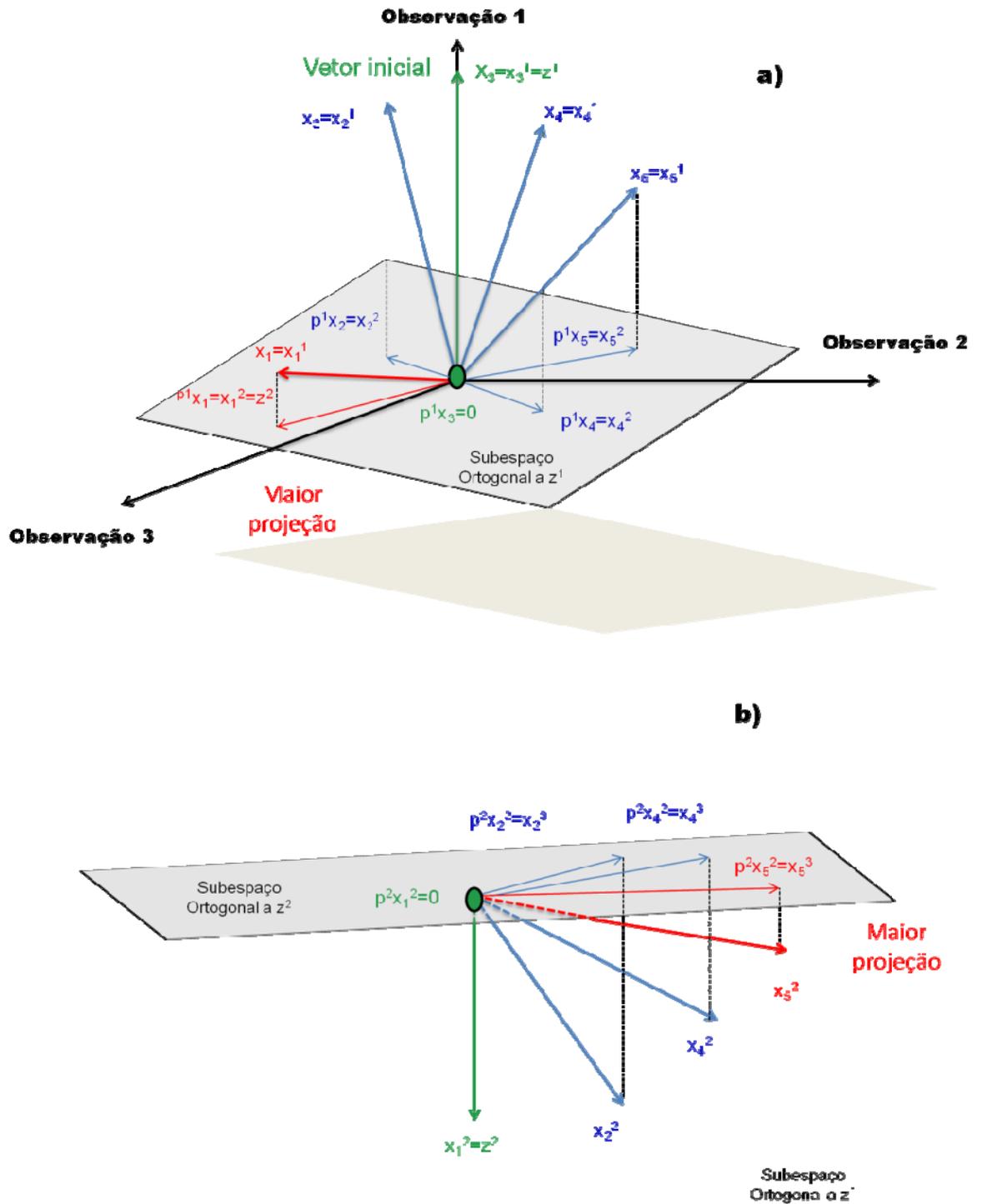


Figura 2.11- Ilustração da seqüência de projeções realizadas pelo SPA. (a): Primeira iteração. (b): Segunda iteração. Nesse exemplo, a cadeia de variáveis que inicia em x_3 deverá ser $\{x_3, x_1, x_5\}$ Adaptado de [62].

Diversas aplicações do SPA a dados espectroscópicos demonstram seu potencial. Determinação de enxofre em diesel [65], determinação de parâmetros de qualidade em óleos vegetais [66], quantificação de biodiesel em diesel [67] e determinação de parâmetros de qualidade de óleos isolantes [68] são alguns exemplos de aplicações do SPA empregando espectroscopia NIR. Outros trabalhos reportam o uso do APS com dados de espectroscopia UV-Vis na determinação simultânea de cátions bivalentes em polivitamínicos [69], quantificação de fenóis em água do mar [70], além de aplicações no contexto de estudos QSAR [71] e QSPR [72].

O bom desempenho do SPA motivou implementações do mesmo em outros contextos diferentes da calibração MLR. Dantas Filho e colaboradores aplicaram o SPA na seleção de um subconjunto de amostras representativa para construção de modelos de calibração [73]. Em 2005, Pontes e colaboradores adaptaram o SPA de modo que este fosse capaz de atuar como ferramenta de seleção de variáveis em problemas de classificação em conjunto com modelos de análise discriminante linear (LDA) [74]. Recentemente, Soares e colaboradores apresentaram uma modificação no SPA-MLR para que, no procedimento de seleção de variáveis, levasse em consideração a presença de interferentes desconhecido. Para atingir este objetivo a função de avaliação do algoritmo foi modificada, além de se avaliar o RMSE passou se a avaliar também o erro estatístico de predição [75].

Capítulo 3

Metodologia

3.0 METODOLOGIA

3.1 SPA em regressão PLS

Um aumento da capacidade preditiva de modelos PLS, oferecida pela seleção de variáveis (**Seção 2.3**), aliada ao bom desempenho do APS em diversas aplicações [66-74], motivaram o desenvolvimento de um algoritmo que incorpore as vantagens do SPA à regressão PLS. Para esse propósito, implementaram-se algumas modificações nas fases 1 e 2 do SPA, sendo que a fase 3 não foi utilizada.

O algoritmo proposto, denominado iSAP-PLS, inicia calculando um modelo PLS global empregando uma quantidade de fatores determinada pelo usuário, e a seguinte saída gráfica é apresentada (**Figura 3.1**):

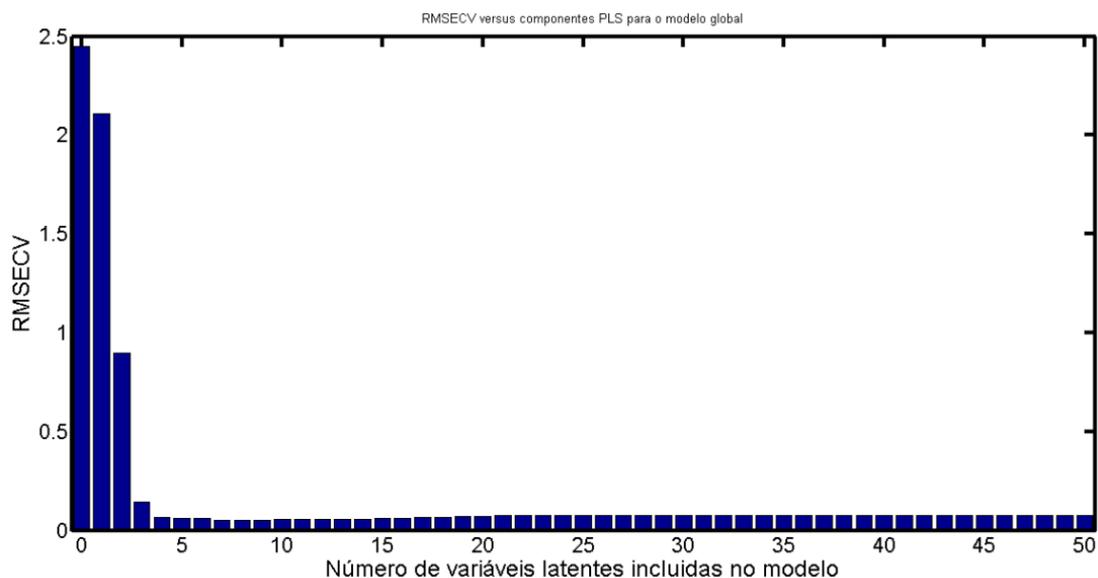


Figura 3.1- Saída gráfico do modelo global PLS calculado no início da execução do algoritmo iSPA-PLS.

Na **Figura 3.1**, são apresentados os valores de RMSECV para o PLS global em função do número de fatores incluídos no modelo. O objetivo dessa etapa é guiar a escolha do número de fatores que

serão usados na avaliação dos intervalos. Para tanto, o ponto de mínimo da curva RMSECV *versus* número de fatores é localizado e comparado ao seu antecessor com base em um teste *F* a 75 % de confiança estatística [76]. Se o RMSECV no ponto mínimo, para *k* fatores, for menor que o RMSECV para *k-1* de acordo com o teste estatístico aplicado *k* fatores, é sugerido como sendo a dimensionalidade inerente ou posto da matriz **Xcal**.

No entanto, se o teste indicar que o RMSECV para *j* fatores não é significativamente menor que o RMSECV para *k-1* fatores, então um novo teste *F* é conduzido comparando o RMSECV para *k-1* com o RMSECV obtido para *k-2*. O procedimento continua até ser encontrada diferença significativa entre os RMSECV comparados. Com base no número de fatores sugeridos conforme o procedimento descrito acima e a saída gráfica do algoritmo (**Figura 3.1**), o usuário indica quantos fatores devem ser utilizados nos modelos iSPA-PLS.

✓ FASE 1

Inicialmente, a matriz de calibração **Xcal**_(m×n) é particionada em *j* matrizes de calibração denominadas de **Ical**_(m×j). Assim como no *iPLS* e suas variações, a quantidade de intervalos é um critério determinado pelo usuário. As matrizes **Ical** possuem igual quantidade de variáveis, contudo em situações em que a razão entre o número de colunas de **Xcal** e a quantidade de intervalos não for exata os primeiros intervalos receberão uma variável a mais.

Para cada intervalo, é escolhido uma variável “representante” do mesmo. Esta escolha é feita determinando a norma de cada variável do intervalo.

De *j=1* ate *j* faça

NOR(j)=norma(Ical^j);

Proximo j

NOR é uma matriz que contém as normas de cada variável dos j intervalos. Na etapa seguinte, é determinada a variável que apresenta a maior norma de cada intervalo, que são denominadas variáveis representante do respectivo intervalo. A matriz **iXcal**_(m×j) contendo apenas as variáveis representantes é empregada na etapa de projeção do SPA como descrito na seção 2.1. A **Figura 3.2** ilustra o procedimento descrito acima.

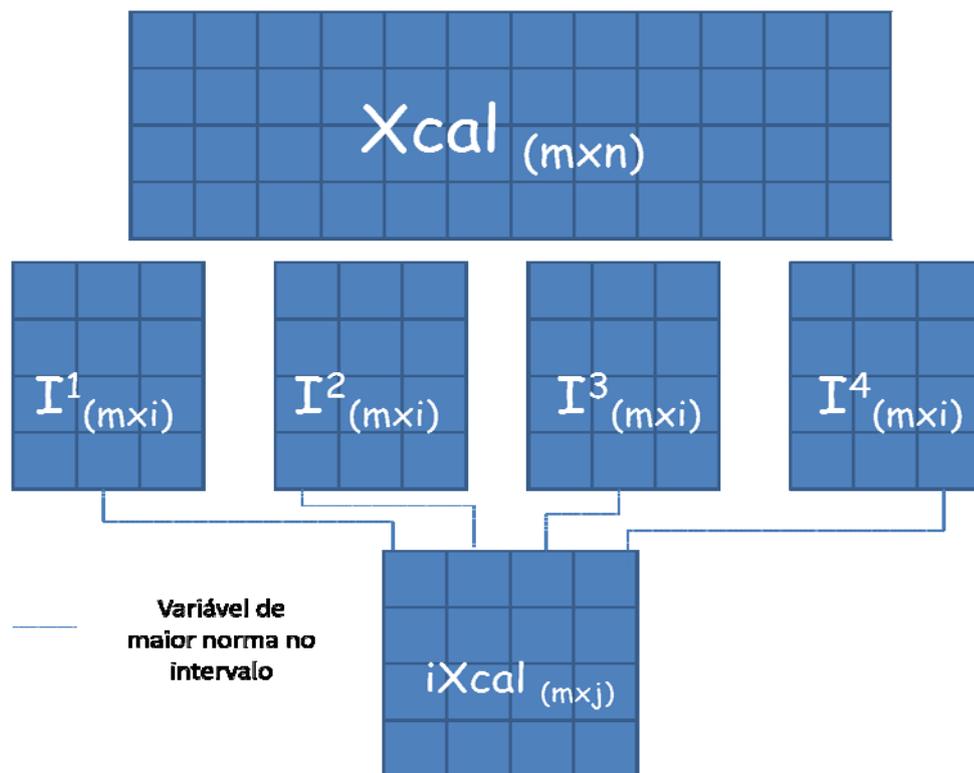


Figura 3.2- Ilustração da montagem da matriz **iXcal** usada na etapa de projeção do SPA.

✓ FASE 2

No final da etapa de projeções é gerada a matriz **SEL** contendo os índices das variáveis representantes que forma cada cadeia de intervalos. Na fase 2 do SPA-MLR após as projeções, é conduzida a avaliação de cada cadeia via regressão MLR. No algoritmo proposto temos:

Dado o valor de 1 de I_{max} (número máximo de intervalos a selecionar);

De $j = 1$ até j faça (j é a quantidade de intervalos)

De $m = 1$ até I_{max} faça

De 1 até K (k número de fatores sugeridos pelo usuário)

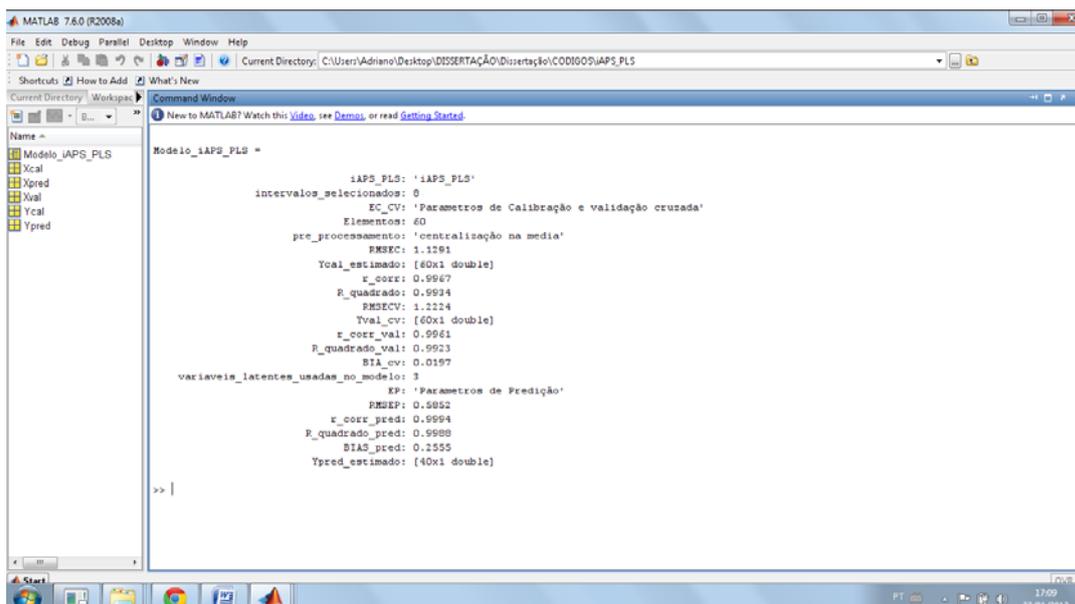
- Use as variáveis representantes com índices **SEL(1,k)**, **SEL(2,k)**, ..., **SEL(m,k)** para construir um modelo PLS empregando os respectivos intervalos, com validação cruzada completa.

Próximo K

Próximo I

Próximo j.

Ao contrario dos outros algoritmos que fazem seleção de intervalos, o iSPA-PLS possibilita selecionar entre 1 e $j-1$ intervalos. Após avaliar todas as cadeias de intervalos, via regressão PLS, o que apresentar menor RMSECV é selecionado, e o seguinte relatório é apresentado na janela de comando do Matlab (**Figura 2.4**).



```

Modelo_iAPS_PLS =
    iAPS_PLS: 'iAPS_PLS'
    intervalos_selecionados: 0
    EC_CV: 'Parametros de Calibração e validação cruzada'
    Elementos: 60
    pre_processamento: 'centralização na media'
    RMSEC: 1.1291
    Ycal_estimado: [60x1 double]
    r_corr: 0.9967
    R_quadrado: 0.9934
    RMSECV: 1.2224
    Yval_cv: [60x1 double]
    r_corr_val: 0.9961
    R_quadrado_val: 0.9923
    BIAS_cv: 0.0197
    variaveis_latentes_usadas_no_modelo: 3
    EP: 'Parametros de Predição'
    RMSEP: 0.5052
    r_corr_pred: 0.9994
    R_quadrado_pred: 0.9988
    BIAS_pred: 0.2555
    Ypred_estimado: [40x1 double]
>> ]
  
```

Figura 3.3- Ilustração do relatório de saída do algoritmo iSPA-PLS.

Para ter acesso a qualquer parâmetro do modelo, é simplesmente digitar o seguinte comando no Matlab [nome do modelo].[variável]. Por exemplo, Modelo_iSPA_PLS.RMSECV. Além do relatório apresentado acima, saídas gráficas também são geradas para complementar os resultados (**Figura 3.4**). Na **Figura 3.3a** é apresentado o gráfico de valor predito versus referência para o conjunto de predição externa, enquanto na **Figura 3.4b** é mostrado o conjunto de intervalos selecionados pelo algoritmo iSPA-PLS.

