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 associado ao Kernel-PLS para calibração multivariada não linear

Valber Elias de Almeida

João Pessoa – PB - Brasil
Julho 2017

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 associado ao Kernel-PLS para calibração multivariada não linear

Valber Elias de Almeida*

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.

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

Co-Orientador: **Prof. Dr. Adriano de Araújo Gomes**

* Bolsista:

CAPES

João Pessoa – PB – Brasil

Julho 2017

A447a Almeida, Valber Elias de.
Algoritmo das projeções sucessivas associado ao Kernel-PLS para calibração multivariada não linear / Valber Elias de Almeida. - João Pessoa, 2017.
93 f.: il. -

Orientador: Mário César Ugulino de Araújo.
Coorientador: Adriano de Araújo Gomes.
Dissertação (Mestrado) - UFPB/CCEN

1. Química Analítica. 2. Calibração não linear. 3. Seleção de Variáveis . 4. Algoritmo das projeções sucessivas. 5. Dados Simulados. I. Titulo.

UFPB/BC CDU: 543(043)

Algoritmo das projeções sucessivas associado ao Kernel-PLS para calibração multivariada não linear.

Dissertação de Mestrado apresentada pelo aluno Valber Elias de Almeida e aprovada pela banca examinadora em 26 de julho de 2017.

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

Prof. Dr. José Germano Veras Neto
Examinador

Profa. Dra. Kátia Messias Bichinho
Examinadora

Dedicatória

A Deus que sempre me guiou por onde quer que eu andasse, e que em todos os desafios da vida se fez presente incondicionalmente, mostrando-me os caminhos por onde eu deveria seguir. E a todos os meus familiares e amigos que se fizeram importantes nesta conquista.

AGRADECIMENTOS

Em primeiro lugar dou graças a Deus por tudo que ele é...

Agradeço a meus pais (Valdir e Edjane) e meus irmãos (Evelyn e Wiliam) pelo que eles representam em minha vida, sempre foram meu porto seguro em todos os momentos, e tornaram toda caminhada até aqui muito mais fácil.

Agradeço a toda minha família por me darem condições de ser quem hoje eu sou, priorizando o amor e os verdadeiros valores da vida.

Agradeço a minha namorada Denise, pela dedicação amor e carinho que tem comigo ao longo de muitos anos.

Agradeço ao meu orientador Mario César, e co-orientador Adriano Araújo, pelos ensinamentos e conhecimentos passados durante o mestrado.

Aos meus orientadores em outras oportunidades, Germano Veras e Paulo Henrique, pela base que me deram para conseguir chegar até aqui.

Aos amigos do "Nosso Grupinho" pela grande amizade que se formou nesse tempo do mestrado e que com certeza se prolongará por toda vida.

A todos os amigos do LAQA pela amizade formada durante o tempo do mestrado.

Aos amigos do LQAQ que mesmo longe não deixaram de se fazer presentes.

Serei eternamente grato a David e Adenilton por tudo que fizeram por mim, não só durante e tempo do mestrado, mais de uma amizade/irmandade que já vem de outras épocas.

Agradeço ao PPGQ-UFPB e aos professores pelos ensinamentos passados durante o mestrado e curso das disciplinas.

Agradeço a CAPES pela concessão da bolsa de estudos.

RM 12:3-7

...Porque pela graça, que me é dada, digo a cada um dentre vós que não saiba mais do que convém saber, mais que saiba com temperança, conforme a medida da fé que Deus repartiu a cada um. Porque assim como em um corpo temos muitos membros, e nem todos os membros têm a mesma operação. Assim nós, que somos muitos, somos um só corpo em cristo, mas individualmente somos membros uns dos outros. De modo que tendo diferentes dons, segundo a graça que nos é dada, se é profecia, seja ela segundo a medida da fé, se é ministério, seja em ministrar; se é ensinar haja dedicação ao ensino...

RESUMO

Neste trabalho é relatado, pela primeira vez, o uso do Algoritmo das Projeções Sucessivas para a seleção de intervalos (*i*SPA) como etapa prévia a modelagem de dados não-lineares por meio *Kernel Partial Least Square* (Kernel-PLS). Esta nova abordagem, ou seja, iSPA-Kernel-PLS, é uma combinação entre a remoção de variáveis não informativas e ou redundantes, promovida pelo SPA e redução de ruído em dados não-lineares por Kernel-PLS. O desempenho do iSPA-Kernel-PLS foi avaliado nos seguintes estudos de caso em que a relação entre concentração e sinal analítico é não linear: dois bancos de dados simulados e um banco de dados envolvendo a quantificação de açúcares totais e grau brix em diferentes etapas do processo de produção de açúcar utilizando espectroscopia de infravermelho próximo em modo de transflectância. Quando comparados com o modelo *full* Kernel-PLS (espectro completo), o iSPA-Kernel-PLS apresentou melhores resultados em termos de RMSE, REP, R^2 e não foi verificado a presença de tendências (*bias*) significativas nas elipses de confiança. Portanto, os resultados obtidos mostram que o método proposto (iSPA-Kernel-PLS) é uma ferramenta útil na calibração não-linear.

Palavras-chave: calibração não linear, seleção de variáveis, algoritmo das projeções sucessivas, dados simulados.

ABSTRACT

In this work is reported, for the first time, the use of the Successive Projection Algorithm for interval selection (*i*SPA) combined to nonlinear data modeling by Kernel Partial Least Square (Kernel-PLS). This new approach, namely *i*SPA-Kernel-PLS, is a linkup between uninformative variable removed by SPA and noise reduction in nonlinear data by Kernel-PLS. The performance of the proposed method was evaluated in three cases of study: (i) two simulated data to quantitation of the analyte in which concentration-analytical signal relation is quadratic and (ii) sugar and brix quantitation in sugar cane process control at different steps using near infrared spectroscopy (NIR) in transflectance mode. The nonlinear relationship between sugar/brix and NIR intensities was confirmed by appropriate statistical tests. When compared with full model (full spectrum), the proposed methods showed better results in terms of RMSE, REP and R^2 for all case. In addition, significant *bias* is always absent in interval selection models based; this information is supported by analysis of elliptical joint confidence region. Therefore, the obtained results show that interval or variable selection, widespread in the linear calibration context, is a useful tool in nonlinear context too.

Keywords: nonlinear calibration, variable selection, successive projections algorithm, simulate data.

LISTA DE FIGURAS

Figura 2.1 - Organograma dos dados da matriz X e do vetor y para utilização na calibração multivariada.....25

Figura 2.2 - Esquema da divisão da matriz **X** e o vetor **y** nos subconjuntos de amostras de calibração ($\mathbf{X}_{cal}$, $\mathbf{y}_{cal}$), validação ($\mathbf{X}_{val}$, $\mathbf{y}_{val}$) e predição ($\mathbf{X}_{pred}$, $\mathbf{y}_{pred}$).....26

Figura 2.3 - Valores de observados *versus* valores preditos ajustados pelo método OLS (amostras, - reta de ajuste**)**.....32

Figura 2.4 - Elipse de confiança EJCR.....35

Figura 3.1 - Perfis gaussianos para cada constituinte da amostra referentes ao banco de dados 1 (analito, constituinte 1, constituinte 2, constituinte 3, constituinte 4, constituinte 5, sinal resultante).....39

Figura 3.2 - Perfis gaussianos para cada constituinte da amostra referentes ao banco de dados 2 (analito, constituinte 1, constituinte 2, constituinte 3, sinal resultante).....41

Figura 3.3 - Script Matlab contendo a rotina do algoritmo iSPA-kernel-PLS.....42

Figura 3.4 - Superfície de resposta correspondente a avaliação dos pares de valores de VL e σ (ponto escolhido PRESS =0,03; VL = 20; σ = 1,0).....44

Figura 3.5 - Relatório de saída do algoritmo iSPA-Kernel-PLS.....45

Figura 3.6 - Saídas gráficas do algoritmo iSPA-Kernel-PLS ((a) Intervalos selecionados; (b) Valores observados versus valores preditos; (c) elipse de confiança EJCR).....46

Figura 3.7 - Fluxograma do algoritmo iSPA-Kernel-PLS.....47

Figura 4.1 - Resposta dos dados simulados 1 das amostras de calibração e predição.....52