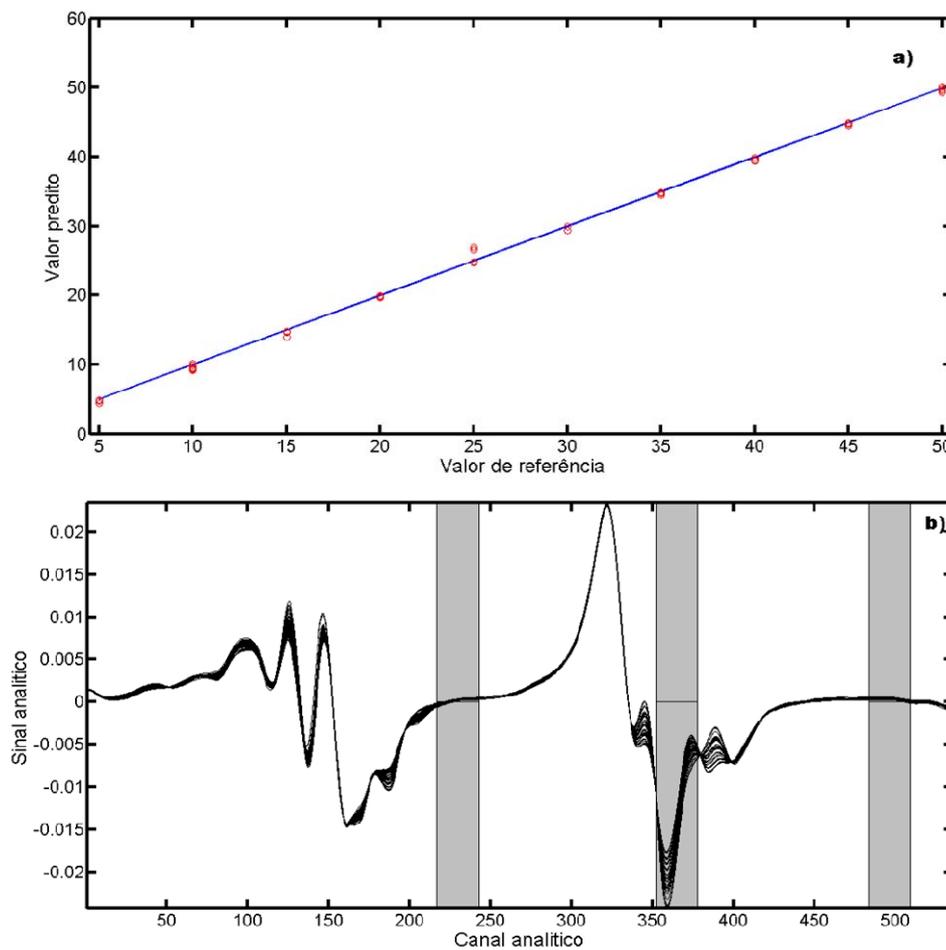


Figura 3.4- Saída gráfica do iSPA-PLS (a) valor predito versus referência, (b) Intervalos selecionados.

Com intuito de apresentar o algoritmo de forma resumida, um fluxograma é apresentado na **Figura 3.5**, mostrando o passo a passo das etapas executadas.

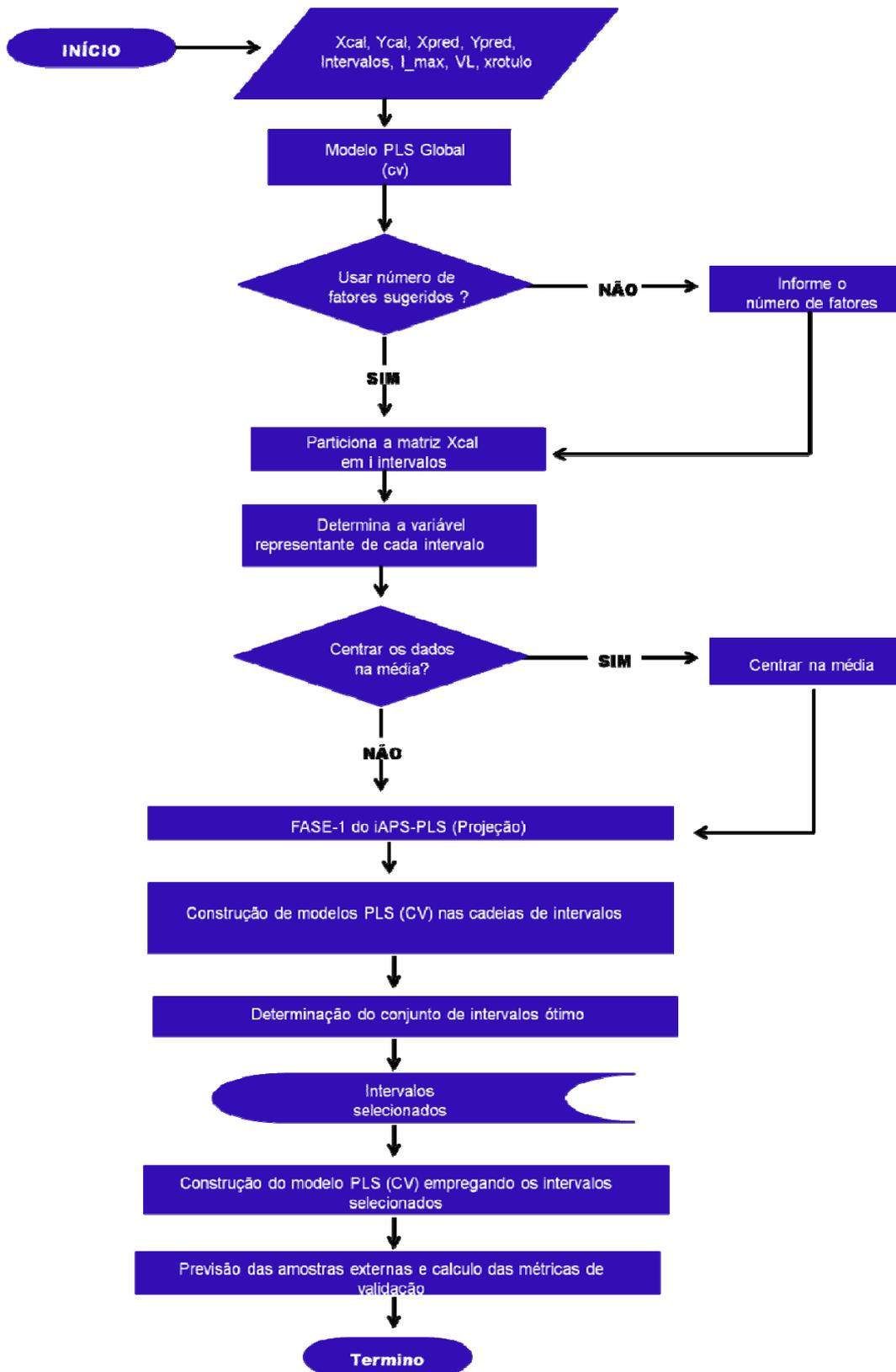


Figura 3.5: Fluxograma do algoritmo iSPA-PLS.

3.2 Algoritmos usados para comparação

3.2.1 GA-PLS

O Algoritmo GA-PLS empregado neste trabalho foi desenvolvido por Leardi e colaboradores e se encontra disponível em <http://www.models.kvl.dk/algorithms>. As configurações do GA estão apresentadas na **Tabela 3.1**.

Tabela 3.1: Configurações do GA ^[48].

Parâmetro do GA	Valor
Tamanho da População	30
Probabilidade de mutação	1%
Quantidade de Gerações	100
Função de avaliação	RMSECV

3.2.2 iPLS e siPLS

Os modelos iPLS e siPLS foram construídos empregando a interface gráfica mostrada na **Figura 3.6** e que se encontra disponível em http://www.models.kvl.dk/go?filename=ItoolsGUI1_01.zip.

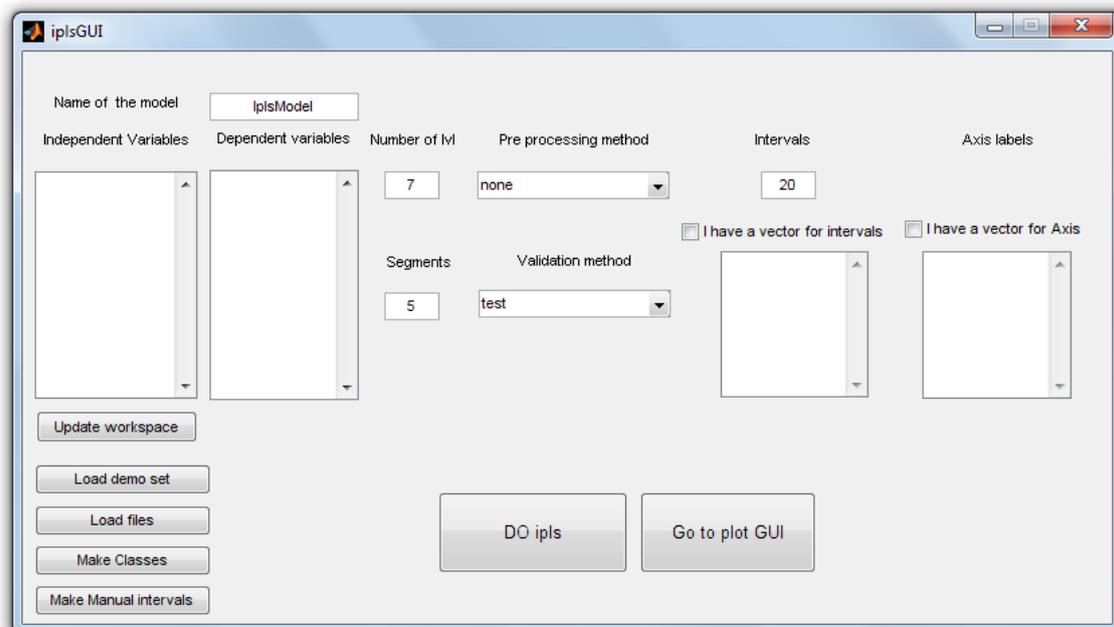


Figura 3.6: Interface gráfica do iToolBox.

3.2.3 Jack-Knife-PLS

Os modelos Jack-Knife-PLS foram construídos empregando o software The Unscrambler versão 9.7.

3.2.4 Comparação dos modelos

Os modelos obtidos foram comparados em termos de RMSEC, RMSECV, RMSEP, coeficiente de determinação da calibração e predição, além do número de fatores PLS empregados no modelo final e o número de variáveis selecionadas.

Os RMSEP foram comparados com base em um teste F a 95 % de confiança estatística. O valor de F calculado ^[25] (F_{cal}) foi obtido empregando a equação abaixo:

$$F_{cal} = \frac{(RMSEP_A)^2}{(RMSEP_B)^2} \quad (29)$$

Em que $RMSEP_A$ e $RMSEP_B$ são respectivamente o RMSEP de maior valor e do menor valor respectivamente. Os valores de F crítico foram obtidos empregando o *toolbox* estatística.

3.3 Estudos de caso

3.3.1 Determinações de corantes alimentícios

Este banco de dados foi construído por Veras e Colaboradores, e empregado na validação de um espectrômetro alternativo [ref].

Nesta seção, apresenta-se um estudo de caso envolvendo a determinação de três corantes alimentícios em amostras sintéticas usando a espectrometria de absorção molecular na região do visível. O objetivo do uso deste conjunto de dados é atestar e explorar as características do algoritmo proposto, uma vez que a matriz é simples e todas as fontes de sinal analítico são conhecidas.

Os três corantes escolhidos foram tartrazina (TRA), amarelo crepúsculo (AC) e vermelho quarenta (V40), cujos espectros de suas soluções puras contra um branco de referência são apresentados na **Figura 3.7**. São também mostradas suas respectivas fórmulas moleculares e estruturais.

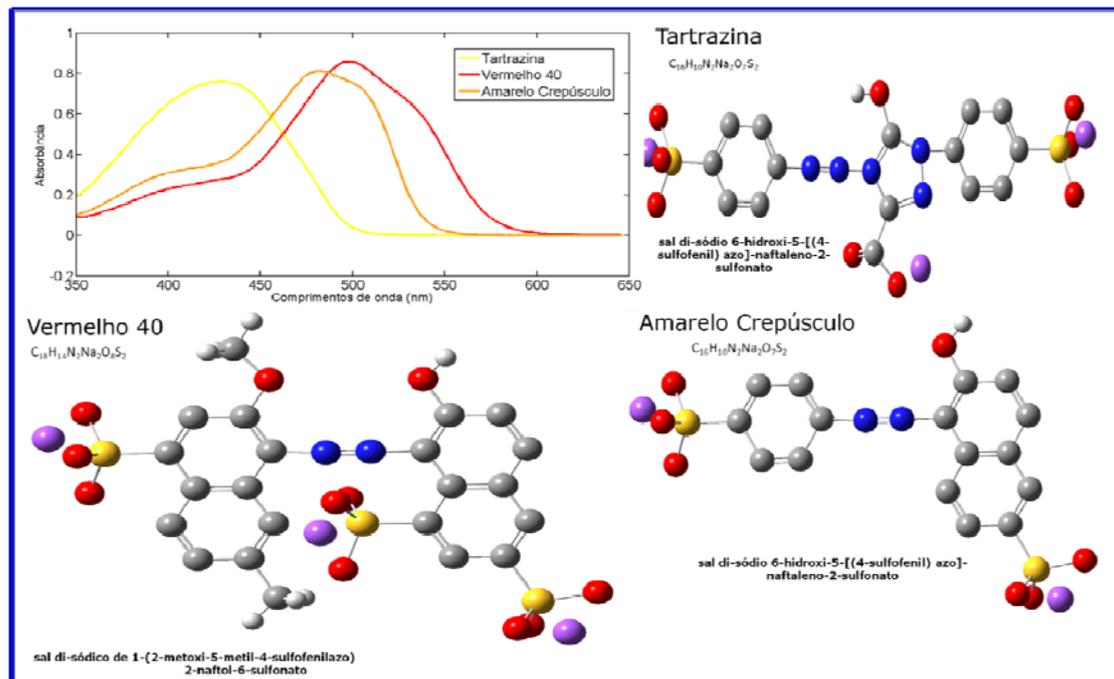


Figura 3.7. Espectros de absorção UV-VIS e estruturas moleculares dos corantes puros.

As misturas de calibração e validação externa foram preparadas empregando um Planejamento Brereton [12] com três fatores em três níveis. Nesse planejamento, o número de experimentos deve ser múltiplo do número de níveis (I) elevado ao quadrado, $N = k * I^2$, onde K é o número de repetições em uma dada mistura.

No presente trabalho, tem-se um total de 81 experimentos ou ensaios para três níveis em triplicatas. Na **Tabela 3.2** são apresentados os níveis escolhidos para cada corante.

Tabela 3.2- Concentração dos corantes do conjunto de calibração [mg L^{-1}].

Corante	Níveis		
	-1	0	1
Tartrazina	2	5	8
Vermelho 40	2	5	8
Amarelo Crepúsculo	2	5	8

Os valores -1, 0 e 1 representam os valores codificados dos níveis que são usados para gerar a matriz de planejamento com base no nível de repetição. Neste caso, escolheu-se o **0**, e em um ciclo gerados (**-1** **1**), com os quais é obtido a primeira coluna da matriz de planejamento e sucessivamente as demais. Na [Tabelas 3.3](#) é apresentada a matriz de planejamento codificado e com valores dos níveis para o conjunto de calibração.

Tabela 3.3- Matriz completa do planejamento Brereton com 03 níveis e 03 fatores das misturas para conjunto de calibração [mg L^{-1}].

Experimento	Fator 01/Tartrazina (nível/conc.)	Fator 02/Amarelo Crepúsculo (nível/conc.)	Fator 03/Vermelho 40 (nível/conc.)
01	0/5	0/5	0/5
02	0/5	-1/2	-1/2
03	-1/2	-1/2	1/8
04	-1/2	1/8	0/5
05	1/8	0/5	1/8
06	0/5	1/8	1/8
07	1/8	1/8	-1/2
08	1/8	-1/2	0/5
09	-1/2	0/5	-1/2

Baseado também em um planejamento Brereton, porém com ensaios sem repetições, um conjunto de misturas ditas de validação externa, composto por 27 amostras foram preparadas de modo a não extrapolar a faixa de concentração modelada pelo conjunto de calibração. A matriz de planejamento é apresentada na [Tabela 3.4](#).

Tabela 3.4- Matriz completa do planejamento Brereton com 03 níveis e 03 fatores das misturas para conjunto de validação externa [mg L^{-1}].

Experimento	Fator 01/Tratrazina (nível/conc.)	Fator 02/Amarelo Crepúsculo (nível/conc.)	Fator 03/Vermelho 40 (nível/conc.)
01	0/5.5	0/5.5	0/5.5
02	0/5.5	-1/4.0	-1/4.0
03	-1/4.0	-1/4.0	1/7.0
04	-1/4.0	1/7.0	0/5.5
05	1/7.0	0/5.5	1/7.0
06	0/5.5	1/7.0	1/7.0
07	1/7.0	1/7.0	-1/4.0
08	1/7.0	-1/4.0	0/5.5
09	-1/4.0	0/5.5	-1/4.0

Após a preparação das misturas dos corantes foi efetuada a aquisição dos espectros com um espectrômetro HP, modelo 8435, equipado com célula de quartzo, caminho óptico de 10 mm e detector de arranjo de fotodiodos com resolução de 1 nm na faixa de 430 a 646 nm.

A matriz de calibração tem dimensões de 81 amostras de medida em 217 comprimentos de onda ($\mathbf{Xcal}_{81 \times 217}$), e as dimensões da matriz de validação externa é de 27 amostras também medido em 217 comprimentos de onda ($\mathbf{Xval}_{27 \times 217}$). A faixa de concentração modelada na calibração está compreendida entre 2 a 8 mg/L, sendo a precisão de preparação das amostras de 0,1 mg/L.

3.3.2 Determinação do teor de proteínas em amostras de trigo

O objetivo deste estudo de caso é avaliar o desempenho do algoritmo iSPA-PLS frente a uma matriz complexa, na determinação do teor de proteína em amostras de trigo. O conjunto de dados (disponível em <http://www.idrc-chambersburg.org/shootout.html>) é composto de 107 espectros na faixa de 400 a 2498 nm com resolução de 2 nm. Contudo, neste trabalho apenas a faixa NIR foi empregada, ou seja, a região compreendida entre 1108 a 2498 nm, totalizando

680 variáveis. O teor de proteína nas amostras varia entre 9.70 e 14.40 %m/m.

O conjunto de amostras foi particionado em calibração e validação externa empregando o algoritmo SPXY, sendo 67 amostras usadas na calibração e 40 na validação externa.

3.3.3 Determinação da qualidade de amostras de extrato de cerveja

Neste estudo de caso é conduzida a determinação de parâmetro de qualidade em amostras de extrato de cerveja, empregando espectrometria VIS-NIR na região de 400 a 2250 nm. Os espectros foram registrados com resolução de 2 nm totalizando 926 variáveis.

Este conjunto de dados encontra-se disponível em <http://www.models.kvl.dk/algorithms>), sendo composto por quarenta amostras de calibração e vinte amostras de predição. A variável resposta (\mathbf{y}) é adimensional ^[43] e expressa uma medida da qualidade da cerveja que varia entre 4,2 a 18,8.

Capítulo 4

Determinações de corantes
alimentícios

4.0 DETERMINAÇÕES DE CORANTES ALIMENTÍCIOS

Na **Figura 4.1** são apresentados os espectros do conjunto de calibração. Devido a não existência de ruído excessivo ou mesmo deslocamentos sistemáticos de linha de base os espectros, os dados foram simplesmente centrados na média das colunas antes da realização dos cálculos de regressão.