Figura 4.2 Otimização de VLs e σ para o *full* Kernel-PLS (ponto escolhido PRESS =0,02; VL = 22; σ = 1,0).....53

Figura 4.3 - Parâmetros de validação do modelo *full* Kernel-PLS: (a) reta de ajuste dos valores preditos versus observados por OLS; (b) elipse EJCR.....54

Figura 4.4 - Parâmetros de predição do modelo *full* Kernel-PLS: (a) reta de ajuste dos valores preditos versus observados por OLS; (b) elipse EJCR.....55

Figura 4.5 - RMSECV em função do número de variáveis dos dados simulados 1.....57

Figura 4.6 - Intervalos selecionados pelo iSPA-Kernel-PLS: para o modelo 7(2) sobre as respostas das amostras (a) e os perfis puros (b), e para o modelo 9(2) sobre as respostas das amostras(c) e os perfis puros (d) para o banco de dados 1.....58

Figura 4.7 - Superfícies de Resposta para otimização dos parâmetros de construção dos modelos: (a) → ponto escolhido para o modelo 7(2) PRESS =0,017; VL = 16; σ = 0,5; (b) → ponto escolhido para o modelo 9(2) PRESS =0,014; VL = 15; σ = 0,5.....59

Figura 4.8 - Parâmetros de predição do modelo iSPA-Kernel-PLS para as amostras de calibração por validação cruzada (reta de ajuste dos valores preditos *versus* observados por OLS para os modelos 7(2) (a) e 9(2) (c), elipse EJCR dos modelos 7(2) (b) e 9(2) (d)(▬ Kernel-PLS; ▬ iSPA-Kernel-PLS).....60

Figura 4.9 - Parâmetros de predição do modelo iSPA-Kernel-PLS para as amostras de calibração por validação cruzada (reta de ajuste dos valores preditos *versus* observados por OLS para os modelos 7(2) (a) e 9(2) (c), elipse EJCR dos modelos 7(2) (b) e 9(2) (d)(▬ Kernel-PLS; ▬ iSPA-Kernel-PLS).....62

Figura 4.10 - Respostas simuladas para o banco de dados 2.....63

Figura 4.11 - Otimização de VLS e σ para o *full* Kernel-PLS (→ ponto escolhido PRESS =0,010; VL = 17; σ = 1,8889).....64

Figura 4.12 - Parâmetros de validação do modelo *full* Kernel-PLS ((a)reta de ajuste dos valores preditos versus observados por OLS, (b) elipse EJCR) para o segundo banco de dados.....65

Figura 4.13 - Parâmetros de predição do modelo *full* Kernel-PLS (a)reta de ajuste dos valores preditos versus observados por OLS, (b) elipse EJCR) para o segundo banco de dados.....66

Figura 4.14 - RMSEV em função do número de variáveis para o segundo banco de dados.....68

Figura 4.15 - Intervalos selecionados pelo iSPA-Kernel-PLS para o modelo 7(3) sobre o os espectros das amostras (a) e os perfis puros (b).....69

Figura 4.16 - Superfícies de Resposta para otimização dos parâmetros de construção do modelo iSPA-Kernel-PLS para o segundo banco de dados.....70

Figura 4.17 - Parâmetros de predição do modelo iSPA-Kernel-PLS para as amostras de validação externa (reta de ajuste dos valores preditos *versus* observados por OLS (a) e elipse EJCR (b) (▬ Kernel-PLS; ▬ iSPA-Kernel-PLS)).....71

Figura 4.18 - Parâmetros de predição do modelo iSPA-Kernel-PLS para as amostras de predição (reta de ajuste dos valores preditos *versus* observados por OLS (a) e elipse EJCR (b) (▬ Kernel-PLS; ▬ iSPA-Kernel-PLS)).....73

Figura 4.19 - Espectros NIR para amostras obtidas na produção de açúcar.....74

Figura 4.20 - Parâmetros de validação do modelo Kernel-PLS para as amostras de validação externa (reta de ajuste dos valores preditos versus observados por OLS para grau brix (a) e açúcares totais (c), elipse EJCR para grau brix (b) e açúcares totais (d)).....76

Figura 4.21 - Parâmetros de predição do modelo Kernel-PLS (reta de ajuste dos valores preditos versus observados por OLS para grau brix (a) e açúcares totais (c), elipse EJCR para grau brix (b) e açúcares totais (d)).....77

Figura 4.22 - Intervalos selecionados pelo iSPA-Kernel-PLS para o modelo 4(2) grau brix (a) e o modelo 7(2) açúcares totais (b).....79

Figura 4.23 - Parâmetros de validação do modelo iSPA-Kernel-PLS (reta de ajuste dos valores preditos *versus* observados por OLS para grau brix (a) e açúcares totais (c), elipse EJCR para grau brix (b) e açúcares totais (d)(▬ Kernel-PLS; ▬ iSPA-Kernel-PLS))....80

Figura 4.24 - parâmetros de predição do modelo iSPA-Kernel-PLS (reta de ajuste dos valores preditos versus observados por OLS para grau brix (a) e açúcares totais (c), elipse EJCR para grau brix (b) e açúcares totais (d))(▬ Kernel-PLS; ▬ iSPA-Kernel-PLS))....81

Figura a1 - Fotografias de amostras em cada umas das etapas (a) moagem (suco), (b) evaporação (xarope), (c) cristalização (massa cozida) e (d) centrifugação (melaço)....92

LISTA DE TABELAS

Tabela 4.1 - Resultados da validação dos modelos iSPA-Kernel-PLS.....56

Tabela 4.2 - Resultado da determinação da concentração por iSPA-Kernel-PLS para as amostras de predição.....61

Tabela 4.3 - Resultados da validação dos modelos iSPA-Kernel-PLS para o banco de dados.....67

LISTA DE ABREVIATURAS E SIGLAS

σ	Largura da transformação gaussiana
ANN	Do inglês, *Artificial Neural Networks*
BP-ANN	Do inglês, *Backpropagation - Artificial Neural Networks*
DCLS	Do inglês, *Direct Classical Least Squares*
EJCR	Do inglês, *Elliptical Joint Confidence Region*
ICLS	Do inglês, *Indirect Classical Least Squares*
*i*SPA	Do inglês, *interval successive Projection Algorithm*
*i*SPA-Kernel-PLS	Do inglês, *interval successive Projection Algorithm* – Kernel – *Partial Least Squares*
Kernel-PLS	Do inglês, Kernel – *Partial Least Squares*
MLR	Do inglês, *Multiple Linear Regression*
NIPALS	Do inglês, *NonLinear Iterative Partial Least Squares*
NIR	Do inglês, *Near Infrared*
PCR	Do inglês, *Principal Component Analysis*
PLS	Do inglês, *Partial Least Squares*
PRESS	Do inglês, Predicted Residual Error Sum of Squares
R^2	Coeficiente de correlação
REP	Do inglês, Relative Error of Predictions

RMSEC	Do inglês, Root Mean Square Error of Calibration
RMSECV	Do inglês, Root Mean Square Error of Cross Validation
RMSEP	Do inglês, Root Mean Square Error of Prediction
RMSEV	Do inglês, Root Mean Square Error of Validation
SDV	Do inglês, Standard Deviation of validation
SPA	Do inglês, Successive Projections Algorithm
SPXY	Do inglês, Sample Set Partitioning Based on Joint X- and Y-blocks
SVM	Do inglês, Support Vector Machine
VL	Do inglês, Latent Variables

SUMÁRIO

1 INTRODUÇÃO.....19

1.1 CARACTERIZAÇÃO GERAL DO PROBLEMA.....19

1.2 OBJETIVO.....21

1.2.1 Objetivo geral.....21

1.2.1 Objetivos específicos.....21

2 FUNDAMENTAÇÃO TEÓRICA.....23

2.1 NOTAÇÃO CIENTÍFICA.....23

2.2 CALIBRAÇÃO MULTIVARIADA.....23

2.2.1 Organização dos dados.....24

2.2.2 Particionamento das amostras.....25

2.2.3 Regressão pelo Método dos Mínimos Quadrados Parciais.....26

2.2.4 Kernel – Mínimos Quarados Parciais.....29

2.2.5 Ferramentas de diagnóstico.....31

3 METODOLOGIA.....37

3.1 DADOS DO INFRAVERMELHO PRÓXIMO.....37

3.1.1 Procedimentos quimiométricos.....37

3.2 DADOS SIMULADOS.....38

3.2.1 Banco de dados 1.....38

3.2.1.1 Simulação das amostras.....38

3.2.1.2 Procedimento quimiométrico.....39

3.2.2 Banco de dados 2.....40

3.2.2.1 Aquisição das amostras simuladas.....40

3.2.2.2 Procedimentos quimiométricos.....42

3.3 CONSTRUÇAO DO ALGORÍTMO iSPA-KERNEL-PLS.....42

4 **RESULTADOS E DISCUSSÃO**.....48

4.1 *INTERVAL* ALGORITMO DAS PROJEÇÕES SUCESSIVAS – KERNEL - MÍNIMOS QUADRADOS PARCIAIS.....49
4.2 ANÁLISE DOS DADOS SIULADOS.....51
4.2.1 Banco de dados 1.....51
4.2.1.1 Kernel – Mínimos Quadrados Parciais.....51
4.2.1.2 *Interval* Algoritmo das Projeções Sucessivas – Kernel - Mínimos Quadrados Parciais.....56
4.2.2 Banco de dados 2.....62
4.2.2.1 Kernel – Mínimos Quadrados Parciais.....62
4.2.2.2 *Interval* Algoritmo das Projeções Sucessivas – Kernel - Mínimos Quadrados Parciais.....67
4.3 ANÁLISE DOS DADOS EXPERIMENTAIS.....74
4.3.1 Kernel – PLS.....75
4.3.2 *Interval* Algoritmo das Projeções Sucessivas – Kernel - Mínimos Quadrados Parciais.....78
5 CONCLUSÕES.....84
6 REFERÊNCIAS.....85
ANEXO 1.....92

Capítulo 1

Introdução

1 INTRODUÇÃO

1.1 CARACTERIZAÇÃO GERAL DO PROBLEMA

Com o avanço tecnológico e eletrônico, novas formas de obtenção de informações químicas, assim como a modernização das já existentes por meio de instrumentos, são constantemente desenvolvidas, fornecendo cada vez mais, uma maior quantidade de dados. Dentre as mais utilizadas destacam-se: técnicas espectrométricas de absorção molecular [1], técnicas espectrométricas de absorção atômica [2], técnicas cromatográficas [3], técnicas eletroanalíticas [4]. No entanto, devido à complexidade e dificuldade de avaliação dos dados, nem sempre estes são facilmente entendidos.

Visando uma melhor interpretação e simplificação destes dados, ferramentas quimiométricas podem ser utilizadas, entretanto a permanência de variáveis redundantes ou não informativas podem vir a prejudicar as modelagens. A utilização de técnicas de seleção de variáveis bem executada neste contexto pode ser de grande utilidade, pois possibilitam a busca de grupos de variáveis mais informativas, reduzindo o espaço dimensional sem prejuízo à construção dos modelos, além de aumentar consideravelmente a parcimônia [5 – 7].

A regressão pelo método dos mínimos quadrados parciais (PLS, do inglês *Partial Least Squares*) [8] tem sido largamente difundida, sendo considerada uma das técnicas mais estabelecidas e utilizadas para fins de regressão multivariada linear. Embora não seja obrigatoriamente necessária a seleção de variáveis, esta estratégia é comprovadamente benéfica, promovendo a obtenção de modelos melhor ajustados e mais objetivos [9 - 11]. A utilização de intervalos na regressão por PLS foi primeiramente implementada dividindo-se o conjunto total de variáveis em pequenos intervalos de variáveis e construindo modelos para cada um deles, sendo selecionado o intervalo com melhor desempenho, chamado iPLS [12]. Gomes e colaboradores em 2013 [13] e em

2014 [14] propuseram a utilização do algoritmo das projeções sucessivas (SPA) [15 -20] para a seleção de intervalos em modelagem PLS para dados de primeira e segunda ordem respectivamente.

Apesar da calibração multivariada linear apresentar um enorme potencial no âmbito da quantificação de analitos, alguns inconvenientes podem invalidar a aplicação de certas técnicas. Um dos principais inconvenientes é verificado quando a relação entre o sinal medido e a propriedade de interesse não se apresenta de forma linear. Para isso, formas alternativas de modelar esta não linearidade foram descritas na literatura, destacando-se dentre outras, redes neurais artificiais (ANN do inglês, *Artificial Neural Network*) [21], máquina de vetores de suporte (SVM) [22], processos de regressão Gaussiano [23] e Kernel – PLS [24 – 26].

A utilização da seleção de variáveis na calibração multivariada não linear é descrita na literatura com baixa frequência, embora alguns trabalhos tenham sido reportados. Benoudjit [27] e colaboradores propuseram selecionar variáveis para a calibração por ANN. Chen e colaboradores [28] propuseram um método de seleção de intervalos indireta, selecionando intervalos por siPLS e utilizando-os na calibração por BP-ANN (do inglês, *Backpropagation – Artificial Neural Network*). Tan e Li [29] sugeriram a seleção de intervalos pela estratégia de informação mútua para regressão por Kernel-PLS. Lee e colaboradores [30] propuseram "*multivariate feature selection*", em dados de espectroscopia no infra vermelho próximo (NIR), para regressão por Kernel-PLS.

1.2 OBJETIVOS

1.2.1 Objetivo geral

O presente trabalho tem como principal objetivo o desenvolvimento de um algoritmo, em ambiente Matlab, para seleção de intervalos, via algoritmo das projeções sucessivas (SPA), para subsequente modelagem *Kernel – Partial Least Squares* (Kernel-PLS) de dados em que em que a relação concentração e sinal analítico é não-linear.

1.2.2 Objetivos específicos

- Desenvolver o algoritmo, em linhas de comando, denominado *i*SPA-Kernel-PLS a fim de construir modelos de calibração multivariada não linear com seleção de intervalos de variáveis;
- Avaliar a capacidade preditiva do algoritmo proposto nos seguintes estudos de caso em que a relação concentração e sinal analítico é não linear: dois bancos de dados simulados e um banco de dados envolvendo a quantificação de açúcares totais e grau brix em diferentes etapas do processo de produção de açúcar utilizando espectroscopia de infravermelho próximo em modo de transflectância.
- Comparar os resultados obtidos pelo método proposto com o método *full* iSPA-Kernel-PLS.

Capítulo 2

Fundamentação teórica

2 FUNDAMENTAÇÃO TEÓRICA

2.1 NOTAÇÃO CIENTÍFICA

Daqui em diante, a notação a ser utilizada será: letras maiúsculas negritas para Matrizes, letras minúsculas negritas para vetores, letras em itálico para escalares e apostrofo “’” para matrizes ou vetores transpostos, A norma euclidiana de um vetor x é denotada por || x || e o acento circunflexo (^) é usado para indicar um valor estimado.

2.2 CALIBRAÇÃO MULTIVARIADA

De modo geral, o objetivo final de todo analista é obter informações sobre alguma propriedade física, química ou físico-química de uma dada amostra, seja esta resposta de natureza qualitativa ou quantitativa. Principalmente, tratando-se de concentração, esta pode não ser medida diretamente, assim o analista interessado em obter essa informação, precisa fazê-la indiretamente, correlacionando-a a uma outra propriedade mensurável da amostra [31].

A utilização da calibração univariada é muito importante tendo em vista a simplicidade matemática necessária para construir os modelos de regressão e, consequentemente, obter a informação desejada. Entretanto, qualquer interferente que venha a estar presente nas amostras pode inutilizar ou invalidar os modelos construídos, sendo necessário um esforço experimental para eliminar essa interferência [32], assim as técnicas de calibração multivariadas vem a superar essas deficiências.

A grande diferença entre as técnicas de calibração multivariada é a forma com que a correlação entre as respostas instrumentais **X** e o vetor de medidas **y** é feita, isto é, a forma de se calcular os coeficientes de regressão.