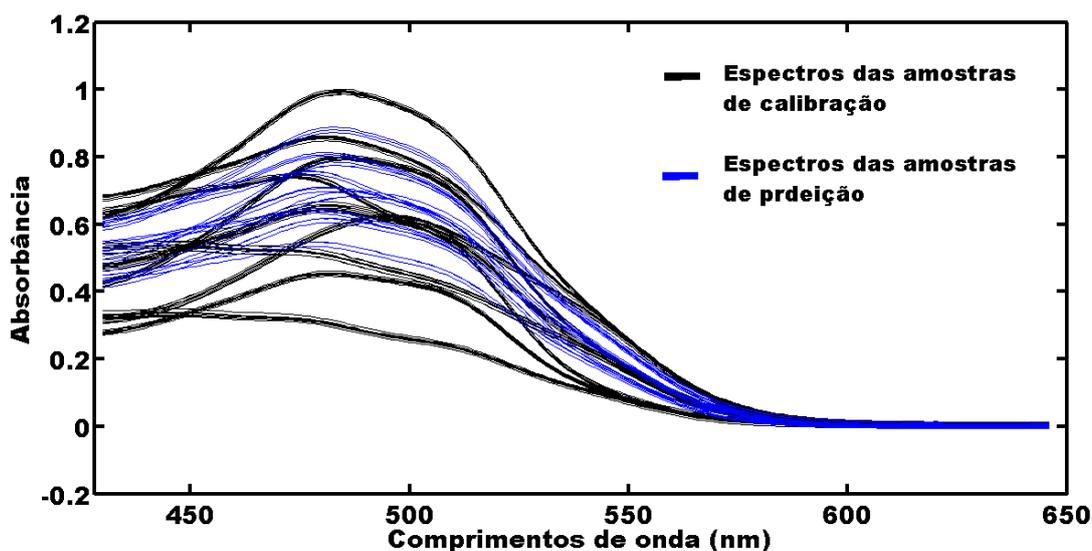


Figura 4.1- Espectros de absorção no visível das amostras das misturas de corantes.

A melhoria obtida pelo processo de seleção de variáveis será avaliada em termos de diminuição dos erros de validação cruzada, predição para um conjunto externo e parcimônia na seleção do número de fatores PLS quando comparado ao modelo global.

Inicialmente será apresentada uma investigação sobre o número de fatores ótimo para o modelo global. De acordo com Leardi^[48-49] o número máximo de fatores para uma modelo que emprega um subconjunto de variáveis não deve ser superior a número de fatores empregado para o modelo global. Na **Figura 4.2** é apresentado o gráfico de RMSECV em função do número de fatores

incluídos modelo no global para cada corante empregado neste estudo de caso.

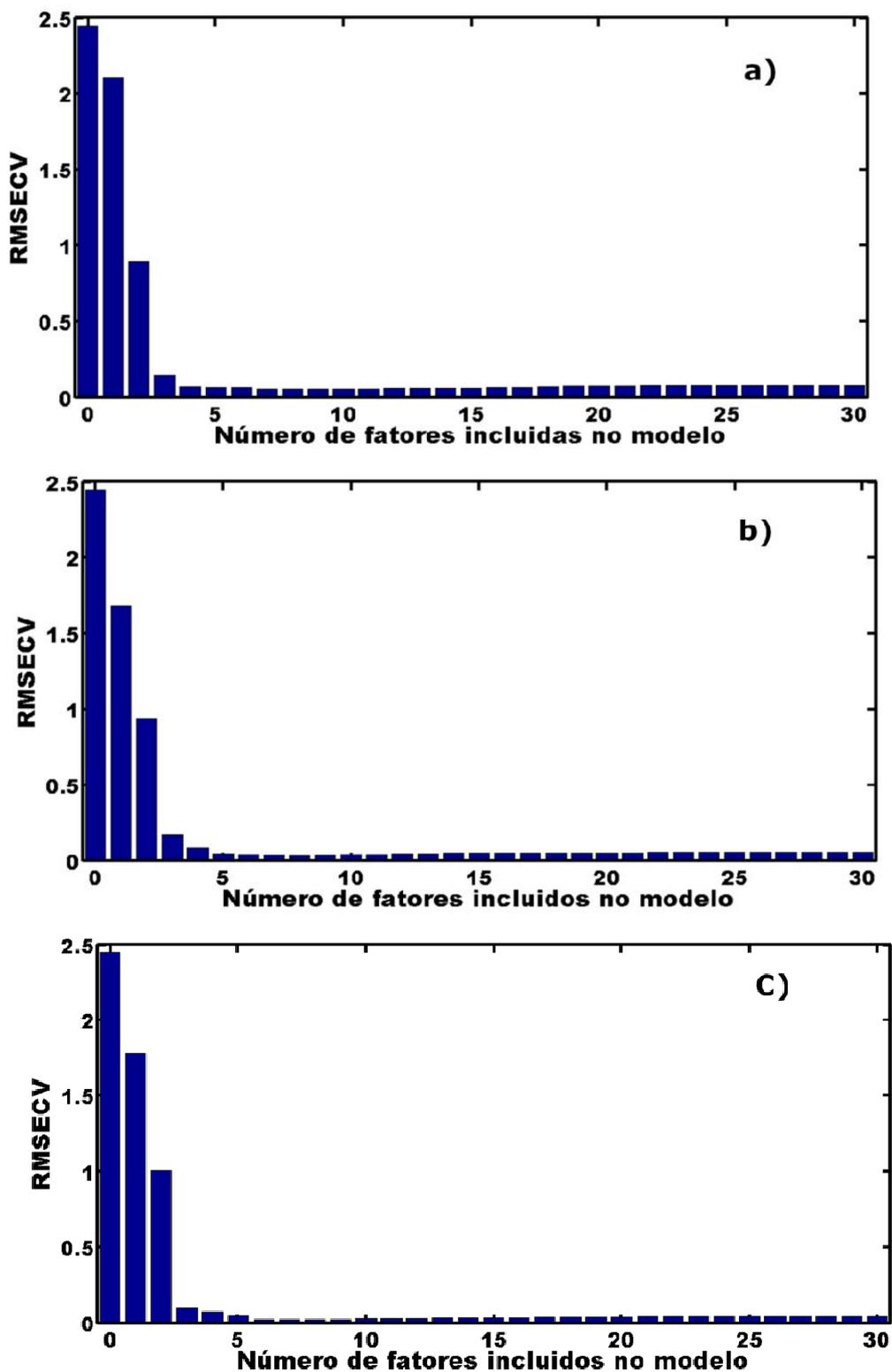


Figura 4.2- RMSECV versus número de fatores PLS incluídos no modelo global (a) amarelo crepúsculo, (b) vermelho 40 e (c) tartrazina.

Com base nos gráficos apresentados na **Figura 4.2**, é possível perceber que após o terceiro fator praticamente não ocorre variação significativa no valor do RMSECV. Este resultado sugere que três fatores são suficientes para descrever o sistema em estudo, este resultado é corroborado pelo gráfico de resíduos obtidos por validação cruzada, que são apresentados na **Figura 4.3**.

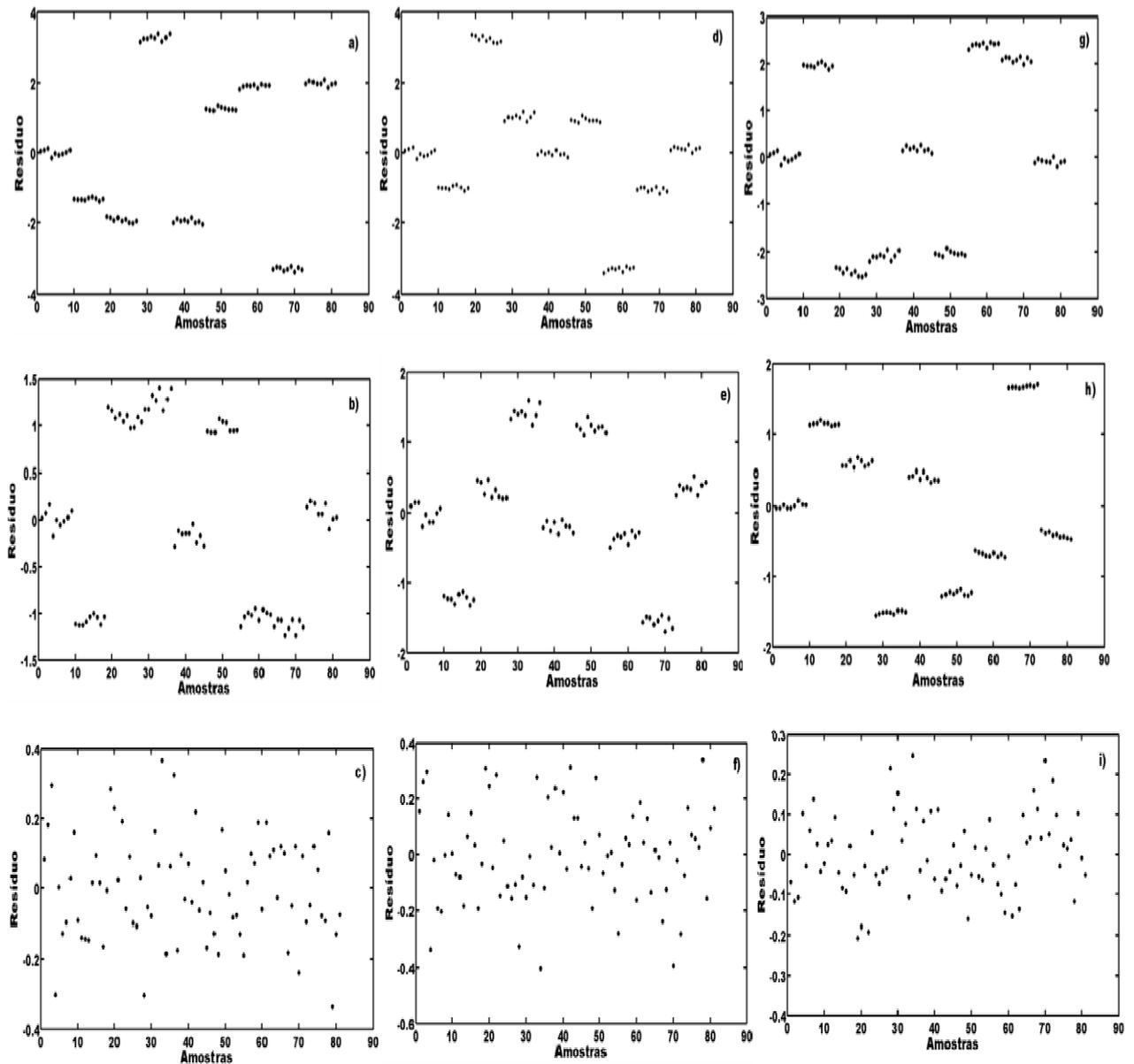


Figura 4.3- Resíduos de validação cruzada para (a) AC um fator, (b) AC 2 fatores, (c) AC 3 fatores, (d) V40 1 fator, (e) V40 2 fatores, (f) V40 3 fatores, (g) TAR 1 fator, (h) TAR 2 fatores e (i) TRA 3 fatores.

A **Figura 4.3** sugere que após o 3 fator na existe mais informação sistemática a ser modelada, permanecendo na matriz de resíduos apenas flutuação aleatórias.

4.1 Quantificação dos corantes

Nas **Tabelas 4.1, 4.2 e 4.3** são apresentados os resultados de calibração e de validação externa dos corantes tartrazina, amarelo crepúsculo e vermelho-40 respectivamente.

Os resultados, obtidos em termos dos erros (RMSEC, RMSECV e RMSEP), encontram-se em concordância com esperado, tendo em vista a simplicidade da matriz. Sendo o principal inconveniente do conjunto de dados a forte multicolineariedade entre as variáveis, o que é contornado pela decomposição dos dados na modelagem PLS.

Contudo os modelos iPLS, siPLS e GA-PLS para os três analitos, apresentaram a escolha de um elevado número de fatores PLS para um sistema que apresenta apenas três compostos absorventes. Supondo a existência de algum efeito sinérgico resultante da mistura dos corantes podemos supor que 4 fatores seria maior valor plausível, entretanto em alguns casos este número chega a ser igual a 10. E os gráficos de resíduos para o modelo global indicam que três fatores são suficientes para modelar adequadamente os dados.

Tabela 4.1- Resumo dos resultados de calibração e validação externa para o corante tartrazina.

Modelo	RMSEC (mg/L)	RMSECV (mg/L)	RMSEP (mg/L)	R²_{cal}	R²_{pred}	Fatores PLS	QVS
PLS	0.0928	0.0979	0.1103	0.9986	0.9919	3	217
PLS-JK	0.0917	0.0967	0.1079	0.9986	0.9922	3	178
GA-PLS	0.0820	0.0866	0.0773	0.9989	0.9960	3	33
iPLS (5^a)	0.0235	0.0306	0.0362	0.9999	0.9991	6	43
iPLS (10^a)	0.0419	0.0523	0.0544	0.9997	0.9980	5	22
iPLS (15^a)	0.0530	0.0579	0.0600	0.9995	0.9976	4	14
siPLS (2^b-5^a)	0.0164	0.0245	0.0298	0.9999	0.9997	9	87
siPLS (2^b-10^a)	0.0186	0.0279	0.0388	0.9999	0.9989	10	44
siPLS (2^b-15^a)	0.0212	0.0327	0.0505	0.9999	0.9983	10	30
siPLS (3^b-5^a)	0.0204	0.0226	0.0321	0.9999	0.9993	6	131
siPLS (3^b-10^a)	0.0179	0.0229	0.0348	0.9999	0.9991	7	66
siPLS (3^b-15^a)	0.0175	0.0257	0.0356	0.9999	0.9992	10	44
iSPA- PLS (5^a)	0.0595	0.0628	0.0337	0.9994	0.9994	3	43
iSPA- PLS (10^a)	0.0574	0.0607	0.0479	0.9995	0.9986	3	44
iSPA- PLS (15^a)	0.0595	0.0630	0.0431	0.9994	0.9988	3	30

^anúmero de intervalos ^bnúmero de combinações

Tabela 4.2. Resumo dos resultados de calibração e validação externa para o corante amarelo crepúsculo.

Modelo	RMSEC (mg/L)	RMSECV (mg/L)	RMSEP (mg/L)	R²_{cal}	R²_{pred}	Fatores PLS	QVS
PLS	0.1387	0.1457	0.2002	0.9968	0.9733	3	217
PLS-JK	0.1383	0.1451	0.1993	0.9968	0.9735	3	162
GA-PLS	0.0507	0.0560	0.0692	0.9996	0.9968	6	14
iPLS (5^a)	0.0683	0.0931	0.1435	0.9993	0.9867	7	44
iPLS (10^a)	0.0739	0.0969	0.1435	0.9991	0.9874	7	22
iPLS (15^a)	0.1025	0.1186	0.1972	0.9982	0.9741	5	16
siPLS (2^b-5^a)	0.0425	0.0512	0.0850	0.9997	0.9952	7	87
siPLS (2^b-10^a)	0.0411	0.0509	0.0823	0.9997	0.9955	7	44
siPLS (2^b-15^a)	0.0452	0.0534	0.1067	0.9996	0.9924	7	29
siPLS (3^b-5^a)	0.0409	0.0506	0.0767	0.9997	0.9961	8	131
siPLS (3^b-10^a)	0.0423	0.0501	0.0801	0.9997	0.9957	7	66
siPLS (3^a-15^b)	0.0434	0.0494	0.0800	0.9997	0.0089	6	44
iSPA-PLS (5^a)	0.1045	0.1109	0.0960	0.9980	0.9974	3	87
iSPA-PLS (10^a)	0.0667	0.0714	0.0836	0.9993	0.9985	3	65
iSPA-PLS (15^a)	0.0583	0.0625	0.0658	0.9994	0.9980	3	59

^anúmero de intervalos ^bnúmero de combinações

Tabela 4.3- Resumo dos resultados de calibração e validação externa para o corante vermelho-40.

Modelo	RMSEC (mg/L)	RMSECV (mg/L)	RMSEP (mg/L)	R²_{cal}	R²_{pred}	Fatores PLS	QVS
PLS	0.16290	0.17100	0.22020	0.9956	0.9678	3	217
PLS-JK	0.16210	0.17010	0.16210	0.9956	0.9956	3	171
GA-PLS	0.02920	0.03440	0.03980	0.9998	0.9989	7	34
iPLS (5)	0.03890	0.05430	0.06770	0.9997	0.9993	7	43
iPLS (10)	0.05210	0.06700	0.09300	0.9995	0.9942	6	22
iPLS (15)	0.06210	0.07020	0.07460	0.9994	0.9990	4	14
siPLS (2-5)	0.03270	0.03530	0.03240	0.9998	0.9993	5	87
siPLS (2-10)	0.03770	0.04080	0.05350	0.9998	0.9993	4	43
siPLS (2-15)	0.03470	0.04110	0.06630	0.9998	0.9970	6	29
siPLS (3-5)	0.03090	0.03510	0.03930	0.9998	0.9989	7	131
siPLS (3-10)	0.03220	0.03460	0.03630	0.9998	0.9991	6	66
siPLS (3-15)	0.03120	0.03470	0.04500	0.9998	0.9986	6	44
iSPA-PLS (5)	0.10100	0.10630	0.11050	0.9983	0.9990	3	43
iSPA-PLS (10)	0.08910	0.09330	0.08800	0.9987	0.9972	3	22
iSPA-PLS (15)	0.07550	0.07920	0.08200	0.9990	0.9988	3	14

^anúmero de intervalos ^bnúmero de combinações

O desempenho do iSPA-PLS foi comprado com os demais métodos por meio de um teste F , conforme descrito no Capítulo 3, e os resultados são apresentados nas **Tabelas 4.4, 4.5 e 4.6** para os corantes tartrazina, amarelo crepúsculo e vermelho 40 respectivamente. O valor de $F_{\text{critico}}(27,27,0.95) = 1.9048$.

Tabela 4.4-Valores de F calculado para o corante tartrazina

	iSPA-PLS (5)	iSPA-PLS (10)	iSPA-PLS (15)
PLS	10.7047	5.2986	6.5446
PLS-JK	10.2495	5.0733	6.2663
GA-PLS	5.2627	2.6050	3.2175
iPLS (5)	1.1539	1.7509	0.7054
iPLS (10)	2.6058	1.2898	1.5931
iPLS (15)	3.1741	1.5711	1.9406
siPLS (2-5)	1.2789	2.5837	2.0918
siPLS (2-10)	1.3256	1.5241	1.2339
siPLS (2-15)	2.2456	1.1115	1.3729
siPLS (3-5)	1.1022	2.2267	1.8028
siPLS (3-10)	1.0663	1.8946	1.5339
siPLS (3-15)	1.1159	1.8104	1.4657

Tabela 4.5-Valores de F calculado para o corante amarelo crepúsculo.

	iSPA-PLS (5)	iSPA-PLS (10)	iSPA-PLS (15)
PLS	4.3490	5.7348	9.2571
PLS-JK	4.3099	5.6833	9.1741
GA-PLS	1.9246	1.4595	1.1060
iPLS (5)	2.2344	2.9464	4.7561
iPLS (10)	0.4475	2.9464	4.7561
iPLS (15)	4.2196	5.5642	8.9818
siPLS (2-5)	0.7840	1.0338	1.6687
siPLS (2-10)	1.3606	1.0318	1.5644
siPLS (2-15)	1.2353	1.6290	2.6295
siPLS (3-5)	1.5666	1.1880	1.3587
siPLS (3-10)	1.4364	1.0893	1.4819
siPLS (3-15)	1.4400	1.0920	1.4782

Com base nos resultados apresentados podemos observar que em todos os casos o algoritmo proposto é melhor que o modelo global, e é comprável aos demais algoritmos. Contudo a situações em que o siPLS ou iPLS apresentam resultados melhores, estes

empregam um elevado número de fatores tornando os modelos mais complexos e passível de sobreajuste.