Inicialmente, os modelos de calibração multivariada foram propostos fazendo uma relação direta com a lei de Lambert-Beer [33], onde o sinal analítico é diretamente

proporcional à soma da concentração dos analitos. Estes são denominados métodos clássicos de calibração, e têm como principal característica a necessidade de se conhecer os perfis puros dos constituintes da amostra, sejam eles de forma direta (DCLS, do inglês *Direct Classical Least Square*) quando os perfis puros são medidos experimentalmente, ou indireta (ICLS, do inglês *Indirect Classical Least Square*) quando os perfis puros são obtidos a partir dos dados espectrais e da concentração [34].

A dificuldade de se conhecer sempre estes perfis puros torna a utilização desses métodos diretos bastante restrita. O desenvolvimento das técnicas de calibração inversa provocou um enorme avanço nas aplicações analíticas. Diferentemente dos métodos diretos, os inversos consideram que a concentração é função do sinal analítico, adotando uma forma "inversa" da lei de Lambert-Beer. Assim, não se faz mais necessário o conhecimento da contribuição do sinal de todas as espécies puras, uma vez que é utilizada na modelagem, a estrutura de variância/covariância da matriz **X** [35].

2.2.1 Organização dos dados

Tem-se claramente que quanto mais informação forem obtidas das amostras, mais facilmente um determinado fenômeno poderá ser entendido e quantificado. Nesse sentido, em casos mais complexos, é mais valioso utilizar um conjunto de medidas do que uma única. Assim possíveis variações ou interferências que possam afetar a medida podem ser detectadas e modeladas matematicamente por meio da calibração multivariada [32]. Os dados utilizados na calibração multivariada são organizados de acordo com a ilustração representada na Figura 2.1.

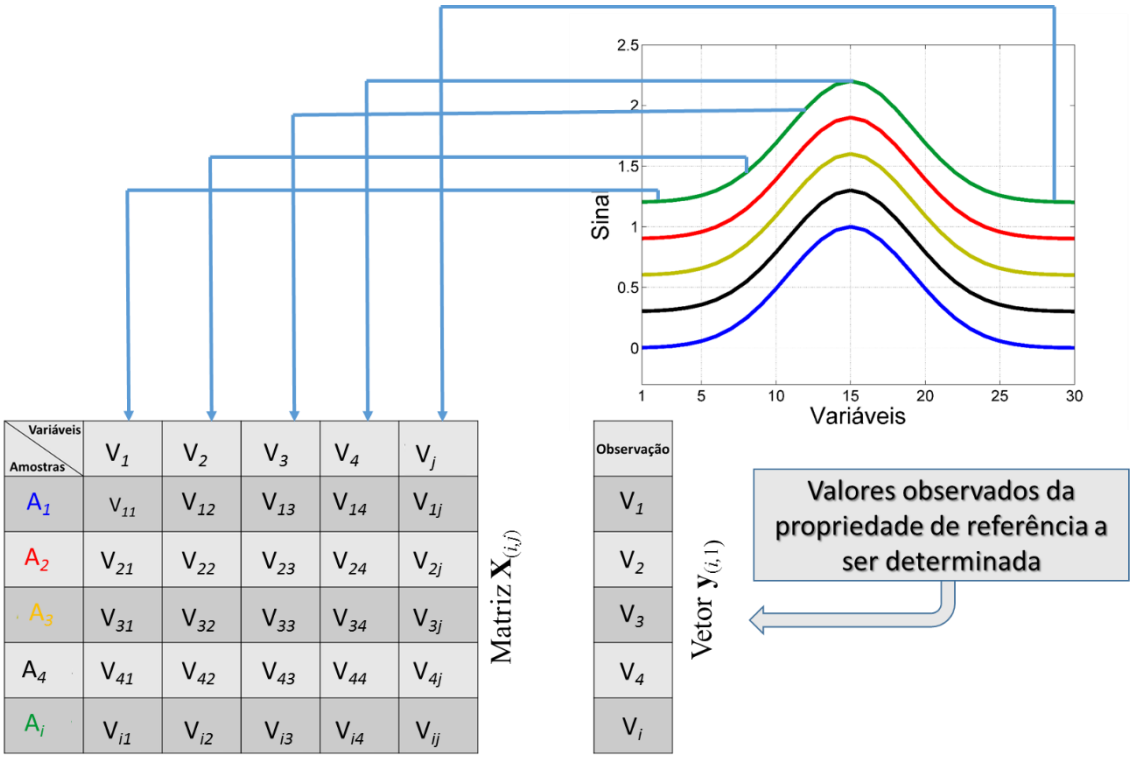

Figura 2.1 - Organograma dos dados da matriz **X** e do vetor **y** para utilização na calibração multivariada.

As informações correspondentes ás variáveis para cada amostra são alocadas lado a lado, formando um vetor linha de informações, por amostra, e cada amostra é colocada uma abaixo da outra perfazendo uma matriz de dados. Por convenção, a matriz que contém os dados instrumentais é denominada como matriz **X**, enquanto que a propriedade medida a ser determinada é denominada vetor **y**.

2.2.2 Particionamento das amostras

Na quimiometria como um todo, e mais especificamente na calibração multivariada, a escolha das amostras que serão utilizadas para a construção dos modelos é crucial para o desempenho destes, uma vez que a sua variabilidade é importante na explicação do fenômeno de estudo. Basicamente, o particionamento das amostras para a construção dos modelos em quimiometria é feito de acordo com a representação na Figura 2.2.

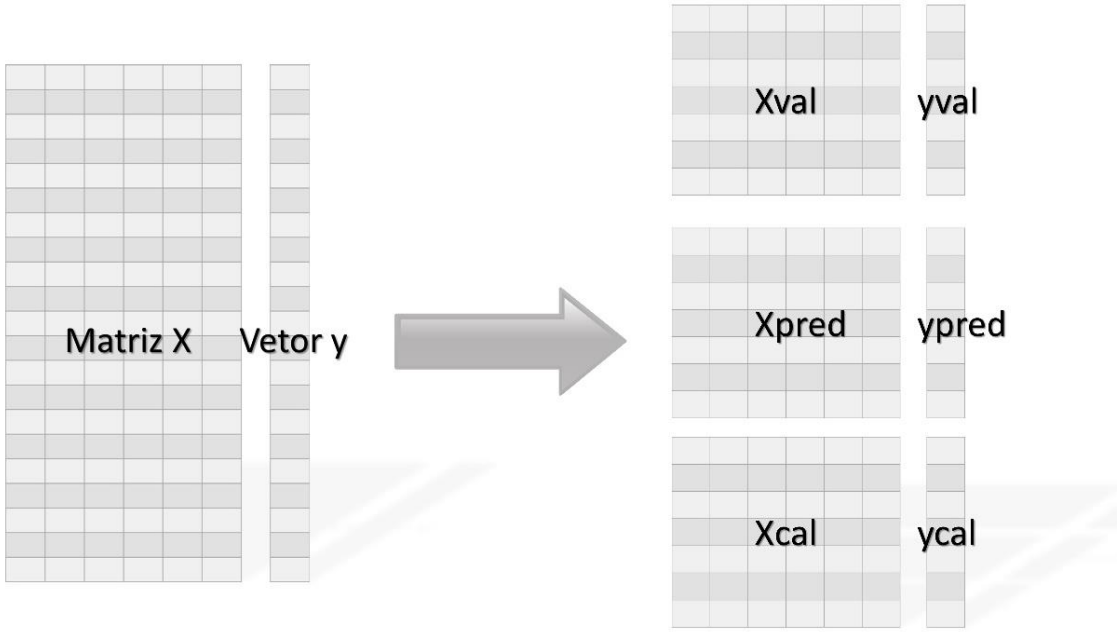

Figura 2.2 - Esquema da divisão da matriz **X** e o vetor **y** nos subconjuntos de amostras de calibração ($\mathbf{X}_{cal}$, $\mathbf{y}_{cal}$), validação ($\mathbf{X}_{val}$, $\mathbf{y}_{val}$) e predição ($\mathbf{X}_{pred}$, $\mathbf{y}_{pred}$).

Onde as amostras utilizadas para a construção do modelo são denominadas grupo ou conjunto de calibração; as amostras utilizadas para a validação do modelo são denominadas como grupo ou conjunto de validação e finalmente as amostras utilizadas para avaliar o desempenho final do modelo são denominadas como pertencentes ao grupo ou conjunto de predição. Geralmente para o particionamento das amostras em calibração, validação e teste, utiliza-se um algoritmo para seleção de amostras, tal como Kernnard Stones [36], Partição de amostras usando as matrizes **X** e **y**. (SPXY, do inglês *Sample Set Partitioning Based on Joint* **X**- *and* **Y**-*blocks* [37].

2.2.3 Regressão pelo Método dos Mínimos Quadrados Parciais

Proposto por Wold e colaboradores no ano de 1975 [38 - 39], a regressão PLS baseia-se na transformação das variáveis originais em variáveis latentes (VLs). Isto é feito por meio de uma decomposição da matriz de dados **X** em *scores* e *loadings* para cada nova variável latente, utilizando as informações da matriz **X** e do vetor **y** para PLS1 (uma única

propriedade) e a matriz **Y** para PLS2 (mais de uma propriedade), correlacionando os *scores* a fim de se obter a informação de interesse por regressão [38].

O método PLS é considerado uma extensão do PCR (do inglês, *Principal Components Regression*). No PCR, a decomposição da matriz de dados é feita visando à maximização da explicação da variância da matriz **X** em detrimento da redução da matriz de resíduos. Entretanto, no método PLS a decomposição busca a maximização da explicação da variância mais correlacionada com o vetor **y** ou matriz **Y**, em detrimento da redução da matriz de resíduos. Ou seja, os fatores obtidos no modelo PLS são otimizados a expressarem a maior parte da variância mais correlacionada com os dados independentes **Y** e no PCR os fatores expressam a maior parte da variância generalizada dos dados [40].

A forma de decomposição da matriz de dados nos componentes PLS é muito diversificada [41-42], embora a mais utilizada faz uso do algoritmo proposto por Wold e colaboradores [43] chamado algoritmo interativo não linear dos mínimos quadrados parciais (NIPALS do inglês, *Non-Linear Interative Partial Least Squares*).

O NIPALS obtém os componentes PLS de forma interativa, primeiramente decompondo as matrizes de dados instrumentais **X** e de respostas **Y**, em *scores* (**T** e **U** para **X** e **Y** respectivamente) e *loadings* (**P** e **Q** para **X** e **Y** respectivamente) por meio dos passos descritos a seguir:

$$\boldsymbol{X} = \boldsymbol{TP'} + \boldsymbol{E_x} = \sum \boldsymbol{t_k p'_k} + \boldsymbol{E_x} \qquad \text{Equação 1}$$

$$\boldsymbol{Y} = \boldsymbol{UQ'} + \boldsymbol{E_y} = \sum \boldsymbol{u_k q'_k} + \boldsymbol{E_y} \qquad \text{Equação 2}$$

Onde $\mathbf{E_x}$ é o resíduo de **X** e $\mathbf{E_y}$ o de **y**. Após obtidos os *scores* de **X,** $\boldsymbol{t_k}$e de **y**, $\boldsymbol{u_k}$, é obtida uma relação linear entre eles:

$$\boldsymbol{u_k} = \boldsymbol{b_k t_k} \qquad \text{Equação 3}$$

Onde **b** corresponde as estimativas dos coeficientes de regressão para *k* fatores, obtido pela Equação 4

$$\boldsymbol{b_k} = \frac{\boldsymbol{u_k'} * \boldsymbol{t_k}}{\boldsymbol{t_k'} * \boldsymbol{t_k}} \qquad \text{Equação 4}$$

Onde $\boldsymbol{t_k}$e $\boldsymbol{u_k}$ são os *scores* de **X** e **y** respectivamente. Os *loadings* obtidos são então normalizados para o comprimento unitário.

Agora a matriz **X** é decomposta nas levando-se em consideração a informação dos vetores **y** de acordo com os seguintes passos:

1. Faz-se $\boldsymbol{y} = \boldsymbol{u_k}$,

   Enquanto não houver convergência

2. Faz-se

   Estimam-se os *loadings* $\mathbf{W}_k$ de **X**:

$$\boldsymbol{W_k'} = \frac{\boldsymbol{u_k'} * \boldsymbol{X}}{\boldsymbol{u_k'} * \boldsymbol{u_k}} \qquad \text{Equação 5}$$

   Onde $\boldsymbol{u_k}$ são os *scores* de **y**. Os *loadings* são então normalizados de acordo com a equação 6 para o valor máximo 1.

$$\boldsymbol{W_{k,norm}'} = \frac{\boldsymbol{W_k'}}{\boldsymbol{norm}(\boldsymbol{W_k'})} \qquad \text{Equação 6}$$

   Os *scores* baseados nos *loadings* calculados na Equação 5 são expressos na forma da Equação 7

$$\boldsymbol{t_k} = \frac{\boldsymbol{X}\boldsymbol{W_{k,norm}}}{\boldsymbol{W_k'}\boldsymbol{W_k}} \qquad \text{Equação 7}$$

   Onde $\boldsymbol{W_{k,norm}}$.é o *loading* normalizado para o *k*-ésimo fator. Os *loadings* de **y** são obtidos de acordo com a Equação 8, e analogamente ao que é feito para os *loadings* de **X** são normalizados pela Equação 9:

$$\boldsymbol{q_k'} = \frac{\boldsymbol{t_k'} * \boldsymbol{Y}}{\boldsymbol{t_k'} * \boldsymbol{t_k}} \qquad \text{Equação 8}$$

$$\boldsymbol{q_{k,norm}'} = \frac{\boldsymbol{q_k'}}{\boldsymbol{norm}(\boldsymbol{q_k'})} \qquad \text{Equação 9}$$

Onde $\boldsymbol{t_k}$ são os *scores* de **X**, e $\boldsymbol{q_k}$ os *loadings* de **y**. Os *scores* de **y** são então obtidos:

$$\boldsymbol{u_k} = \frac{\boldsymbol{Yq_k}}{\boldsymbol{q_k'} * \boldsymbol{q_k}} \qquad \text{Equação 10}$$

Onde $\boldsymbol{q_k}$ corresponde aos *loadings* de **y**.

3. Compara-se então os *scores* $\boldsymbol{u_k}$ obtidos pela Equação 10 e pela Equação 3. Em caso de convergência, segue-se para o passo 4. Caso contrário, retorna-se ao passo 2 até a convergência.
4. Uma vez que os valores de **t** não são ortogonais, os valores de **p'** são substituídos por **w'** e um passo extra é realizado após a convergência visando ortogonalizar os valores de **t**.

$$\boldsymbol{p_k} = \frac{\boldsymbol{t_k'} * \boldsymbol{X}}{\boldsymbol{t_k'}\,\boldsymbol{t_k}} \qquad \text{Equação 11}$$

Onde $\boldsymbol{t_k}$ são os *scores* de **X**, e $\boldsymbol{p_k}$ são os *loadings* de **X** ortogonalizados.

5. A fração correspondente ao k-ésimo fator é eliminada subtraindo das matrizes originais **X** e **y** o produto dos respectivos *scores* e *loadings*:

$$\boldsymbol{X_k} = \boldsymbol{X_{k-1}} - \boldsymbol{t_k p_k} \qquad \text{Equação 12}$$

$$\boldsymbol{Y_k} = \boldsymbol{Y_{k-1}} - \boldsymbol{u_k q_k} \qquad \text{Equação 13}$$

Os passos de 2 a 5 são repetidos até que o limite *k* de fatores sejam calculados.

2.2.4 Kernel - Mínimos Quadrados Parciais

Apesar do PLS suportar certa não linearidade, quando esta se apresenta de forma muito evidenciada, chegando a prejudicar a capacidade preditiva do modelo, é necessária a correção desse inconveniente. Muitas técnicas de calibração multivariada não linear são reportadas na literatura, entretanto a técnica Kernel-PLS é bastante utilizada, devido

principalmente à simplicidade de operação e na geração de modelos relativamente simples.

A técnica Kernel-PLS é uma extensão da técnica PLS linear para um universo de modelagem não linear [44-45]. Consiste basicamente na utilização de uma etapa prévia de linearização dos dados e a utilização desses dados na calibração linear por PLS.