Tabela 4.6-Valores de F calculado para o corante vermelho 40.

	iSPA-PLS (5)	iSPA-PLS (10)	iSPA-PLS (15)
PLS	4.3490	5.7348	9.2571
PLS-JK	4.3099	5.6833	9.1741
GA-PLS	1.9246	1.4595	1.1060
iPLS (5)	2.2344	2.9464	4.7561
iPLS (10)	0.4475	2.9464	4.7561
iPLS (15)	4.2196	5.5642	8.9818
siPLS (2-5)	0.7840	1.0338	1.6687
siPLS (2-10)	1.3606	1.0318	1.5644
siPLS (2-15)	1.2353	1.6290	2.6295
siPLS (3-5)	1.5666	1.1880	1.3587
siPLS (3-10)	1.4364	1.0893	1.4819
siPLS (3-15)	1.4400	1.0920	1.4782

Nas **Figuras 4.4, 4.5 e 4.6** são apresentados as variáveis selecionadas pelo GA e pelo Jack-Knife. É possível perceber que enquanto o Jack-Knife seleciona um elevado número de variáveis, levando a resultados similares ao modelo global, o GA seleciona poucas variáveis, contudo emprega um numero bem maior de fatores.

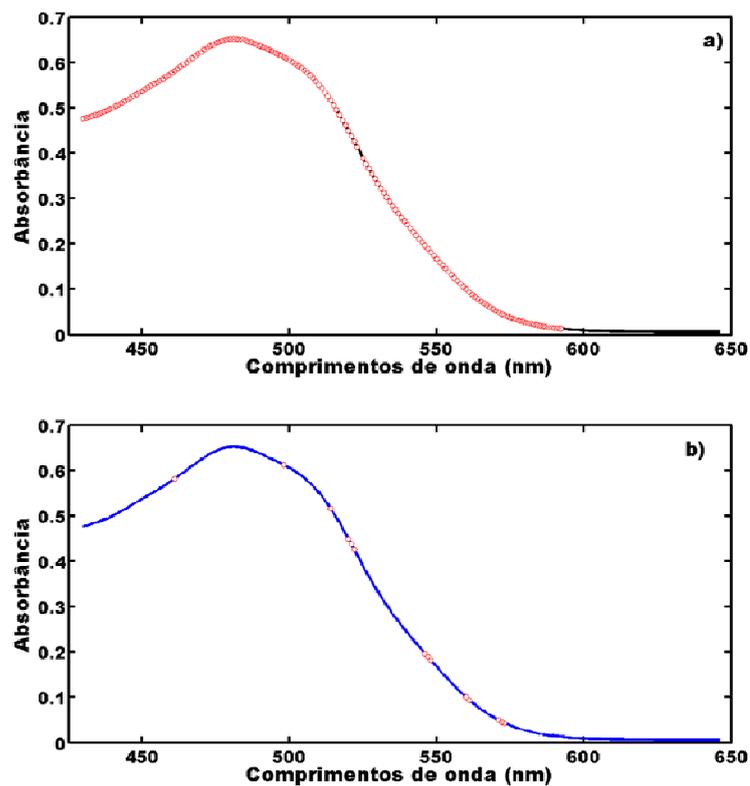


Figura 4.4- Variáveis selecionadas pelos algoritmos (a) Jack-Knife e (b) GA para o corante AC.

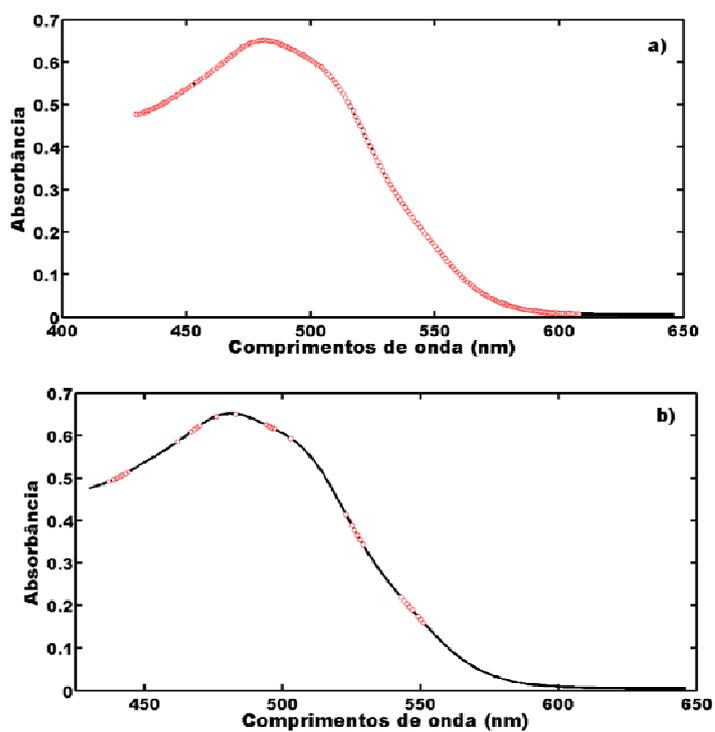


Figura 4.5- Variáveis selecionadas pelos algoritmos (a) Jack-Knife e (b) GA para o corante TAR.

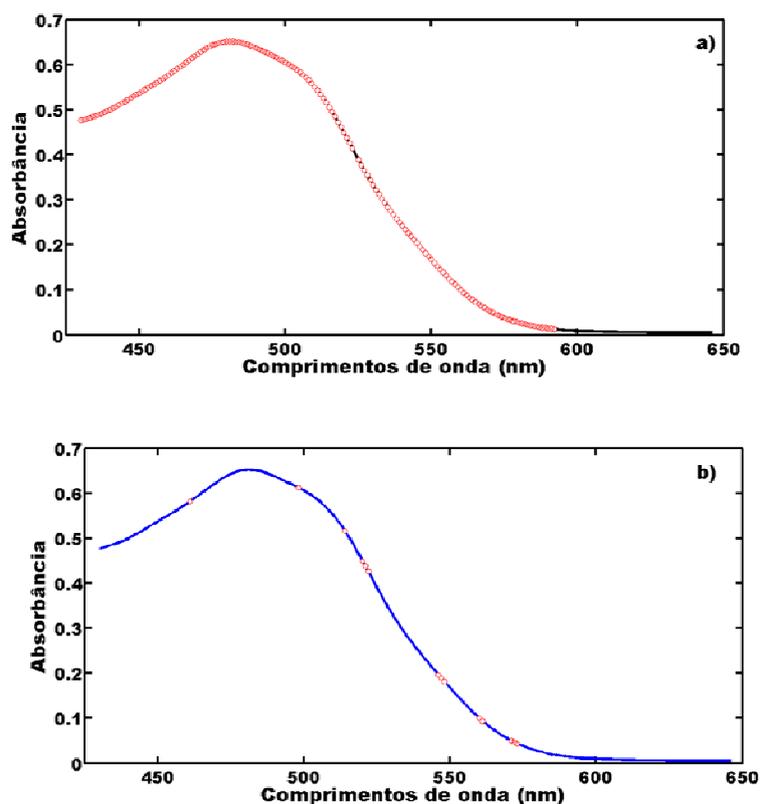


Figura 4.6- Variáveis selecionadas pelos algoritmos (a) Jack-Knife e (b) GA para o corante V40 .

Nas **Figuras 4.7, 4.8 e 4.9** são apresentados os intervalos selecionados pelos algoritmos *iPLS*, *siPLS* e *iSPA-PLS*. A região em torno de 550 nm parece ser a mais informativa para efeito de quantificação dos corantes, pois para todos os casos intervalos foram selecionados nesta região.

Para todos os corantes os resultados em termos de erros foram similares mesmo empregando faixas distintas do espectro, contudo as regiões selecionados pelo *iSPA-PLS* conduziu a modelos mais parcimoniosos em termos de fatores PLS.

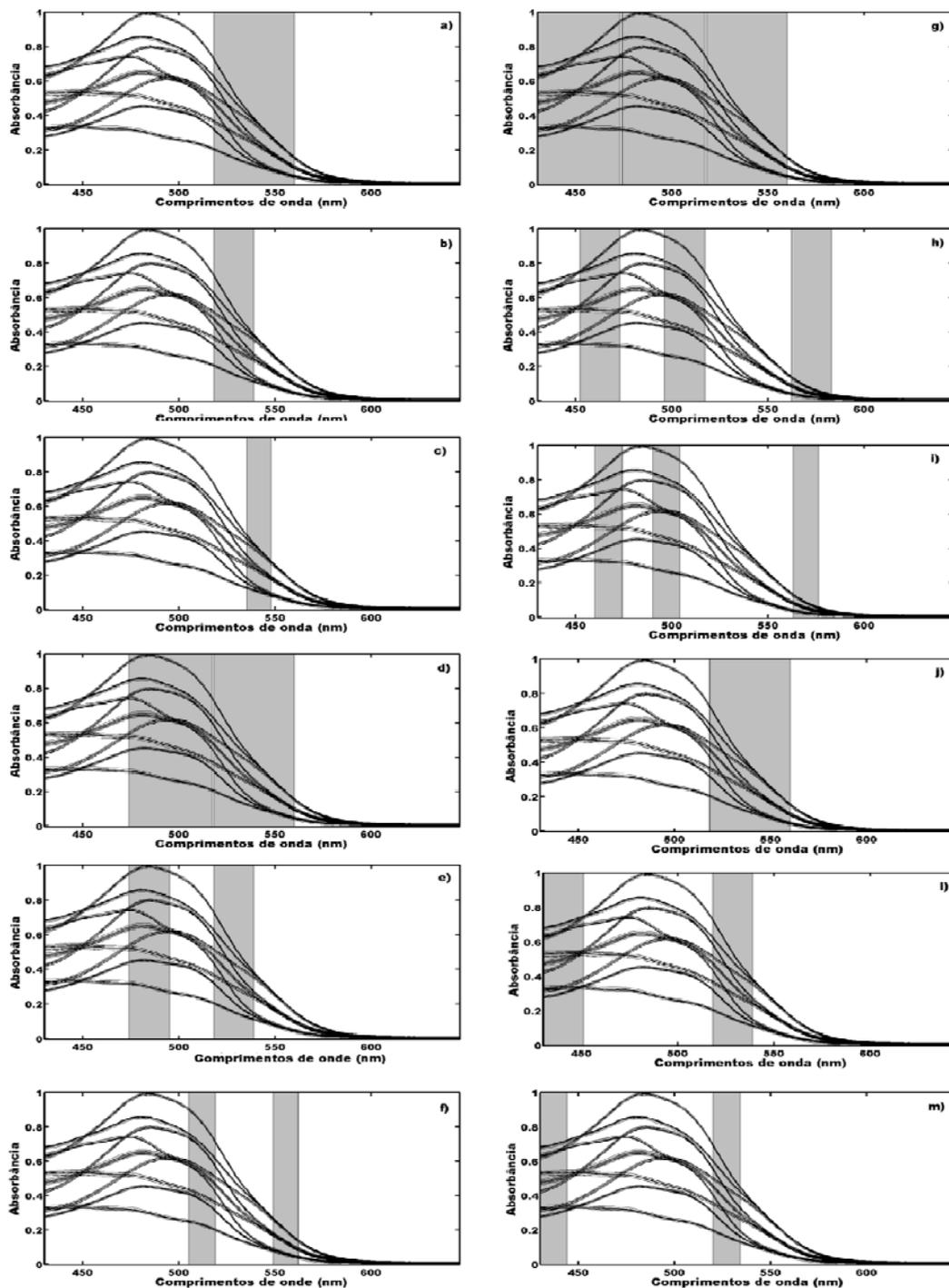


Figura 4.7: Variáveis selecionadas pelos algoritmos: (a) iPLS-5 intervalos, (b) iPLS-10 intervalos, (c) iPLS-15 intervalos, (d) siPLS-5 intervalos em combinação de 2, (e) siPLS-10 intervalos em combinação de 2, (f) siPLS-15 intervalos em combinação de 2, (g) siPLS-5 intervalos em combinação de 3, (h) siPLS-10 intervalos em combinação de 3, (i) siPLS-15 intervalos em combinação de 3, (j) iAPS-PLS-5 intervalos, (l) iAPS-PLS-10 intervalos e (m) iAPS-PLS-15 intervalos para o corante AC.

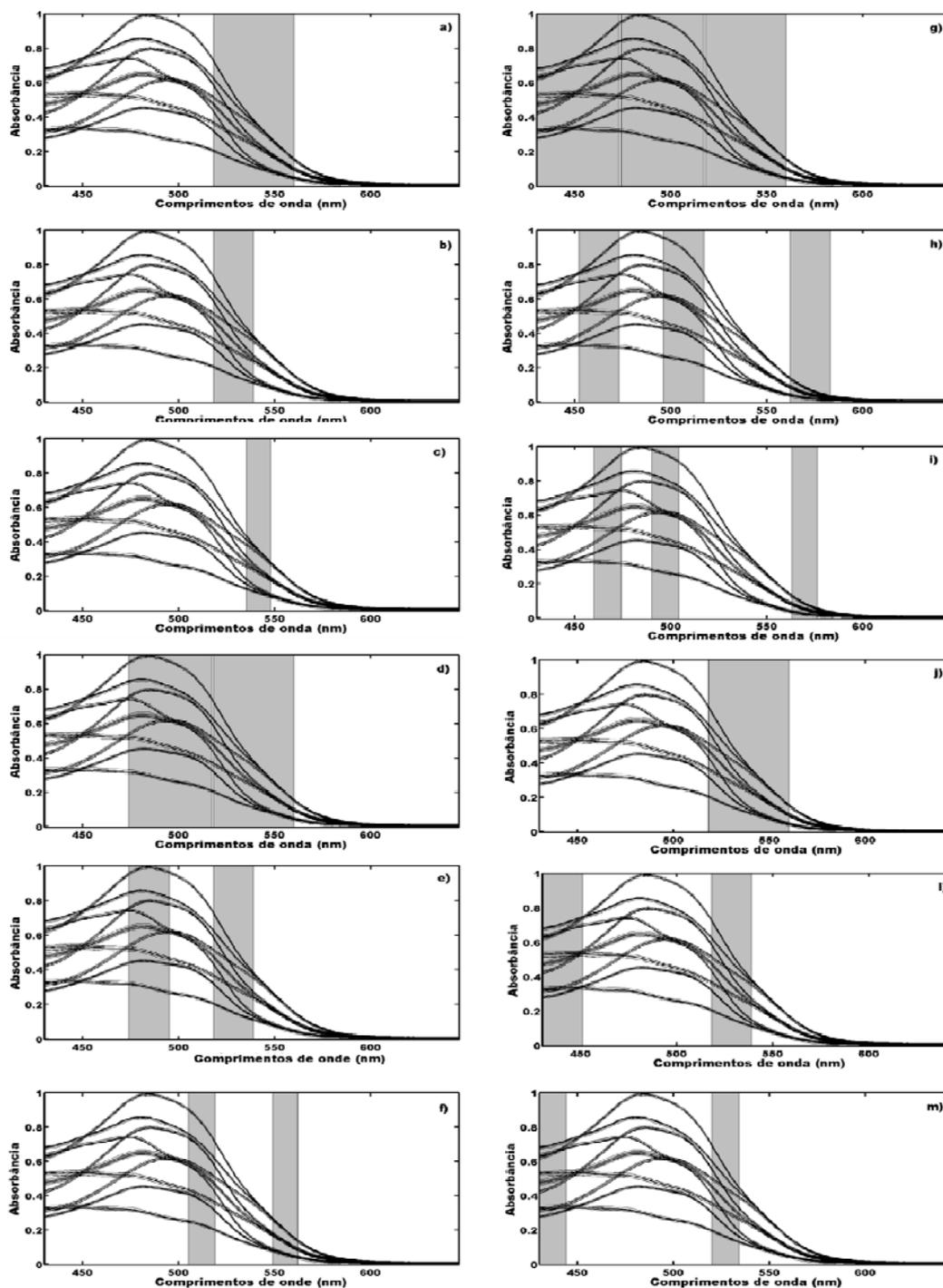


Figura 4.8: Variáveis selecionadas pelos algoritmos: (a) iPLS-5 intervalos, (b) iPLS-10 intervalos, (c) iPLS-15 intervalos, (d) siPLS-5 intervalos em combinação de 2, (e) siPLS-10 intervalos em combinação de 2, (f) siPLS-15 intervalos em combinação de 2, (g) siPLS-5 intervalos em combinação de 3, (h) siPLS-10 intervalos em combinação de 3, (i) siPLS-15 intervalos em combinação de 3, (j) iAPS-PLS-5 intervalos, (l) iAPS-PLS-10 intervalos e (m) iAPS-PLS-15 intervalos para o corante TAR.