A matriz Kernel **K** de tamanho N_{cal} x N_{cal} é gerada a partir da projeção dos dados originais **X** em um espaço gaussiano não linear, representado pela expressão [44-45]:

$$\mathbf{K} = exp\ (-(||\mathbf{x}_{caln} - \mathbf{x}_{caln,\ ...,\ Ncal}||)^2/\sigma^2) \qquad \text{Equação 14}$$

Onde $\mathbf{x}_{caln}$ *é* o vetor correspondente à cada linha da matriz de calibração dos dados originais centrados na média, N_{cal} é o número total de amostras de calibração, e σ é o parâmetro de largura da transformação gaussiana.

Obtida a matriz **K** com os dados transformados e corrigidos os problemas relacionados aos desvios de linearidade, constroem-se os modelos PLS lineares entre **K** e o vetor **y** de respostas, obtendo seus respectivos coeficientes de regressão. Ainda relacionado à construção desses modelos PLS, é necessária a otimização do número de VLs visando evitar sobreajuste do modelo, e da largura σ. Para isso, García-Reiriz e colaboradores [45] propuseram realizar esta otimização por meio de uma validação cruzada *leave-one-out*.

Esta validação consiste em construir modelos utilizando o total de amostras de calibração menos uma, e utilizando esta deixada de fora para o teste do modelo, até que todas as amostras tenham sido deixadas para o teste ao menos uma vez. Este processo é repetido a cada par VL/σ, avaliando-o sempre em termos do PRESS (*Predicted Residual Error Sum of Squares*). Encontrado o ponto correspondente ao mínimo PRESS, avalia-se por meio de um teste estatístico, se a redução da quantidade de VLs promove um aumento

significativo do erro. Quando houver aumento significativo o número de VL anterior e o σ correspondente é escolhido.

2.2.5 Ferramentas de diagnóstico

A determinação do número *k* de fatores a serem incluídos no modelo de calibração multivariada é feita avaliando ferramentas de diagnóstico, que são medidas do comportamento, ajuste e da qualidade destes. Depois de construídos, a próxima etapa é a validação do modelo, que nada mais é do que a verificação da conformidade do seu funcionamento, para isso faz-se uso das amostras do conjunto de validação que podem ser tanto um conjunto externo quanto as próprias amostras de calibração[46].

Uma ferramenta bastante difundida na avaliação da qualidade de modelos de regressão é o método baseado no ajuste de uma reta entre os valores preditos e os valores observados pelo método OLS (Figura 2.3). É possível a partir desse método obter informações como coeficiente de correlação de *peasron (*r), coeficiente de determinação (R^2) e coeficientes angular (inclinação) e linear (intercepto) [47]. A flutuação das amostras em torno da reta de ajuste diz respeito a como o modelo foi capaz de predizer a informação de interesse, em detrimento ao valor real.

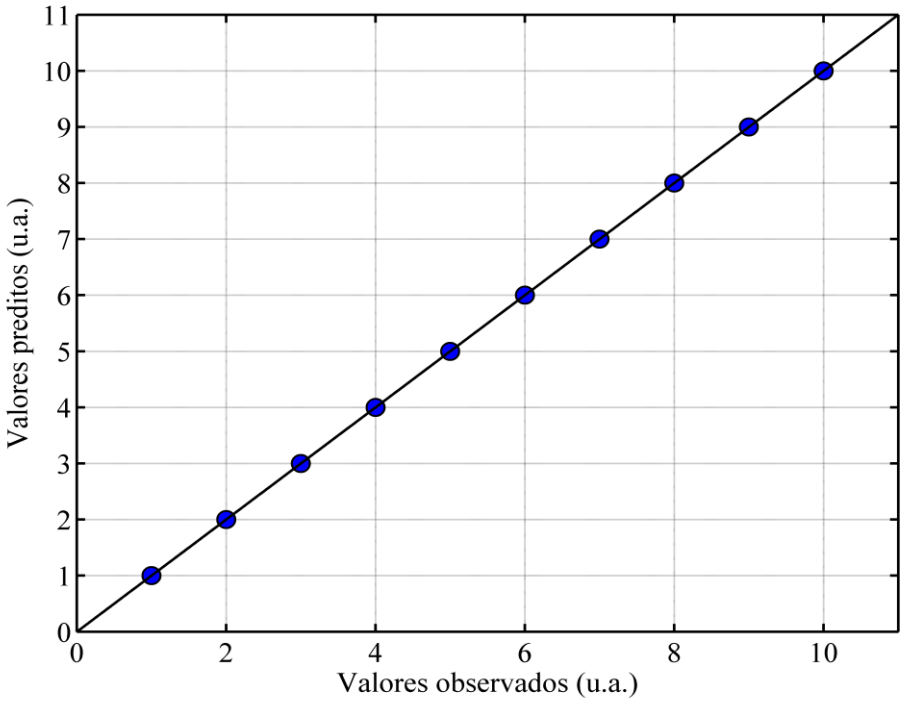

Figura 2.3 - Valores de observados *versus* valores preditos ajustados pelo método OLS (● amostras, - reta de ajuste**).**

Espera-se de forma ideal que o valor predito seja exatamente igual ao valor observado, nesse caso a reta deve cruzar o eixo de valores observados no valor zero (intercepto), e a reta de ajuste deve possuir inclinação igual a um, os coeficientes de determinação e correlação igual a 1 e todas as amostras devem estar localizadas sobre a reta de ajuste. Embora na prática isso não seja alcançado, quanto mais os valores obtidos se aproximarem destes ideais, maiores são os indícios que o modelo pode estar bem ajustado.

Basicamente, em calibração multivariada considera-se como base para os cálculos das métricas de desempenho o desvio (valor observado $(\boldsymbol{y_i}) -$ valor predito $(\hat{\boldsymbol{y}}_i)$), também chamado de resíduo. Esse valor demonstra o quanto o modelo foi capaz de predizer a concentração das amostras a partir dos dados instrumentais. Uma vez que em calibração multivariada utiliza-se uma grande quantidade de amostras, a avaliação dos resíduos um a um pode não ser uma tarefa trivial. Assim, a soma quadrática dos resíduos (*PRESS*, do inglês *Predicted Residual Error Sum of Squares*) mostra-se uma ferramenta

capaz de sumarizar a informação contida nos resíduos, assim como a raiz quadrada do erro médio quadrático de calibração (RMSEC, do inglês *Root Mean Square Error Of Calibration*), demonstrados nas equações 15 e 16 respectivamente [35].

$$\boldsymbol{PRESS} = \sum_{i=1}^{n}(y_{ci} - \hat{y}_{ci})^2 \qquad \text{Equação 15}$$

$$\boldsymbol{RMSEC} = \sqrt{\frac{PRESS}{(n_c - k - 1)}} \qquad \text{Equação 16}$$

Onde $\hat{y}_{ci}$ é o valor predito pelo modelo para o conjunto de calibração, y_{ci} é o valor observado para o conjunto de calibração, k é o número de fatores utilizados na construção do modelo, n_c é a quantidade de amostras contidas no conjunto de calibração e o valor 1 é o grau de liberdade perdido devido a centralização dos dados na média das colunas.

Uma característica do RMSEC é a utilização do termo correspondente ao número de fatores utilizados no modelo, fato que certamente influencia nos resultados, podendo ocasionar subajuste por número insuficiente de fatores, ou sobreajuste, por excesso de fatores [48].

Como forma de evitar prováveis erros na determinação do número de fatores dos modelos considerando o RMSEC, pode-se fazer uso da raiz quadrada do erro médio quadrático de validação cruzada (RMSECV, do inglês *Root Mean Square Error of Cross Validation*) e da raiz quadrada do erro médio quadrático de validação (RMSEV, do inglês *Root Mean Square Error of Cross Validation*). De forma similar a estes, a raiz quadrada do erro médio quadrático de predição (RMSEP, do inglês *Root Mean Square Error of Prediction*) também pode ser calculado, obtendo-se assim uma estimativa do erro associado à utilização do modelo para prever as propriedades das amostras de predição. Estas estimativas são apresentadas a seguir como Equação 17.

$$\boldsymbol{RMSEV, RMSECV\ e\ RMSEP} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{m}} \qquad \text{Equação 17}$$

Onde $\hat{y}_i$ é o valor predito pelo modelo para o conjunto de validação cruzada, validação externa e predição respectivamente, y_i é o valor observado para os conjuntos de calibração, validação externa e predição respectivamente, *m* é a quantidade de amostras de calibração, validação e predição respectivamente. Uma fração do RMSECV, RMSEV e RMSEP faz referência ao erro sistemático (*bias*) Equação 18, expresso em termos da razão do somatório dos desvios pela quantidade de amostra em cada um dos conjuntos respectivamente [35].

$$\boldsymbol{bias} = \frac{\sum(y_i - \hat{y}_i)}{m} \qquad \text{Equação 18}$$

Onde y_i corresponde aos valores observados, $\hat{y}_i$ os valores preditos e *m* o número de amostras. Idealmente, o somatório dos desvios sejam sempre zero, fato que corresponde à situação onde os valores preditos são exatamente iguais aos valores observados. Porém, devido a flutuações aleatórias nos dados e erros experimentais, dificilmente este panorama é obtido. Quanto mais esse valor se aproximar de zero, mais o modelo construído mostra-se bem ajustado. Segundo a norma ASTM E1655-00 [49], um modelo bem ajustado é aquele que não possui *bias* significativo, uma forma de verificar essa significância, é por meio do teste – *t* para amostras de validação externa a 95% de confiança. Para o cálculo de *t* é necessário a obtenção do valor do desvio padrão dos erros de validação (SDV, do inglês *Standard Deviation of Validation*) utilizando a Equação 19.

$$\boldsymbol{SDV} = \sqrt{\frac{\sum[(y_i - \hat{y}_i) - \boldsymbol{bias}]^2}{(m_v - 1)}} \qquad \text{Equação 19}$$

O cálculo de *t* é então realizado por meio da Equação 20. O valor de t obtido é comparado com o valor crítico tabelado de acordo com o grau de liberdade associado. Se o valor calculado for maior que o crítico significa que o modelo pode possuir erro sistemático.

$$t_{bias} = \frac{|bias|\sqrt{m_v}}{SDV} \qquad \text{Equação 20}$$

Outra forma auxiliar de verificar a conformidade dos modelos de regressão é utilizando a região elíptica de confiança conjunta Figura 2.4 (EJCR, do inglês *Elliptical Joint Confidence Region*). A partir da reta de ajuste obtida pelo método OLS, estimam-se os intervalos de confiança conjunta dos coeficientes angulares e lineares para cada modelo [50-51].

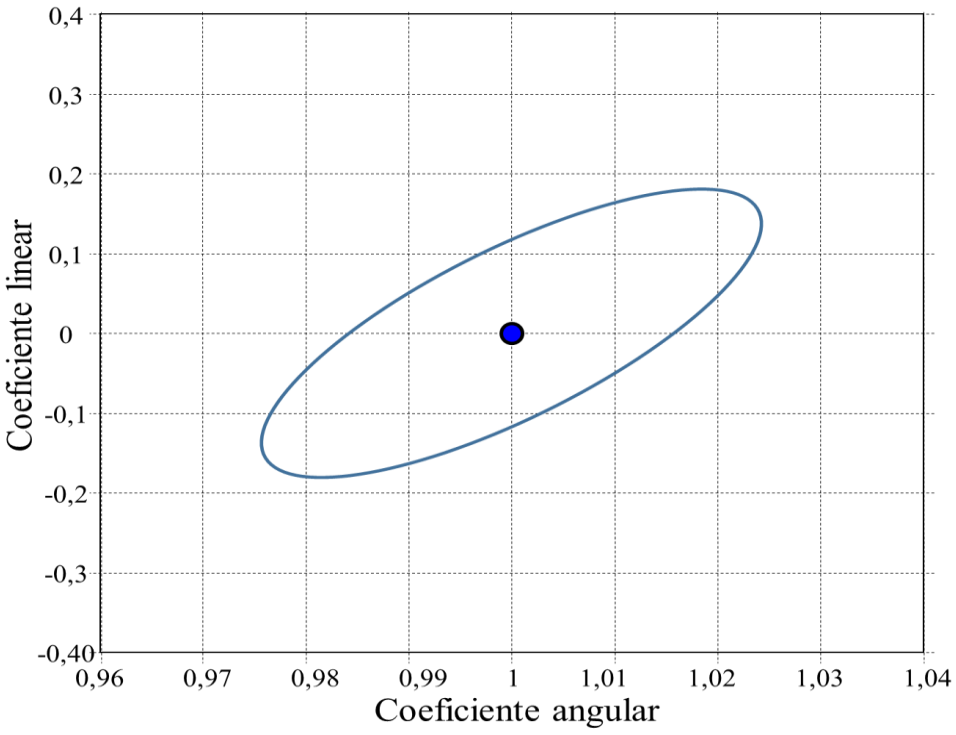

Figura 2.4 - Elipse de confiança EJCR.

Capítulo 3

Metodologia

3 METODOLOGIA

Neste trabalho foram utilizados dois bancos de dados simulados distintos para a verificação do funcionamento do algoritmo em diferentes condições. Em um primeiro caso os constituintes estão sobrepostos ao analito e o analito possui uma única banda de resposta. No segundo banco de dados simulados os constituintes estão parcialmente sobrepostos ao analito, e o analito possui duas bandas de resposta. E em seguida, o algoritmo foi testado em um sistema real, a partir de dados disponibilizados na internet, para a quantificação de dois analitos.

3.1 DADOS DO INFRAVERMELHO PRÓXIMO

Em busca da avaliação do desempenho do algoritmo proposto em um sistema real, foi utilizado um banco de dados disponível ao público eletronicamente (http://www.models.life.ku.dk/nirsugarcanedata), constituído por 1.797 amostras medidas utilizando espetroscopia NIR na faixa entre 400 e 1.888 nm com resolução espectral de 2 nm, perfazendo 745 variáveis. Os dados de referência medidos foram teor de açúcares totais e grau brix. Para mais detalhes a respeito do banco de dados consultar Anexo 1.

3.1.1 Procedimentos quimiométricos

As amostras foram divididas utilizando-se o algoritmo SPXY [34] em calibração (1.000 amostras), validação (397 amostras) e predição (400 amostras). O iSPA-Kernel-PLS foi avaliado para quantificação dos dois analitos, dividindo as variáveis em W intervalos de 1 a 10, número de VLs variando de 1 a 45 para grau brix e de 1 a 50 para teor de açúcares totais e a largura da transformação gaussiana σ variando de 0,5 a 3,0.

3.2 DADOS SIMULADOS

3.2.1 Banco de dados 1

O primeiro banco de dados simulados foi elaborado visando a simulação de um sistema para a quantificação da concentração de um analito cuja concentração é correlacionada não linearmente com o sinal. Este sistema simula um cenário que pode vir a ocorrer em espectros de absorção molecular, com sobreposições totais e parciais.

3.2.1.1 Simulação das amostras

O espectro simulado foi construído como uma soma do perfil de seis gaussianas, das quais uma é referente ao analito, contendo uma única banda, e as outras cinco correspondentes aos constituintes contidos na amostra. Os perfis das gaussianas foram obtidos de acordo com a Equação 21.

$$f(x) = ae^{\frac{-(x-b)^2}{c}} \qquad \text{Equação 21}$$

Onde a, b e c são constantes arbitrárias relacionadas à altura, centro e largura de um perfil gaussiano, respectivamente. Para cada perfil gaussiano f(x), foram atribuídos valores de “x“ na faixa de 1 a 100 com intervalos de 0.1, perfazendo um total de 991 variáveis. Os perfis gaussianos para cada constituinte são apresentados na Figura 3.1.

A escolha dos constituintes e sua posição no espectro foi feita para ajudar a entender como o algoritmo poderá responder num sistema real. Encontra-se dois constituintes parcialmente sobrepostos (verde e amarelo) ao analito (azul) e um totalmente sobreposto (ciano). Além disso, foi adicionada uma região onde não se verifica contribuições de analito, apenas de dois constituintes (vermelho e magenta), região que não deve ser selecionada pelo algoritmo, uma vez que não traz informação útil na construção do modelo.