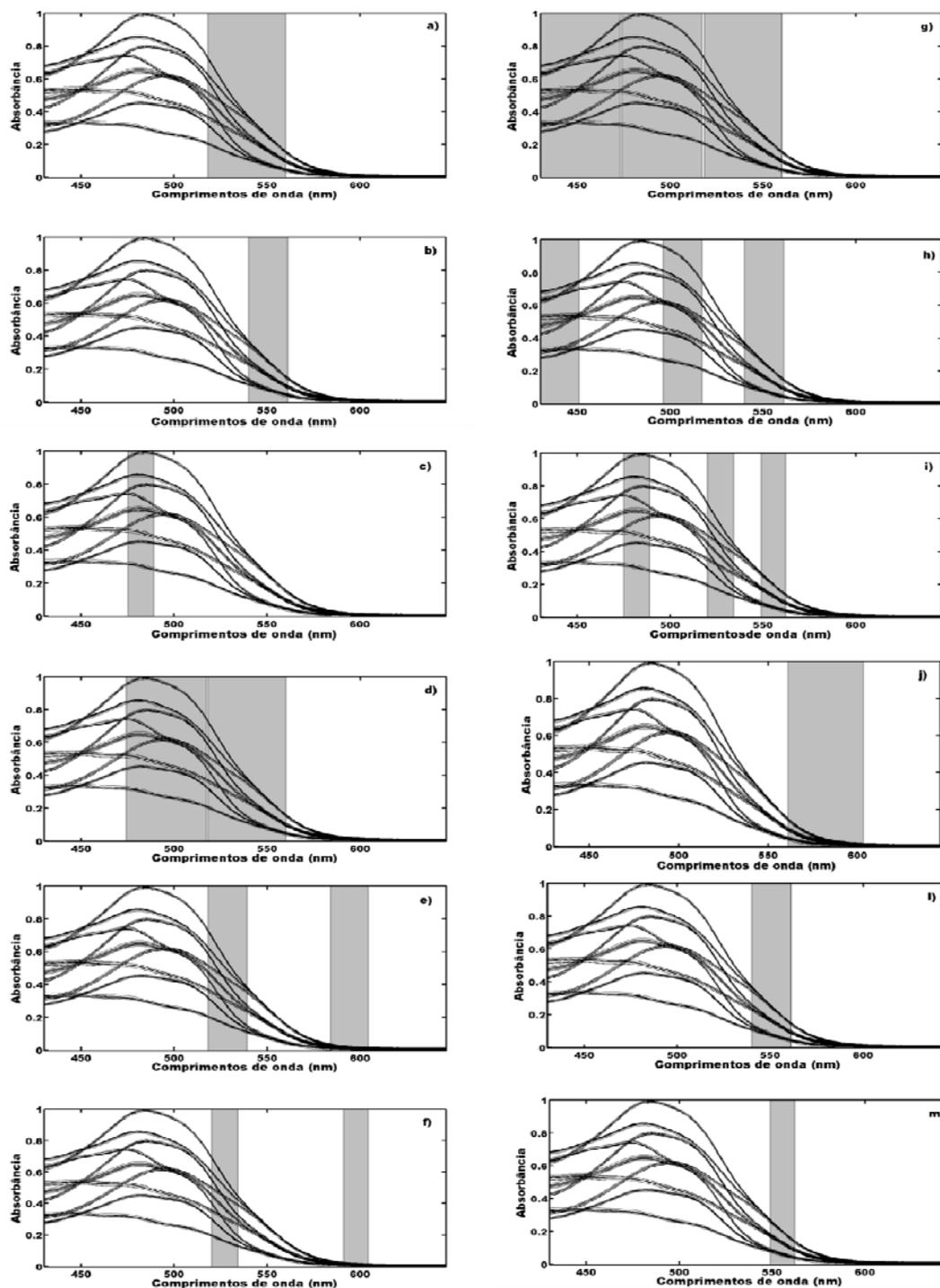


Figura 4.9: Variáveis selecionadas pelos algoritmos: (a) iPLS-5 intervalos, (b) iPLS-10 intervalos, (c) iPLS-15 intervalos, (d) siPLS-5 intervalos em combinação de 2, (e) siPLS-10 intervalos em combinação de 2, (f) siPLS-15 intervalos em combinação de 2, (g) siPLS-5 intervalos em combinação de 3, (h) siPLS-10 intervalos em combinação de 3, (i) siPLS-15 intervalos em combinação de 3, (j) iAPS-PLS-5 intervalos, (l) iAPS-PLS-10 intervalos e (m) iAPS-PLS-15 intervalos para o corante V40.

CAPÍTULO 5

Determinação
do teor de proteínas
em amostras de
trigo

5.0 DETERMINAÇÃO DO TEOR DE PROTEÍNA EM AMOSTRAS DE TRIGO

Na **Figura 5.1** os espectros brutos são apresentados, e como podem ser observados, os dados apresentam forte perfil de linha de base e para contornar este inconveniente foi empregado processo derivativo de primeira ordem com algoritmo Savitzky-Golay [77]. Para isso, ajustou-se um polinômio de segunda ordem e janela móvel de 17 pontos. Os espectros derivativos são mostrados na **Figura 5.2** abaixo.

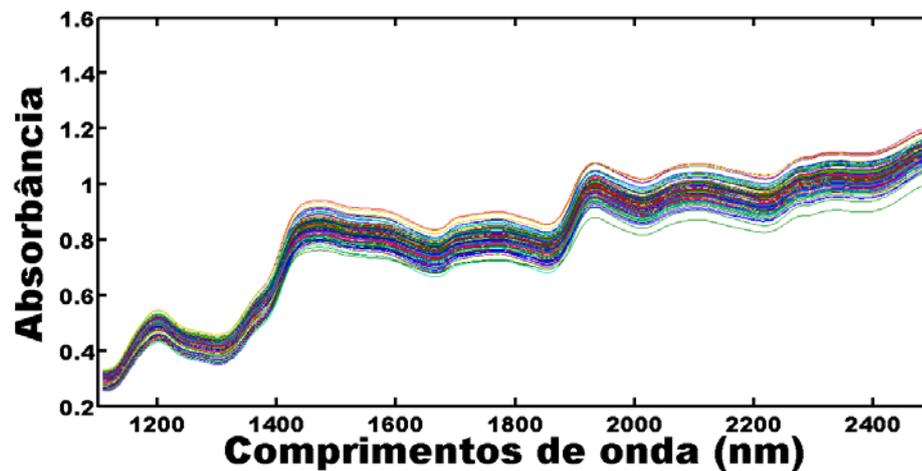


Figura 5.1: Espectros brutos das amostras de trigo.

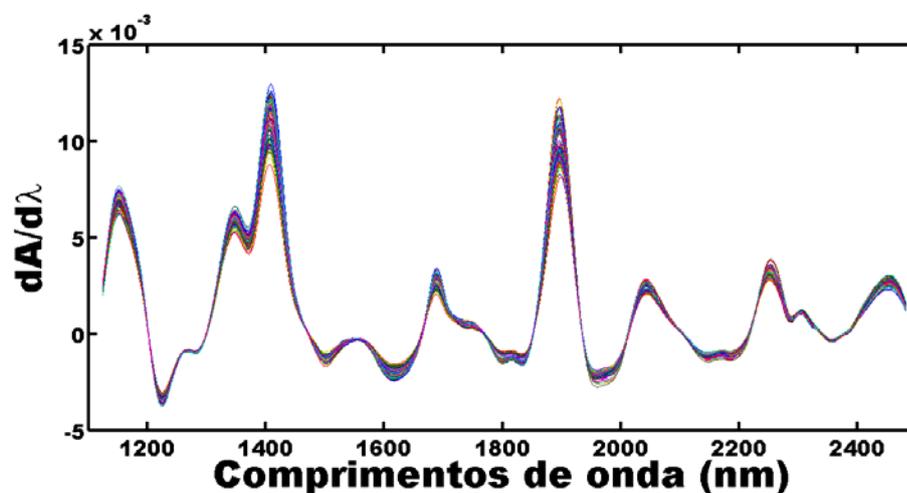


Figura 5.2: Espectros derivativos das amostras de trigo.

Com base no gráfico dos espectros derivativos é possível observar que o pré-processamento empregado corrigiu o deslocamento de linha de base satisfatoriamente.

Utilizando os espectros resultantes do processo derivativo, assim como no estudo de casos dos corantes, o posto do modelo global também foi investigado para este caso. Na **Figura 5.3** é apresentado o gráfico de RMSECV para o modelo global em função do número de fatores incluídos no modelo.

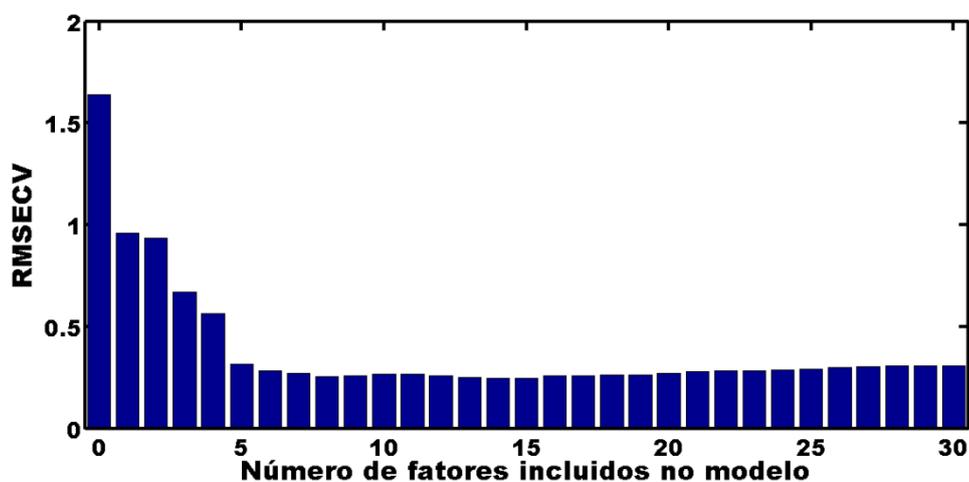


Figura 5.3- RMSECV versus número de fatores PLS incluídos no modelo global

Com base na **Figura 5.3**, é possível observar que ocorre decaimento do RMSECV até o quinto fator, e que após este valor não há diminuição pronunciada do erro de validação cruzada. Portanto cinco fatores parecem descrever satisfatoriamente o conjunto de dados. Para este caso os gráficos de resíduos não parecem informativos, portanto não serão apresentados.

5.1 Quantificação do teor de proteína

A quantificação do teor de proteína em trigo foi realizada empregando modelo PLS global e o algoritmo proposto. Para fins de comparação, modelos com seleção de variáveis individuais (PLS Jack-

Knife e GA-PLS) e em intervalos (*iPLS* e *siPLS*) foram testados. Na **Tabela 5.1** é apresentada um resumo dos parâmetros estatísticos obtidos nas fases de calibração e predição.

Tabela 5.1- Quantificação de proteína em trigo: parâmetros estatísticos.

Modelo	RMSEC (% m/m)	RMSECV (% m/m)	RMSEP (% m/m)	R^2_{cal}	R^2_{pred}	Fatores PLS	QVS
PLS	0.25	0.32	0.25	0.9766	0.9732	5	680
PLS-JK	0.23	0.28	0.24	0.9797	0.9767	5	312
GA-PLS	0.20	0.22	0.26	0.9955	0.9639	7	33
<i>iPLS</i> (10 ^a)	0.16	0.21	0.20	0.9905	0.99	11	68
<i>iPLS</i> (15 ^a)	0.19	0.22	0.23	0.987	0.9706	8	48
<i>iPLS</i> (20 ^a)	0.18	0.21	0.21	0.9876	0.9772	6	34
<i>siPLS</i> (2 ^b -10 ^a)	0.14	0.20	0.20	0.9982	0.979	13	136
<i>siPLS</i> (2 ^b -15 ^a)	0.16	0.20	0.18	0.9904	0.9833	10	92
<i>siPLS</i> (2 ^b -20 ^a)	0.16	0.20	0.17	0.9906	0.9841	10	68
<i>siPLS</i> (3 ^b -10 ^a)	0.14	0.20	0.22	0.9923	0.9731	11	199
<i>siPLS</i> (3-15)	0.14	0.19	0.24	0.9922	0.9689	11	138
<i>siPLS</i> (3-20)	0.13	0.18	0.21	0.9938	0.9769	14	102
<i>iSPA-PLS</i> (10 ^a)	0.20	0.22	0.24	0.9857	0.9687	5	204
<i>iSPA-PLS</i> (15 ^a)	0.19	0.22	0.24	0.9857	0.9687	5	138
<i>iSPA-PLS</i> (20 ^a)	0.19	0.21	0.23	0.9869	0.9742	5	136

^anúmero de intervalos ^bnúmero de combinações

Em linhas gerais, os métodos de seleção de variáveis produziram menores erros de predição quando comparado ao modelo global, sendo os melhores resultados obtidos para modelos *siPLS*. Entretanto o método proposto obteve resultados similares do ponto de vista estatístico, como pode ser observado na **Tabela 5.2** ($F_{critico}(40,40,0.95) = 1.6928$), contudo o *iSPA-PLS* para todos os

casos empregou um número menor de fatores, levando a modelos mais parcimoniosos

Tabela 5.2: Valores de F calculado para comparação dos modelos iSPA-PLS com os demais algoritmos.

	iSPA-PLS (10)	iSPA-PLS (15)	iSPA-PLS (20)
PLS	1.0522	1.1031	1.2252
PLS-JK	1.0634	1.0144	1.0949
AG-PLS	1.1335	1.1883	1.3199
iPLS (10)	1.5615	1.4895	1.3411
iPLS (15)	1.0844	1.0344	1.0737
iPLS (20)	1.3996	1.3351	1.2020
siPLS (2-10)	1.5178	1.4478	1.3035
siPLS (2-15)	1.9137	1.8254	1.6435
siPLS (2-20)	1.9964	1.9044	1.7146
siPLS (3-10)	1.1842	1.1296	1.0170
siPLS (3-15)	1.0233	1.0245	1.1378
siPLS (3-20)	1.3716	1.3083	1.1779

Quando comparado ao modelo global, o iSAP-PLS não promoveu melhoria estatisticamente significativa do erro de predição, tão pouco emprega um número menor de fatores. Toda via utiliza uma faixa menor do espectro, portanto o tempo de análise para amostras de predição será reduzido, o que representa uma vantagem do ponto de vista prático.

Comparando os algoritmos em termos de variáveis selecionadas, o Jack-Knife (**Figura 5.4a**) selecionou variável praticamente por todo espectro, quase em faixas contínuas, levando a resultados em termos de erros de predição similar aos resultados obtidos para os métodos de seleção de intervalos. O GA (**Figura 5.4b**) selecionou variáveis basicamente associado aos sobretons NH de proteínas ^[78], e obteve resultados ligeiramente maiores que o modelo PLS global.

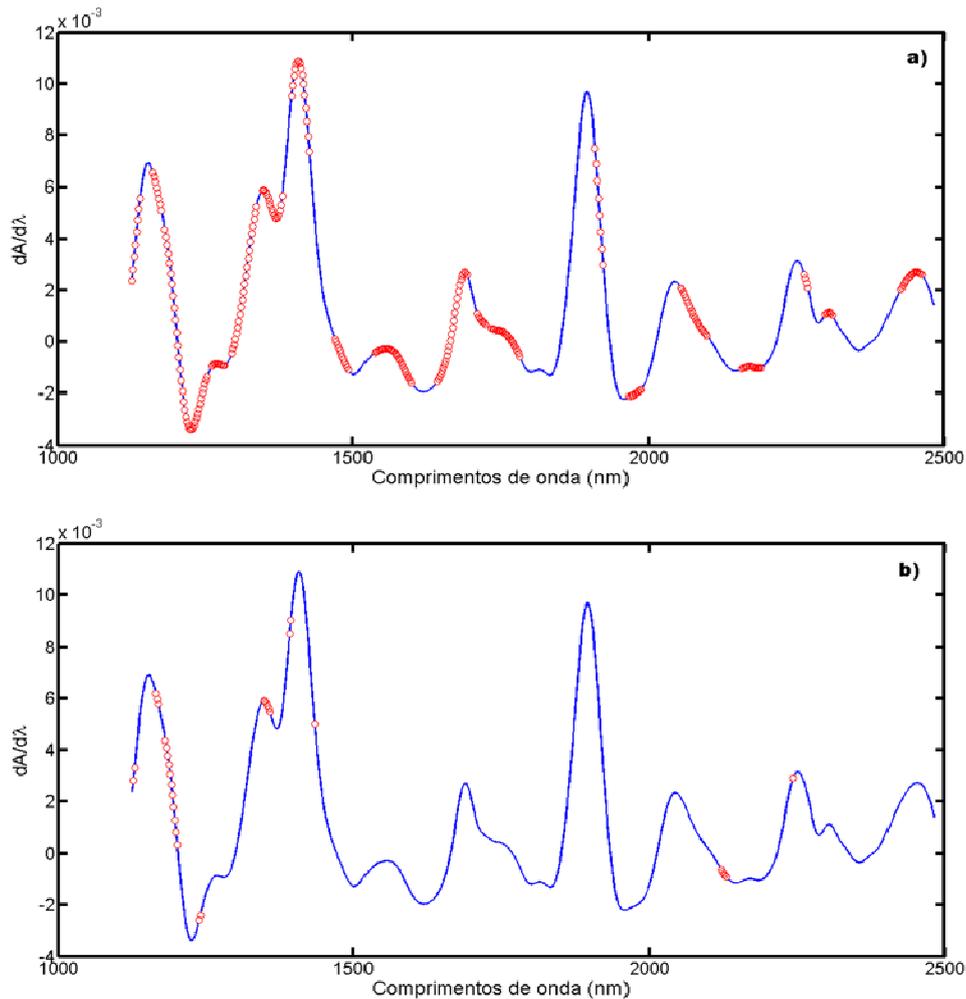


Figura 5.4: Variáveis selecionadas pelos algoritmos (a) Jack-Knife e (b) GA.

A seleção de intervalos com o iPLS, conduziu a seleção de uma faixa espectral (**Figura 5.5 a,b e c**) em torno do terceiro sobreton NH (1200 nm) [78]. A seleção desta faixa é compreensível uma vez que existe uma relação de proporcionalidade entre NH e teor e proteína [1] relação entre proteínas e a quantidade de nitrogênio em uma amostra.

Empregando o método siPLS em combinação de dois intervalos (**Figuras 5.5 d, e e f**), para o caso em que o espectro estava particionado em dez faixas, o siPLS selecionou o mesmo intervalo selecionado pelo iPLS. O segundo intervalo selecionado foi na região de primeiro sobreton CH (1700 nm) [78]. Com o aumento do número

de intervalos em que o espectro foi particionado, a primeira faixa se manteve na região do terceiro sobreton NH. Contudo, foi observado um deslocamento do segundo intervalo selecionado para região do terceiro sobreton CH.

Já para os modelos siPLS em combinação de três intervalos (**Figuras 3.4 g, h e i**), as faixas selecionados não apresentaram grande deslocamento em função do número de intervalos com que os espectros foram particionados. Neste caso, foram sempre selecionados dois intervalos na região de 1200 a 1400 nm e outra por volta de 2200 nm.

Ao contrario do iPLS, que seleciona um único intervalo e do siPLS que obrigatoriamente seleciona 2, 3 ou 4 intervalos conforme especificado previamente, o iSPA-PLS pode selecionar de 1 a $j-1$ intervalos, onde j representa a quantidade de intervalos. Caso um único intervalo seja a solução mais adequada um único intervalo será selecionado, mas se outros intervalos forem benéficos ao modelo esses também terão oportunidade de serem selecionados.

Para o caso em que os espectros foram particionados em dez intervalos (**Figura 5.45 j**) o iSPA-PLS selecionou três intervalos, contudo havia possibilidade de selecionar entre 1 e 9, dois dos quais são similares ao resultado do siPLS (**Figura 5.5 g**). O terceiro intervalo está localizado na região de primeiro sobreton CH por volta de 1700 nm.

Comparando o caso em que os espectros estavam particionados em quinze intervalos, o iSPA-PLS também apresenta como solução um subconjunto de intervalos que mostra melhores resultados que o siPLS e o iPLS em termos de parcimônia e equivalente poder de predição. Este algoritmo também selecionou três intervalos. Finalmente, quando os espectros foram particionados em vinte

intervalos o iSPA-PLS selecionou 4 intervalos, obtendo o menor erro entre os modelos iSPA-PLS.

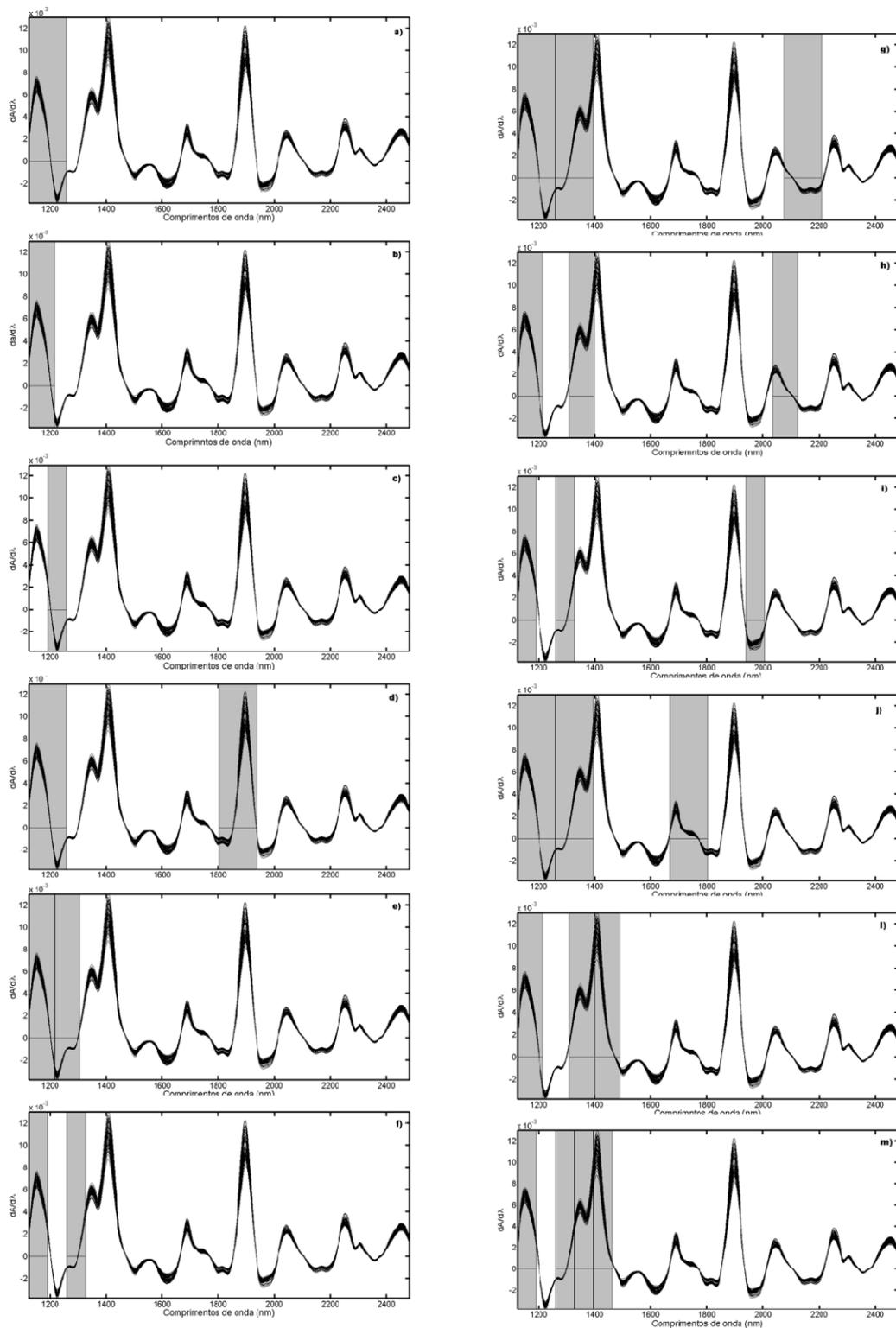


Figura 5.5: Variáveis selecionadas pelos algoritmos: (a) iPLS-10 intervalos, (b) iPLS-15 intervalos, (c) iPLS-20 intervalos, (d) siPLS-10 intervalos em combinação de 2, (e) siPLS-15 intervalos em combinação de 2, (f) siPLS-20 intervalos em combinação de 2, (g) siPLS-10 intervalos em combinação de 3, (h)

siPLS-15 intervalos em combinação de 3, (i) siPLS-20 intervalos em combinação de 3, (j) iAPS-PLS-10 intervalos, (l) iAPS-PLS-15 intervalos e (m) iAPS-PLS-20 intervalos.

CAPÍTULO 6

Determinação da
qualidade de amostras de
extrato de cervejas

6.0 DETERMINAÇÃO DA QUALIDADE DE AMOSTRAS DE EXTRATO DE CERVEJA

Na **Figura 6.1** são apresentados os espectros das amostras de extrato de cerveja. Observando esses dados, é possível ver um perfil suave em função dos comprimentos de onda. Na região do visível ocorre uma maior variabilidade nos espectros, enquanto na extrema direita os espectros são extremamente ruidosos. Outro problema associado a estes dados é o perfil de linha de base característico.

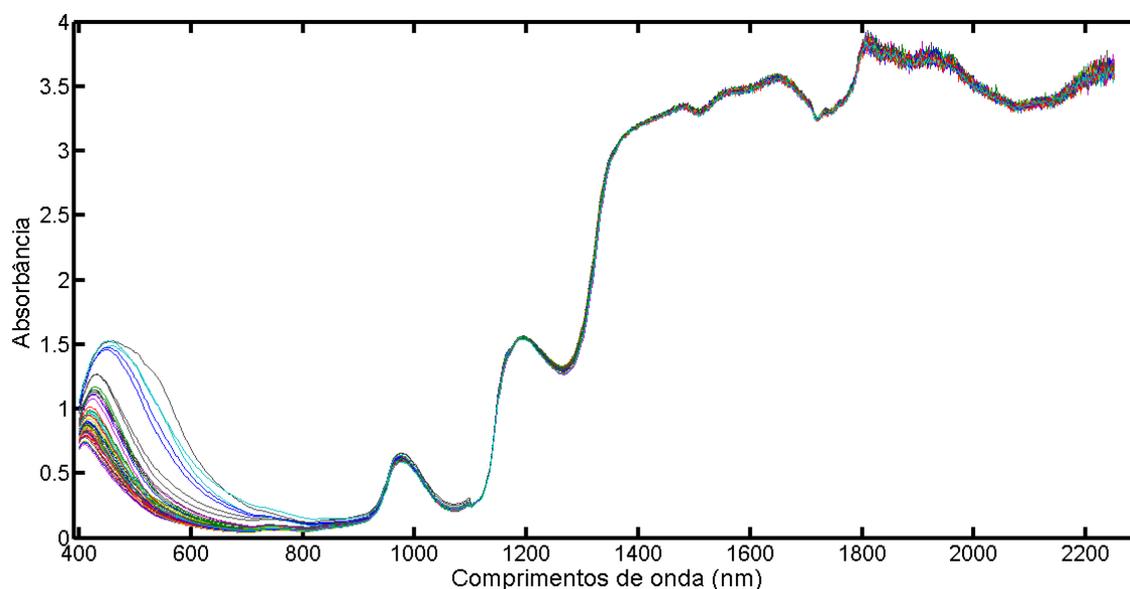


Figura 6.1: Espectros brutos das amostras de extrato de cervejas.

Qualquer especialista ou mesmo usuário da quimiometria chegaria à conclusão de que a faixa compreendida entre 1500 e 2250, não seria adequada para a construção de modelos de calibração devido ao nível de ruído e altos valores de absorbância. Portanto, esta deveria ser removida antes de qualquer tratamento do conjunto de dados. Neste capítulo, deseja-se observar e ilustrar a capacidade do algoritmo em não selecionar regiões pobres em correlação com o parâmetro de interesse bem como regiões ruidosas.

Na **Figura 6.2** é apresentado um gráfico do coeficiente de correlação (r) em função do parâmetro de interesse (vetor **y_{cal}**) e cada variável da matriz **X_{cal}**.

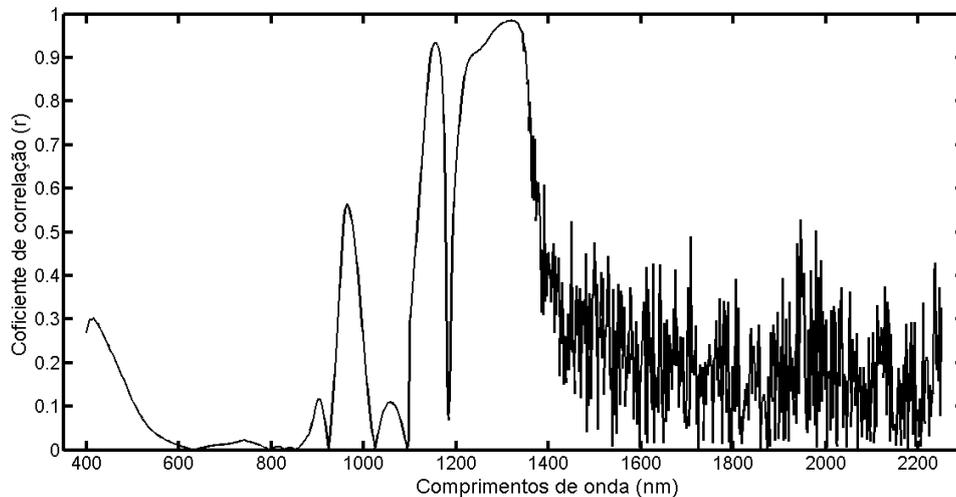


Figura 6.2: coeficiente de correlação entre **y_{cal}** e **X_{cal}**.

Com base na figura acima, observa-se claramente que nem todo canal analítico guarda correlação com o vetor da variável dependente. Inclusive, existe variáveis com correlação, pelo menos linear, próximo do zero, e há regiões onde o valor de r atinge praticamente seu valor máximo, que é igual a 1. Portanto já temos uma ideia prévia da região em que devem estar às variáveis selecionadas, ou em último caso a maioria delas.

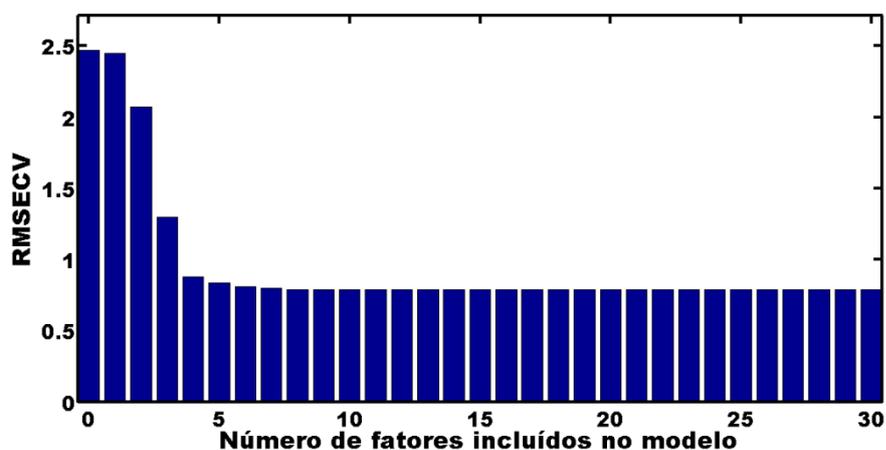


Figura 6.3- RMSECV versus número de fatores PLS incluídos no modelo global

Com relação ao posto do modelo global, como avaliado para os outros casos, com base na **Figura 6.3**, é possível perceber que cinco fatores descrevem satisfatoriamente a relação entre os espectros e a propriedade calibrada.

Na **Tabela 6.1**, é apresentada do resumo dos parâmetros estatísticos obtidos na fase de calibração e predição.

Tabela 6.1- Parâmetros estatísticos para calibração e predição da qualidade em amostras de cervejas.

Modelo	RMSEC	RMSECV	RMSEP	R ² _{cal}	R ² _{pred}	Fatores PLS	QVS
PLS	0.18	0.83	0.74	0.9945	0.9092	5	926
PLS-JK	0.26	0.41	0.27	0.9887	0.9874	4	104
GA-PLS	0.11	0.13	0.17	0.9957	0.9956	6	19
iPLS (5)	0.18	0.17	0.30	0.9976	0.9871	6	185
iPLS (10)	0.08	0.14	0.13	0.9990	0.9973	7	93
iPLS (20)	0.08	0.14	0.15	0.9988	0.9981	7	47
siPLS (2-5)	0.10	0.23	0.14	0.9982	0.9966	5	370
siPLS (2-10)	0.08	0.12	0.09	0.9989	0.9986	8	186
siPLS (2-20)	0.08	0.14	0.13	0.9989	0.9972	7	92
siPLS (3-5)	0.09	0.24	0.27	0.9984	0.9880	7	556
siPLS (3-10)	0.06	0.11	0.14	0.9994	0.9966	10	279
siPLS (3-20)	0.09	0.11	0.18	0.9987	0.9944	7	140
iAPS-PLS (5)	0.02	0.21	0.19	0.9955	0.9943	2	185
iAPS-PLS (10)	0.08	0.14	0.13	0.9990	0.9973	7	93
iAPS-PLS (20)	0.11	0.15	0.13	0.9977	0.9974	5	47

^aNúmero de intervalos, ^bcombinações

Neste caso, em que os dados foram empregados na forma bruta, todos os algoritmos mostram notável melhoria com relação ao modelo global. Os melhores resultados em termos de RMSEP foram obtidos para os modelos siPLS em combinação de dois intervalos.

Contudo para os casos de partição do espectro em 10 e 20 intervalos um número de fatores maior que o posto do modelo global é empregado. Comparando os modelos por meio do erro médio para um conjunto externo de amostras por meio de um teste F (**Tabela 6.2**), sendo o valor de $F_{\text{critico}}(20,20,0.95) = 2.1242$, é possível perceber que o método proposto é capaz de obter resultados muito melhores que o modelo global, comparável aos melhores resultados obtidos com os métodos já consolidados. Entretanto oferece a vantagem de ser mais parcimonioso, merecendo destaque para o caso iSPA-PLS (5) em que com apenas 2 fatores foi obtido RMSEP de 0.19, enquanto o RMSEP do modelo global empregando 5 fatores é 0.74 .

Tabela 6.2- Valores de F calculado para comparação dos modelos iSPA-PLS com os demais algoritmos.

	iSPA-PLS (5)	iSPA-PLS (10)	iSPA-PLS (15)
PLS	15.4863	30.9999	31.9603
PLS-JK	2.1571	4.3180	4.4517
AG-PLS	1.2391	1.6155	1.6655
iPLS (5)	2.4913	4.9870	5.1415
iPLS (10)	2.0018	1.0000	1.0310
iPLS (20)	1.5300	1.3083	1.3489
siPLS (2-5)	1.7011	1.1768	1.2132
siPLS (2-10)	4.1723	2.0843	2.0217
siPLS (2-20)	2.0991	1.0486	1.0171
siPLS (3-5)	2.0469	4.0975	4.2244
siPLS (3-10)	1.7106	1.1702	1.2065
siPLS (3-20)	1.0523	1.9024	1.9613

E por fim uma análise das variáveis selecionadas por cada algoritmo mostra que o Jack-Knife (**Figura 6.4ª**) que utiliza o conceito de estabilidade dos coeficientes de regressão levou a seleção

de variáveis em uma região de baixa relação sinal ruído, comprometendo o desempenho do respectivo modelo. O GA (**Figura 6.4b**) selecionou apenas variáveis em regiões de elevada correlação com a variável y .

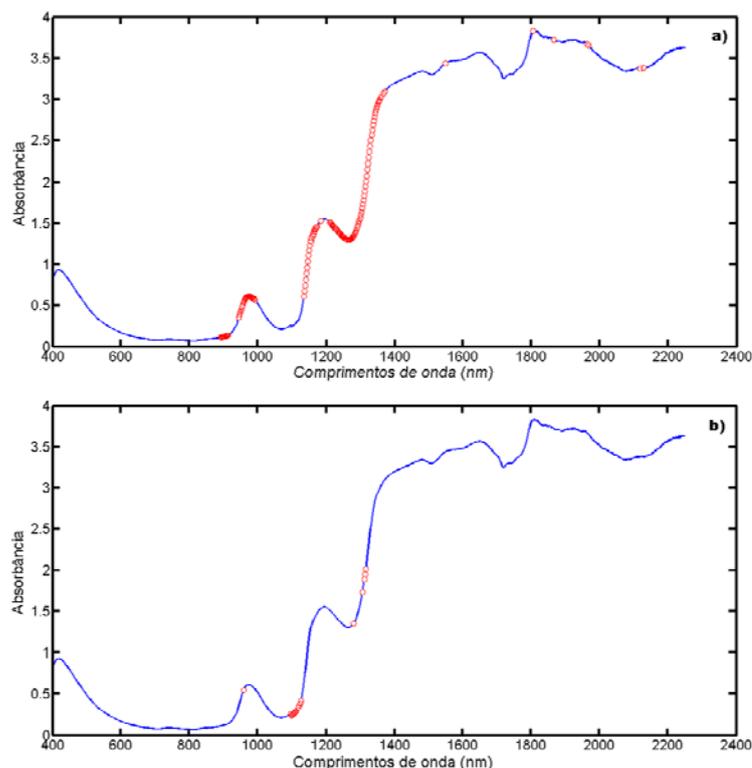


Figura 6.4- Variáveis selecionas (a) pelo Jack-Knie e (b) GA.

Os algoritmos que selecionam intervalos, não selecionaram variáveis em regiões ruidosa ou extremamente pobres em correlação com y . Esta característica pode ser apontada como uma vantagem dos métodos que selecionam intervalos, quando comparados aos que selecionam variáveis individuais.

Conforme observado na **Figura 6.1**, os espectro NIR dos extratos de cervejas apresentam duas regiões de correlação considerável com y . A primeira por volta de 1000 nm com valor de r por volta de 0.6 e outra região mais larga entre 1200 e 1400 com valores de correlação chegando próximo de 1.

Para os modelos iPLS que empregam uma única faixa do espectro, na situação em que os espectros estavam particionados em apenas cinco faixas, o intervalo selecionado (**Figura 6.5 a**) está fora da região onde a correlação entre **ycal** e **Xcal** é mais elevada. Isso ocorreu, possivelmente, pelo fato do tamanho do intervalo definido ser muito largo. Em outras palavras, o intervalo que continhas variáveis com elevada correlação também continha variáveis muito ruidosas.

Com o estreitamento dos intervalos (**Figuras 6.5 b e c**), o intervalo selecionado se desloca para região entre 1200 e 1400 com elevados valores de correlação. Os valores de RMSEP (**Tabela 6.1**) caem bastante quando comparamos os modelos gerados com o intervalo espectral da **Figura 6.5 a** com os das **Figuras 6.4 b e c**.

Os modelos siPLS com combinação de 2 (**Figuras 6.5 d, e e f**) e 3 (**Figuras 6.5 g, h e i**) intervalos para todos os casos selecionaram faixas espectrais na região de correlação elevada. Contudo, empregou-se sempre um número mais elevado de fatores obtendo, conseqüentemente, menores valores de erros de predição.

Os modelos iSPA-PLS selecionou, para todos os casos, apenas um intervalo, de modo a selecionar a região de elevada correlação entre **Xcal** e **ycal**, bem como utilizar o mínimo possível de fatores.

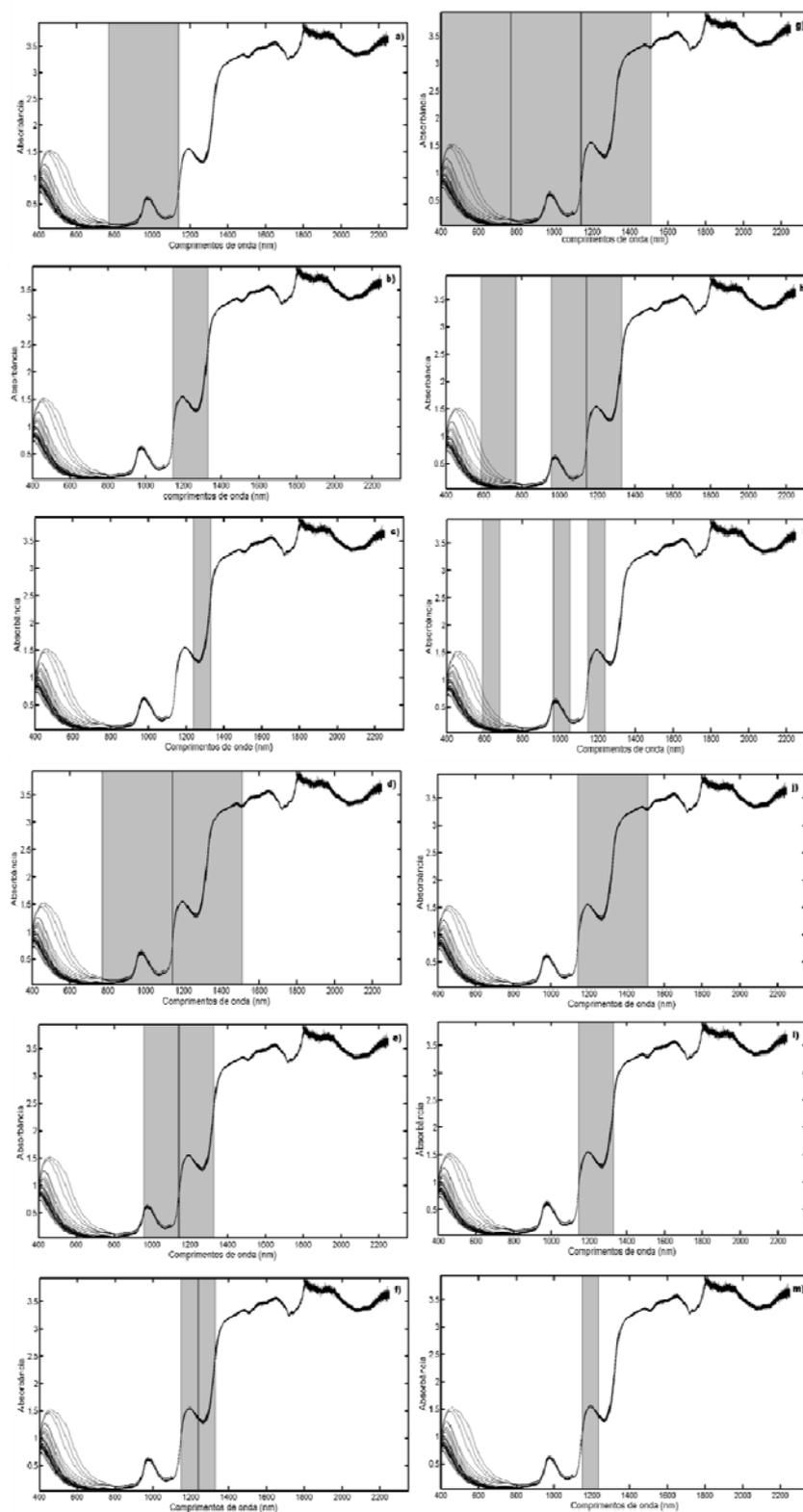


Figura 6.5: Variáveis selecionadas pelos algoritmos: (a) iPLS-5 intervalos, (b) iPLS-10 intervalos, (c) iPLS-20 intervalo, (d) siPLS-5 intervalos em combinação de 2, (e) siPLS-10 intervalos em combinação de 2, (f) siPLS-20 intervalos em combinação de 2, (g) siPLS-5 intervalos em combinação de 3, (h) siPLS-10 intervalos em combinação de 3, (i) siPLS-20 intervalos em combinação de 3, (j) iAPS-PLS-5 intervalos, (l) iAPS-PLS-10 intervalos e (m) iAPS-PLS-20 intervalos.

CAPÍTULO 7

Conclusão

7.0 CONCLUSÃO

Neste trabalho foi apresentada uma nova proposta de seleção de intervalos em regressão por mínimos quadrados parciais (PLS). O algoritmo proposto explorou a filosofia e funcionalidade do SPA, o qual é amplamente utilizado no contexto de calibração MLR e em problemas de classificação acoplado ao LDA.

O Algoritmo proposto, denominado iSPA-PLS, foi avaliado em três estudos de caso:

- ✓ Determinação simultânea de três corantes alimentícios em amostras sintéticas, em que foi possível perceber que para os três analitos o iSPA-PLS foi mais parcimonioso no número de fatores empregados nos modelos finais e obteve resultados melhores que o modelo global e equivalente aos demais métodos.
- ✓ Determinação do teor de proteína em trigo. Neste caso, o método proposto foi mais parcimonioso que os demais modelos que utilizaram seleção de variáveis.
- ✓ Determinação da qualidade de amostras de cervejas. Neste estudo de caso, avaliou-se a capacidade de selecionar boas faixas espectrais mesmo frente a espectros brutos, com regiões ruidosas e perfil de linha de base. Neste caso, o iSPA-PLS se mostrou eficaz ao selecionar sempre faixas em regiões de elevada correlação com a variável dependente. Isso produziu modelos superiores ao modelo global e não foi estatisticamente diferente dos modelos siPLS ou iPLS, embora estes empregue um número elevado de fatores. Por outro lado, o iSPA-PLS empregou apenas um número de fatores menor ou igual a número de fatores considerado ótimo para o modelo global.

- ✓ Em todos os casos, constatou-se que a seleção em intervalos se mostrou mais eficaz que a seleção de variáveis individuais na regressão PLS.

Por conseguinte, o iSPA-PLS pode ser considerado uma estratégia alternativa válida para fazer seleção de intervalos em regressão PLS, permitindo superar os inconvenientes do iPLS (seleciona uma única faixa) e do siPLS (seleciona sempre mais de uma faixa).

7.1 Propostas futuras

Como propostas de possíveis melhorias e aplicações da nova estratégia, pode-se destacar:

- ✓ Fazer a seleção de intervalos guiada por série de teste;
- ✓ Implementar o código iAPS-PLS em uma interface gráfica;
- ✓ Implementar na saída do algoritmo figuras de mérito baseadas no cálculo do NAS.
- ✓ Testar o desempenho do algoritmo em variáveis não espectrais;

REFERÊNCIAS

- [1] SKOOG, D. A. E LEARY, J. J. *Principles of instrumental analysis*. 6. ed. New York : Saunders College Publishing, **1992**
- [2] PIMENTEL, M. F., BARROS NETO, B. *Calibração: Uma revisão para Químicos Analíticos*. Química Nova. **19: 268, 1999.**
- [3] BARROS NETO, B., PIMENTEL, M. F., ARAÚJO, M. C. U. *Recomendações para calibração em química analítica- Parte I. Fundamentos e calibração com um componente (Calibração Univariada)*. Química Nova. **25: 856, 2002.**
- [4] INMETRO, *Orientação sobre validação de métodos de ensaios químicos*. DOQ-CGCRE/0008; **2003.**
- [5] BOKOBZA, L. *Near Infrared spectroscopy*. NIR Publications.
- [6] FERREIRA, M. M. C., et al. *Quimiometria I: Calibração Multivariada, Um Tutorial*. Química Nova. **22: 724, 1999.**
- [7] BRERETON, R. G. *Introduction to multivariate calibration in analytical chemistry*. Analyst. **125: 2125, 2000.**
- [8] NAES, TORMOD, et al. *A User-Friendly Guide to Multivariate Calibration and Classification*. Chichester, UK : NIR Publications, **2002.**
- [9] BEEBE, K. R., PELL, R. J. e SEASHOLTZ, M. B. *Chemometrics: A Practical Guide*. New York: John Wiley & Sons, **1998.**
- [10] DRAPER, N. R., SMITH, H. *Applied Regression Analysis*. New York: John Wiley & Sons, **1998.**
- [11] VALDERRAMA, P. *Calibração multivariada de primeira e segunda ordem e figuras de mérito na quantificação de enantiômeros por espectroscopia*. Tese de Doutorado. Campinas: UNICAMP, **2009.**
- [12] BRERETON, R. G. *Chemometrics: data Analysis for the laboratory and chemical plant*. New York: John Wiley & Sons, **2003.**

- [13] MARTENS, H., Naes T. *Multivariate Calibration*. John Wiley: New York, **1989**.
- [14] DE LATHAUWER, L., DE MOOR, B., VANDEWALLET, J. J. *Chemom.* **14: 123, 2000**.
- [15] GELADI, P., KOWALSKI, B. R. Partial Least-square: A tutorial. *Anal. Chim. Acat.* **185: 17, 1986**.
- [16] WOLD, S., et al. *Some recent developments in PLS modeling*. *Chemom Intell. Lab. Syst.* **58: 131, 2001**.
- [17] WOLD, S., STÖSTRÖM, M., ERIKSSON, L. *PLS-regression: a basic tool of chemometrics*. *Chemom Intell. Lab. Syst.* **58: 109, 2001**.
- [18] BORGES NETO, W. *Parâmetros de qualidade de lubrificantes e óleos de oliva através de espectroscopia vibracional, calibração multivariada e seleção de variáveis..* Tese de Doutorado. Campinas: UNICAMP, **2005**.
- [19] ANDERSSON, M. A comparison of nine PLS1 algorithms. *J. Chemom.* **23: 518, 2009**.
- [20] BOOKSH, K. S., KOWALSKI, B R. *Theory of Analytical Chemistry*. *Anal. Chem.* **66: 782, 1994**.
- [21] ALDERRAMA, P., BRAGA, J. W. B., POPPI, R. J. Estado da Arte das figuras de mérito em calibração multivariada. *Química Nova.* **32: 1278, 2009**.
- [22] Kennard, R. W.; Stone, L. A. *Computer Aided Design of Experiments*. *Technometrics*, **11: 137, 1969**.
- [23] GALVÃO, R. K. H., et al. *A method for calibration and validation subset partitioning*. *Talanta.* **67: 736, 2005**.
- [24] SNEE, R. D. *Validation of Regression Models: Methods and Examples*. *Technometrics*, **19: 415, 1977**.

- [25] NETO, B.B; SCARMINIO, I.S.; BRUNS, R.E. *Como fazer experimentos: pesquisa e desenvolvimento na ciência e na indústria*. 2. ed. Campinas: Editora da UNICAMP, 2003.
- [26] Annual Book of ASTM Standards; Standards practices for infrared, multivariate, quantitative analysis, E1655, vol 03.06, ASTM International: West Conshohocken, **2000**.
- [27] GONZALEZ, A. G., HERRADOR, M. A., ASUERO, A. G. *Intra-laboratory testing of method accuracy from recovery assays*. *Talanta*. **48: 729, 1999**.
- [28] HAWKINS, D. M. *The Problem of Overfitting*. *J. Chem. Inf. Comput. Sci* **44: 1, 2004**.
- [29] LIRA, L. F. B., et al. *Prediction of properties of diesel/biodiesel blends by infrared spectroscopy and multivariate calibration*. *Fuel* **89: 405, 2010**.
- [30] BARTHUS, R. C., POPPI, R. J. *Determination of the total unsaturation in vegetable oil by Fourier transform RAMAN spectroscopy and multivariate calibration*. *Vibrational Spectroscopy* **26: 99, 2001**.
- [31] ILLIAMS, P. e NORRIS, K. *Near-Infrared Technology in the Agricultural and Food Industries*. St. Paul, USA : Amer Assn of Cereal Chemists, **2001**.
- [32] SANTOS, E. O., et al. *Determination of degree of polymerization of insulating paper using near infrared spectroscopy and multivariate calibration*. *Vibrational Spectroscopy*. **52: 154, 2010**.
- [33] MOREIRA, E. D. T., et al. *Near infrared reflectance spectrometry classification of cigarettes using the successive projections algorithm for variable selection*. *Talanta*. **79: 1260, 2009**.
- [34] WU, D., et al. *Exploring Near and Midinfrared Spectroscopy to Predict Trace Iron and Zinc Contents in Powdered Milk*. *J. Agric. Food Chem.* **57: 1697, 2009**.

- [35] HENDL, O., et al. *A rapid and simple method for the determination of iodine values using derivative Fourier transform infrared measurements.* Anal. Chim. Acta **427: 75, 2001.**
- [36] WANG, L., et al. *Feasibility study of quantifying and discriminating soybean oil adulteration in camellia oils by attenuated total reflectance MIR and fiber optic diffuse reflectance NIR.* Food Chemistry **95: 529, 2006.**
- [37] SIMÕES, S. S. *Desenvolvimento de métodos validados para a determinação de captopril usando espectrometria NIRR e calibração multivariada.* Tese de doutorado, João Pessoa, UFPB, 2008.
- [38] BALABIM, R. M., SAFIEVA, R. Z. *Gasoline classification by source and type based on near infrared (NIR) spectroscopy data.* Fuel **87: 1096, 2008.**
- [39] ANDERSEN, C. M., BRO, R. *Variable selection in regression—a tutorial.* J. Chemom. Special Issue Article, **2010.**
- [40] FORINA, M., et al. *Selection of useful predictors in multivariate calibration.* Anal. Bioanal. Chem. **380: 397, 2004.**
- [41] GHASEMI, I., NIAZI, A., LEARDI, R. *Genetic-algorithm-based wavelength selection in multicomponent spectrophotometric determination by PLS: application on copper and zinc mixture.* Talanta **59: 311, 2003.**
- [42] SPIEGELMEN, C. H., et al. *Theoretical Justification of Wavelength Selection in PLS Calibration: Development of a New Algorithm.* Anal. Chem. **70: 35, 1998.**
- [43] HÖSKULDSSON, A. *Variable and subset selection in PLS regression.* Chemom. Int. Lab. Syst. **55: 23, 2001.**
- [44] FERREIRA, M. M. C., MONTANARI, C. A., GAUDIO, A. C. *Seleção de Variáveis em QSAR.* Química Nova **3: 439, 2002.**
- [45] COSTA FILHO, C. A., POPPI, R. J. *Algoritmo Genético em química.* Química Nova **22: 405, 1999.**

- [46] WEIJER, A. P., et al. *Using genetic algorithms for an artificial neural network model inversion*. Chemom. Int. Lab. Syst. **20: 45, 1993**.
- [47] LUCASIUS, C.B., KATEMAN, G. *Understanding and using genetic algorithms Part 1. Concepts, properties and context*. Chemom. Int. Lab. Syst. **19: 1, 1993**.
- [48] LEARDI, R., GONZÁLEZ, A. L. *Genetic algorithms applied to feature selection in PLS regression: how and when to use them*. Chemom. Int. Lab. Syst. **41: 195, 1998**.
- [49] LEARDI, R. *Application of genetic algorithm–PLS for feature selection in spectral data sets*. J. Chemom. **14: 643, 2000**.
- [50] CENTER, V., et al. *Elimination of Uninformative Variables for Multivariate Calibration*. Anal. Chem. **68: 3851, 1996**.
- [51] EFRON, B. *The Jack-knife, the bootstrap and other resampling plans*. Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics, **1982**.
- [52] MARTENS, H., MARTENS, M. *Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR)*. Food Quality and Preference **11: 5, 2000**.
- [53] SHAMSIPUR, M., et al. *Ant colony optimisation: a powerful tool for wavelength selection*. J. Chemom. **20: 146, 2006**.
- [54] ALLEGRINI, F., OLIVIERI, A. C. *A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy/partial least-squares analysis*. Anal. Chim. Acta **689: 18, 2011**.
- [55] NORGAARD, L., et al. *Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy*. Applied Spectroscopy. **54: 413, 2000**.

- [56] PIERNA, J. A. F., et al. *A Backward Variable Selection method for PLS regression (BVSPLS)*. Anal. Chim. Acta **642: 89, 2009**.
- [57] LEARDI, R. e NORGAARD, L. *Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions*. J. Chemometrics. **18: 486, 2004**.
- [58] NØRGAARD, L. *iToolbox Manual*, **2005**.
- [59] TEOFILO, R. F., *Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression*. J. Chemom. **23: 32, 2009**.
- [60] HAGEMAN, J. A., et al. *Wavelength selection with Tabu Search*. J. Chemom. **23: 32, 2009**.
- [61] FORINA, M., CASOLINO, C., MILLAN, C. P. *ITERATIVE PREDICTOR WEIGHTING (IPW) PLS: A TECHNIQUE FOR THE ELIMINATION OF USELESS PREDICTORS IN REGRESSION PROBLEMS*. J Chemom. **13: 165, 1999**.
- [62] ARAÚJO, M. C. U., et al. *The successive projections algorithm for variable selection in spectroscopic multicomponent analysis*. Chemometrics and Intelligent Laboratory Systems. **57: 65, 2001**.
- [63] GALVÃO, R. K. H., et al. *Aspects of the successive projections algorithm for variable selection in multivariate calibration applied to plasma emission spectrometry*. Analytica Chimica Acta. **443: 107, 2001**.
- [64] GALVÃO, R. K. H., et al. *Cross-Validation for the Selection of Spectral Variables Using the Successive Projections Algorithm*. Journal of Brazilian Chemical Society. **18: 1580, 2007**.
- [65] BREITKREITZ, M. C., et al. *Determination of total sulfur in diesel fuel employing NIR spectroscopy and multivariate calibration*. Analyst. **128: 1204, 2003**.

- [66] PEREIRA, A. F. C., et al. *NIR spectrometric determination of quality parameters in vegetable oils using iPLS and variable selection*. Food Research International. **41: 341, 2008**.
- [67] FERNANDES, D. D. S., et al. *Determination of biodiesel content in biodiesel/diesel blends using NIR and visible spectroscopy with variable selection*. Talanta **87: 30, 2011**.
- [68] PONTES, M. J. C., et al. *Determining the quality of insulating oils using near infrared spectroscopy and wavelength selection*. Microchemical journal **98: 254, 2011**.
- [69] DANTAS FILHO, H. A., et al. *Simultaneous Spectrometric Determination of Cu^{2+} , Mn^{2+} and Zn^{2+} in Polivitaminic/Polimineral Drug Using SPA and GA Algorithms for Variable Selection*. J. Braz. Chem. Soc. **16: 58, 2005**.
- [70] DI NEZIO, M. S., et al. *Successive projections algorithm improving the multivariate simultaneous direct spectrophotometric determination of five phenolic compounds in sea water*. Microchemical Journal. **85: 194, 2007**.
- [71] GOODARZI, M., FREITAS, M. P. e JENSEN, R. *Feature Selection and Linear/Nonlinear Regression Methods for the Accurate Prediction of Glycogen Synthase Kinase-3 Inhibitory Activities*. J. Chem. Inf. Model. **49: 824, 2009**.
- [72] GOUDARZI, N., et al. *QSPR Modeling of Soil Sorption Coefficients (K_{oc}) of Pesticides Using SPA-ANN and SPA-MLR*. J. Agric. Food Chem. **57: 7153, 2009**.
- [73] DANTAS FILHO, H. A., et al. *A strategy for selecting calibration samples for multivariate modelling*. Chemom. and Intell. Lab. Syst. **72: 83, 2004**.
- [74] PONTES, M. J. C., et al. *The successive projections algorithm for spectral variable selection in classification problems*. Chemometrics and Intelligent Laboratory Systems. **78: 11, 2005**.

[75] SOARES, S. F. C., et al. *A modification of the successive projections algorithm for spectral variable selection in the presence of unknown interferences*. Anal. Chim. Acta **689: 22, 2011**.

[76] GALVÃO, R. K. H, et al. *A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm*. Chemometrics and Intelligent Laboratory Systems. **92: 83, 2008**.

[77] SAVITZKY, A., GOLAY, M. J. E. *Smoothing and Differentiation of Data by Simplified Least Squares Procedures*. Anal. Chem. **36: 1627, 1964**.

[78] XIAOBO, Z., et al. *Variables selection methods in near-infrared spectroscopy*. Anal. Chim. Acta **667: 14, 2010**.

Apêndice

Código Principal

```
%ALGORITMO DAS PROJEÇÕES SUCESSIVAS APLICADO A SELEÇÃO DE VARIÁVEIS EM
REGRESSÃO PLS.
%LINHA DE COMANDO
%Modelo_iAPS_PLS=iAPS_PLS(Xcal,Ycal,Xpred,Ypred,intervalos,I_max,VL,xr
otulo)
% O Algoritmo iAPS_PLS realiação seleção de intervalos baseado no
criterio de % projeções do APS convencional acoplado a modelos de
regressão PLS
%Dados de entrada
%Xcal: Matriz (MxJ) das variáveis independentes do conjunto de
calibração.
%Ycal: Vetor coluna (Mx1) contendo o valor do parametro referencia do
conjunto decalibração
%Xpred: Matriz (NxJ) das variáveis independentes do conjunto de
predição.
%Ypred: Vetor coluna (Nx1) contendo o valor do parametro referencia do
conjunto decalibração
%intervalos: quantidade de intervalos que o espectro deve ser dividido
%I_max: número máximo de intervalos que devem ser selecionados
%VL: número de variáveis latentes que devem ser empregado no calculo
do %modelo PLS global
%xrotulo: entrada opcional, corresponde ao rotulo de X (comprimentos
de onda, por exemplo)
%%Versão 1.0 04/12/2011
%Desenvolvido por: Adriano de Araújo Gomes.
```

function

```
Modelo_iAPS_PLS=iAPS_PLS(Xcal,Ycal,Xpred,Ypred,intervalos,I_max,VL,xro
tulo)
clc
%%
intervals=intervalos;
N1=1;
N2=I_max;
%%
if nargin==7
    xrotulo=1:size(Xcal,2);
end
%
[Nmis_cal,Nlambdas]=size(Xcal); % determinando as dimensões de Xcal.

[VL_opt]=full_pls_cross_val(Xcal,Ycal,VL);%calculando PLS full
spectrum
disp(' ')
disp('número de variáveis latentes sugeridas '),VL_opt
Numero_de_VL_usada=input('Qual o número de variáveis latentes deve ser
usado? ');
VL=Numero_de_VL_usada;
%a=input('Qual o número de variáveis latentes deve ser usado? ');
%Particionando o espectro em i intervalos
[X]=partitioner(Xcal,intervals); %emprega uma função auxiliar
normas=[];
norm_max=[];
for i=1:size(X,2)
    a=X{1,i};
    x=Xcal(:,a);
    for j=1:size(x,2)
        b=norm(x(:,j));
        normas=[normas b];
    end
end
```

```

        [A index_norm_max]=max(normas);
    end

    norm_max=[norm_max index_norm_max];
end
iXcal=Xcal(:,norm_max);
% Aplicando o SPA
[L] = cadeias(iXcal,N1,N2); % cadeias é uma função auxiliar.
%%
[iNmis_cal,iNlambdas]=size(iXcal);
R = zeros(1,N2);
rmsep = [];
Lopt = zeros(N2,N2);
for N = N1:N2 % Para a cadeia de comprimento N
    for i = 1:iNlambdas % partindo da variavel i
        lambdas = L(1:N,i); % Variaveis da cadeia
        z=[];
        for b = 1:size(lambdas,1);
            g=lambdas(b,1);
            z=[z,X{1,g}];
        end
        % Respostas instrumentais nas variaveis da cadeia
        %intervals_selected=X{1,lambdas};
        Xcal2=Xcal(:,z);

        [index_RMSECV_min,RMSECV_min,RMSECV,ypred]=reg_pls(Xcal2,Ycal,VL);
        %regressão PLS
        rmsep(N,i)=RMSECV_min;% Salvar o erro associado ao número
        ótimo de VL!

    end
    [R(N) imin] = min(rmsep(N,:));
    N
    Lopt(1:N,N)=L(1:N,imin);
end

[Rbest,Nbest] = min(R(N1:N2));
Nbest = Nbest+N1-1;
rmsepopt = rmsep(Nbest,:);
l = (Lopt(1:Nbest,Nbest))'

%%
%% construção do modelo PLS para as "l" variáveis selecionadas e
previsão
clc
% redefinindo as matrizes espectras apenas as variáveis selecionadas.
h=[];
for f = 1:size(l,2)
    e=l(1,f);
    h=[h,X{1,e}];
end

Xcal2=Xcal(:,h);
%
[index_RMSECV_min,RMSECV_min,RMSECV,ypred]=reg_pls(Xcal2,Ycal,VL);
VL=index_RMSECV_min;
%Centralizando na média
my=mean(Ycal);
sy=1;
%amostras de calibração

```

```

Xcm=Xcal2-ones(Nmis_cal,1)*mean(Xcal2);
Ycm=Ycal-ones(Nmis_cal,1)*mean(Ycal);
%amostras de previsão externa
%Estimando o modelo PLS para as variáveis selecionadas
[wmat,umat,pmat,cmat,dvet,pnorm]=pls_r(Xcm,Ycm,VL);
[Ycal_est]=pls_p(Xcm,wmat,pnorm,pmat,cmat,dvet,VL,sy,my);
Ycal_est=Ycal_est(:,VL);
%Calculando as metricas de calibração
%1= Raiz do erros médio quadratico de calibração (RMSEC)
%RMSEC= sqrt((sumsqr(Ycal-ypr_cal))/Nmis_cal-(a+1))(equação de
referencia)
Elementos=size(Xcal,1);
RMSEC= sqrt((sumsqr(Ycal-Ycal_est))/Nmis_cal);
%2= Coeficiente de dorelação entre valor de referencia e valor
previsto
r=corrcoef(Ycal,Ycal_est);
r=r(2,1);
%3= Coeficinete de determinação
Rquadrado=r^2;
%Metricas de validação
Yval=[]; %Iniciando a matriz
[m n]=size(Xcal);% definindo as dimensões de Xcal
for k=1:m
    X=[Xcal2(1:k-1,:);Xcal2(k+1:m,:)];%cross-validation leve-one-out
em X
    Xcm=X-ones(m-1,1)*mean(X);%centrando na média
    x=Xcm;% definindo x para reg_ga
    Y=[Ycal(1:k-1,:);Ycal(k+1:m,:)];%cross-validation leve-one-out em
Y
    my=mean(Y); %calculando a média de Ycal para cada iteração do
processo de validação
    Ycm=Y-ones(m-1,1)*mean(Y); % cetrando Ycal mádia
    y=Ycm; % definindo y para reg_ga
    [wmat,umat,pmat,cmat,dvet,pnorm]=pls_r(x,y,VL); %calculo os
parametros de regrssão
    [C I]=setdiff(Xcal2,X,'rows');% selecionando a amostras removida
para validação
    x=C-mean(X);% centrando a iesima amostra de validação cruzada na
média do conjunto de amostras remecessentes
    [ypr]=pls_p(x,wmat,pnorm,pmat,cmat,dvet,VL,sy,my); %calculo de
Yprevisto
    Yval=[Yval; ypr]; %atualizando ypred a cada iteração do processo
de validação
end
Yval=Yval(:,VL);
RMSECV= sqrt((sumsqr(Ycal-Yval))/m)
r_val=corrcoef(Ycal,Yval);
r_val=r_val(2,1);
Rquadrado_val=r_val^2;
bias=abs((sum(Ycal-Yval))/m);

%Previsão para um conjunto externo de amostras
Xcal2=Xcal(:,h);
Xcm=Xcal2-ones(Nmis_cal,1)*mean(Xcal2);
Ycm=Ycal-ones(Nmis_cal,1)*mean(Ycal);
Xpred2=Xpred(:,h);
Xpred_cm=Xpred2-ones(size(Xpred2,1),1)*mean(Xcal2);
[wmat,umat,pmat,cmat,dvet,pnorm]=pls_r(Xcm,Ycm,VL);
my=mean(Ycal);
sy=1;
[Ypred_est]=pls_p(Xpred_cm,wmat,pnorm,pmat,cmat,dvet,VL,sy,my);

```

```

[Ypred_est]=pls_p(Xpred_cm,wmat,pnorm,pmat,cmat,dvet,VL,sy,my);
Ypred_est=Ypred_est(:,VL);
%
Elementos=size(Xpred,1);
RMSEP= sqrt((sumsqr(Ypred-Ypred_est))/size(Xpred,1))
r_pred=corrcoef(Ypred,Ypred_est);
r_pred=r_pred(2,1);
Rquadrado_pred=r_pred^2;
bias_pred=abs((sum(Ypred-Ypred_est))/size(Xpred,1));
%%
clc
%Saída
Modelo_iAPS_PLS.iAPS_PLS='iAPS_PLS';
Modelo_iAPS_PLS.intervalos_selecionados=1;
Modelo_iAPS_PLS.EC_CV='Parametros de Calibração e validação cruzada';
Modelo_iAPS_PLS.Elementos=size(Xcal,1);
Modelo_iAPS_PLS.pre_processamento = 'centralização na media';
Modelo_iAPS_PLS.RMSEC= RMSEC;
Modelo_iAPS_PLS.Ycal_estimado=Ycal_est;
Modelo_iAPS_PLS.r_corr=r;
Modelo_iAPS_PLS.R_quadrado=Rquadrado;
Modelo_iAPS_PLS.RMSECV= RMSECV;
Modelo_iAPS_PLS.Yval_cv=Yval;
Modelo_iAPS_PLS.r_corr_val= r_val;
Modelo_iAPS_PLS.R_quadrado_val=Rquadrado_val;
Modelo_iAPS_PLS.BIA_cv=bias;
Modelo_iAPS_PLS.variaveis_latentes_usadas_no_modelo=VL;
Modelo_iAPS_PLS.EP='Parametros de Predição';
Modelo_iAPS_PLS.RMSEP= RMSEP;
Modelo_iAPS_PLS.r_corr_pred= r_pred;
Modelo_iAPS_PLS.R_quadrado_pred=Rquadrado_pred;
Modelo_iAPS_PLS.BIAS_pred=bias_pred;
Modelo_iAPS_PLS.Ypred_estimado=Ypred_est;
%%
%Saída Grafica
%%
esp_m=mean(Xcal);
figure1 = figure('Color',[1 1 1]);
plot(xrotulo,esp_m);
title(' Intervalos selecionados pelo iAPS-PLS')
xlabel(' Canal analitico')
ylabel(' Sinal analitico')
hold on
plot(xrotulo(1,h),esp_m(1,h),'r*')
hold off
figure2 = figure('Color',[1 1 1]);
subplot(2,1,1),plot(Ycal,Ycal)
hold on
subplot(2,1,1),plot(Ycal,Ycal_est,'*r')
title(' Conjunto de calibração')
xlabel(' Valor de referencia')
ylabel(' Valor predito')
hold off
subplot(2,1,2),plot(Ycal,Ycal)
hold on
subplot(2,1,2),plot(Ycal,Yval,'*r')
title(' Validação cruzada')
xlabel(' Valor de referencia')
ylabel(' Valor predito')
figure3 = figure('Color',[1 1 1]);
plot(Ypred,Ypred)

```

```

title(' Conjunto de Predição')
xlabel(' Valor de referencia')
ylabel('Valor predito')
hold on
plot(Ypred,Ypred_est,'*r')

```

Função Auxiliar-1

```

function [J]=full_pls_cross_val(Xcal,Ycal,A)
% A função "full_pls_cross_val" calcula os fatores de um modelos
% PLS-cross-validation empregando algoritmo de NIPAS.
%Adriano de Araújo Gomes-LAQA (25/10/2011).
%%
RMSECV=[];%Iniciando a matriz
sy=1;% valor padrão para dados não autoescalonados
ypred=[]; %Iniciando a matriz
[m n]=size(Xcal);% definindo as dimensões de Xcal
if A>n %(limitando o número de PC's)
    A=n;
end
for k=1:m
    X=[Xcal(1:k-1,:);Xcal(k+1:m,:)];%cross-validation leve-one-out em
X
    Xcm=X-ones(m-1,1)*mean(X);%centrando na média
    x=Xcm;% definindo x para reg_ga
    Y=[Ycal(1:k-1,:);Ycal(k+1:m,:)];%cross-validation leve-one-out em
Y
    my=mean(Y); %calculando a média de Ycal para cada iteração do
processo de validação
    Ycm=Y-ones(m-1,1)*mean(Y); % cetrando Ycal média
    y=Ycm; % definindo y para reg_ga
    [wmat,umat,pmat,cmat,dvet,pnorm]=pls_r(x,y,A); %calculo os
parametros de regrssão
    [C I]=setdiff(Xcal,X,'rows');% selecionando a amostras removida
para validação
    x=C-mean(X);% centrando a iesima amostra de validação cruzada na
média do conjunto de amostras remecessentes
    [ypr]=pls_p(x,wmat,pnorm,pmat,cmat,dvet,A,sy,my); %calculo de
Yprevisto
    ypred=[ypred; ypr]; %atualizando ypred a cada iteração do processo
de validação

end
for i=1:A
    E= sqrt((sumsqr(ypred(:,i)-Ycal))/m);%calculando o RMSECV para
1,2,...,A variaveis latentes
    RMSECV=[RMSECV E]; %armazennado RMSECV
end
%%
%Determinação do Número Otimo de VL's.
alpha = 0.25;% significancia alpha = 0.25
fcrit = finv(1-alpha,m,m); %determinado o valor de fcritico.
(conferir os graus de liberdade)
rmsecvmax = min(RMSECV)*sqrt(fcrit);
disp(' ')
J= min(find(RMSECV < rmsecvmax)) ;
%%
%Saída grafica

```

```

figure1 = figure('Color',[1 1 1]);
ypred_pc_o=mean(Ycal)*ones(m,1);
rmsecv_pc_o=sqrt((sumsq(ypred_pc_o-Ycal))/m);
index=0:A;
erro_cv=[rmsecv_pc_o RMSECV];
bar(index,erro_cv)
title('RMSECV versus componentes PLS para o modelo global')
xlabel('Número de fatores incluídos no modelo')
ylabel('RMSECV')

```

Função auxiliar 2

```

function [X]=partitioner(Xcal,intervals)
% Função empregada para particionar as variáveis em I intervalos.
%%
m=size(Xcal,2);% determinando o número de variáveis
vars_left_over=mod(m,intervals);
N=fix(m/intervals);
startint=[(1:(N+1):(vars_left_over-1)*(N+1)+1)' ; ((vars_left_over-
1)*(N+1)+1+1+N:N:m)'];% Início de cada intervalo
endint=[startint(2:intervals)-1; m]; % Final de Intervalo
X=cell(1,intervals); %Iniciando X
for i=1:intervals
    x=startint(i,1):endint(i,1);
    X{i}=x;
end

```

Função auxiliar 3

```

% [wmat,umat,pmat,cmat,dvet,pnorm]=pls_r(x,y,A);
function [wmat,umat,pmat,cmat,dvet,pnorm]=pls_r(x,y,A);

[rx,cx]=size(x); %dimensão de X
[ry,cy]=size(y);%dimensão de Y
a=0;% Iniciando o contador do 1 while

while a<A; % obtenção de cada PC
    a=a+1;% implementando contador do 1 while
    u1=y(:,1); %definine u1 como primeira coluna de y
    w1=ones(cx,1);% inicia a matriz w1
    n=0; % iniciando o contador de 2 while
    scat=0;
    wesold=10^40;%?????
    while scat==0; % Obtenção dos parametros de cada vl
        n=n+1;%contador do segundo "while"
        w=(x'*u1)/(u1'*u1); % Iniciando os loadings de X
        w=w/norm(w); % normalizando os loadings de X
        wes=sum(abs(w-w1)); % critério de convergencia
        w1=w; % atualização dos loadings
        t=x*w; %calculando os scores de X
        c=(t'*y)/(t'*t);
        c=c';
        cnor=c/norm(c);% nomalizando o veotr de regressão
        u1=y*cnor;
        if (wes<cx*0.00000001)|(n>20&(abs(wes-wesold)<0.000001));
            scat=1;
        end
    end
end

```

```

        end;
        wesold=wes;
    end;
    p=(t'*x)/(t'*t);
    p=p';
    pnor=p/norm(p);
    tnew=t*norm(p);
    wnew=w*norm(p);
    d=(tnew'*ul)/(tnew'*tnew);
    x=x-t*p';
    y=y-tnew*d*cnor';
    wmat(:,a)=w;
    tmat(:,a)=tnew; %escores de X nrmalizados pelos loading de X
    pmat(:,a)=pnor;% loadings de X
    cmat(:,a)=cnor;
    dvet(a,1)=d;
    umat(:,a)=ul;
    pnorm(a,1)=norm(p);
end;

```

Função auxiliar 4

```

% [ypr]=pr_p(x,wmat,pnorm,pmat,cmat,dvet,A,sy,my);
%

function[ypr]=pls_p(x,wmat,pnorm,pmat,cmat,dvet,A,sy,my);

[rx,cx]=size(x);%determina as dimensões de X

tmat2=[]; %Inicia a matriz tmat2

for a=1:A %loop para todas as componentes principais
    t=x*wmat(:,a); %calculando os scores da iesima amosra de previsão
na PC a
    t=t*pnorm(a,1);
    x=x-t*(pmat(:,a))';% calculando os residuos de X
    tmat2(:,a)=t; %autalizando a matriz de escores das amostras de
previsão
end;

ypr=[];% Inicializa Y pr.
ypr1=((tmat2).*(ones(rx,1)*dvet(1:A)')).*(ones(rx,1)*cmat(1:A)); %
Estima Y
if A>1;
    ypr1=(cumsum(ypr1'))';% soma cumulativa
end;
ypr=ypr1*sy+my;%soma a média e multiplica pelo desvio padrão de Y

```