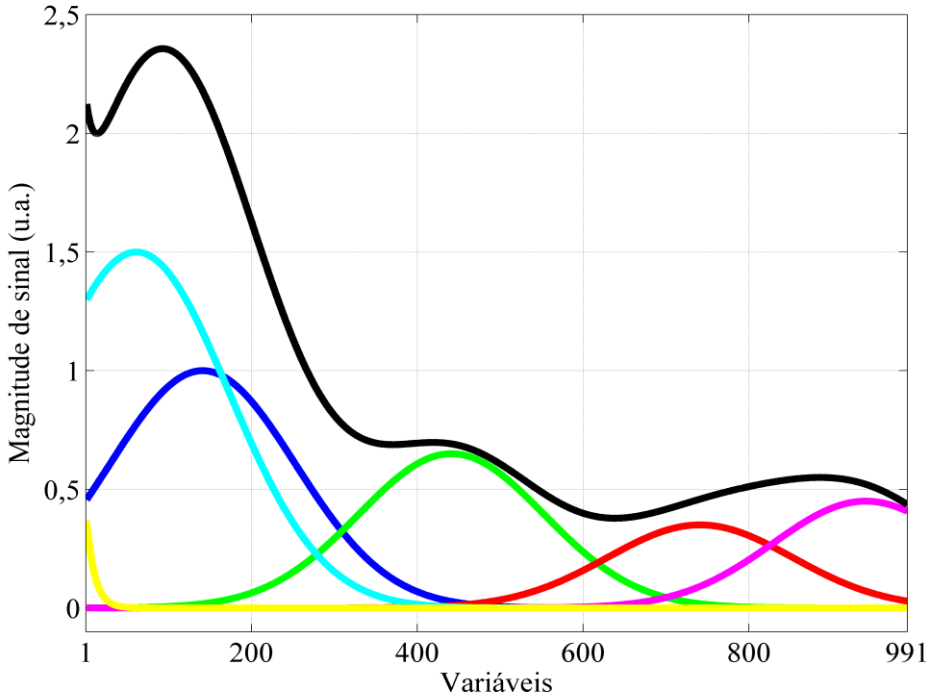

Figura 3.1 - Perfis gaussianos para cada constituinte da amostra referentes ao banco de dados 1 (▬ analito, ▬ constituinte 1, ▬ constituinte 2, ▬ constituinte 3, ▬ constituinte 4, ▬ constituinte 5, ▬ sinal resultante).

O conjunto de calibração $\mathbf{X_{cal}}$ compreende 100 amostras, em que a concentração do analito varia de 0,0053 a 0,9976 unidades arbitrárias. Um conjunto de predição independente $\mathbf{X_{pred}}$, contendo 50 amostras, foi construído de forma semelhante na faixa de concentração de 0,0170 a 0,9293 unidades. Para todos os casos, o ruído atribuído ao sinal e à concentração foi de 1% do sinal máximo.

3.2.1.2 Procedimentos quimiométricos

O iSPA-Kernel-PLS foi avaliado para quantificação do analito dividindo-se as variáveis em W intervalos variando de 1 a 10, número de VL variando de 1 a 45, e a o σ variando de 0,5 a 2,0.

3.2.2 Banco de dados 2

O segundo banco de dados simulados foi elaborado visando a simulação de um sistema para a quantificação da concentração de um analito hipotético em que sua concentração é correlacionada não linearmente com a resposta. Este sistema simula um caso em que os demais constituintes se sobrepõem parcialmente com o analito.

3.2.2.1 Aquisição das amostras simuladas

A simulação das amostras para o segundo banco de dados foi feita de forma similar à que foi feito para o primeiro, diferindo apenas na posição e quantidade de bandas de sinal dos perfis do analito e dos constituintes. Neste caso particular, as amostras são formadas por um analito e três outros constituintes, para mimetizar a matriz, onde o analito (azul) possui duas bandas e cada um dos constituintes, se sobrepõe parcialmente a cada uma das bandas do analito. Dois constituintes possuem duas bandas (vermelho e ciano) e um possui uma banda (verde) (Figura 3.2). A posição das bandas correspondentes a cada constituinte foi escolhida de modo que haja regiões de sobreposição entre o analito e os constituintes, e regiões onde as bandas correspondentes ao analito estejam livres. O sinal do analito foi simulado, de modo que este apresente duas bandas intercaladas, sobrepostas parcialmente a outros dois constituintes.

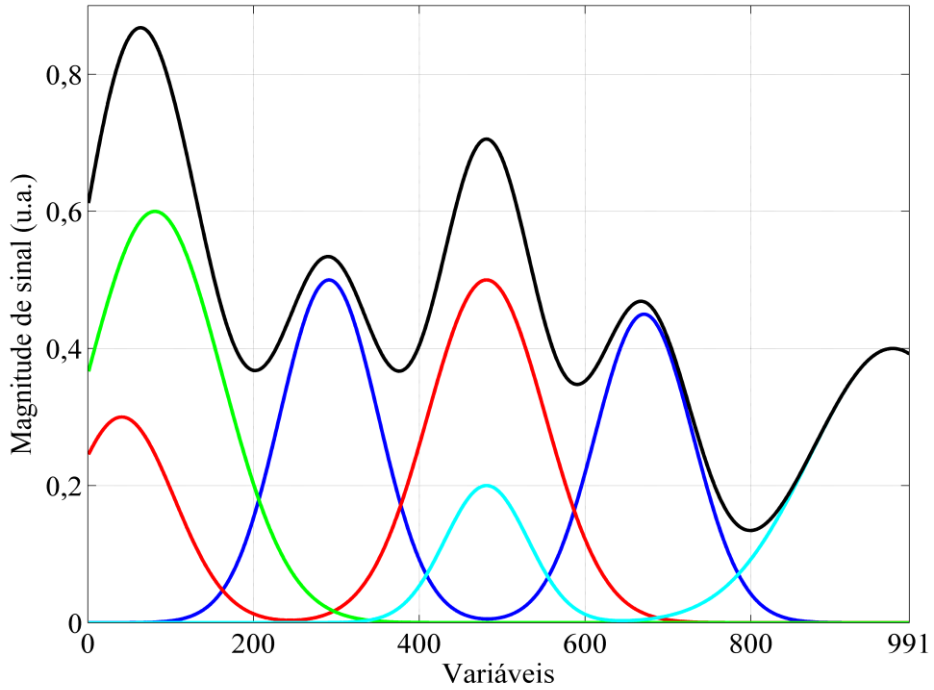

Figura 3.2 - Perfis gaussianos para cada constituinte da amostra referentes ao banco de dados 2 (analito, constituinte 1, constituinte 2, constituinte 3, sinal resultante).

Espera-se, nesse caso, que o algoritmo seja capaz de evitar as regiões onde não são observadas informações relevantes referente ao analito, selecionando as regiões intercaladas que contêm o sinal do analito, ou seja, em torno das variáveis 300 e 700.

O conjunto de calibração $\mathbf{X}_{cal}$ compreende 100 amostras, em que a concentração do analito varia de 0,0258 a 1,9537 unidades aleatoriamente. Um conjunto de validação independente $\mathbf{X}_{val}$, com 50 amostras, foi construído de forma semelhante na faixa de concentração de 0,3478 a 1,6570 unidades e um conjunto de predição independente $\mathbf{X}_{pred}$, com 50 amostras, na faixa de concentração de 0,4223 a 1,5436 unidades. Para todos os casos, o ruído atribuído ao sinal e a concentração foi de 1% do sinal máximo.

3.2.2.2 Procedimentos quimiométricos

O iSPA-Kernel-PLS foi avaliado para quantificação do analito dividindo as variáveis em W intervalos variando de 1 a 10, número de VLs variando de 1 a 30, e a largura da transformação gaussiana σ variando de 0,5 a 3,0.

3.3 CONSTRUÇÃO DO ALGORÍTMO iSPA-KERNEL-PLS

O desenvolvimento do algoritmo e todo o tratamento dos dados foram feitos em ambiente Matlab® 2010b. O algoritmo desenvolvido foi feito em forma de linhas de comando em arquivos do Matlab em arquivos de extensão ".m", como demonstrado na figura 3.3.

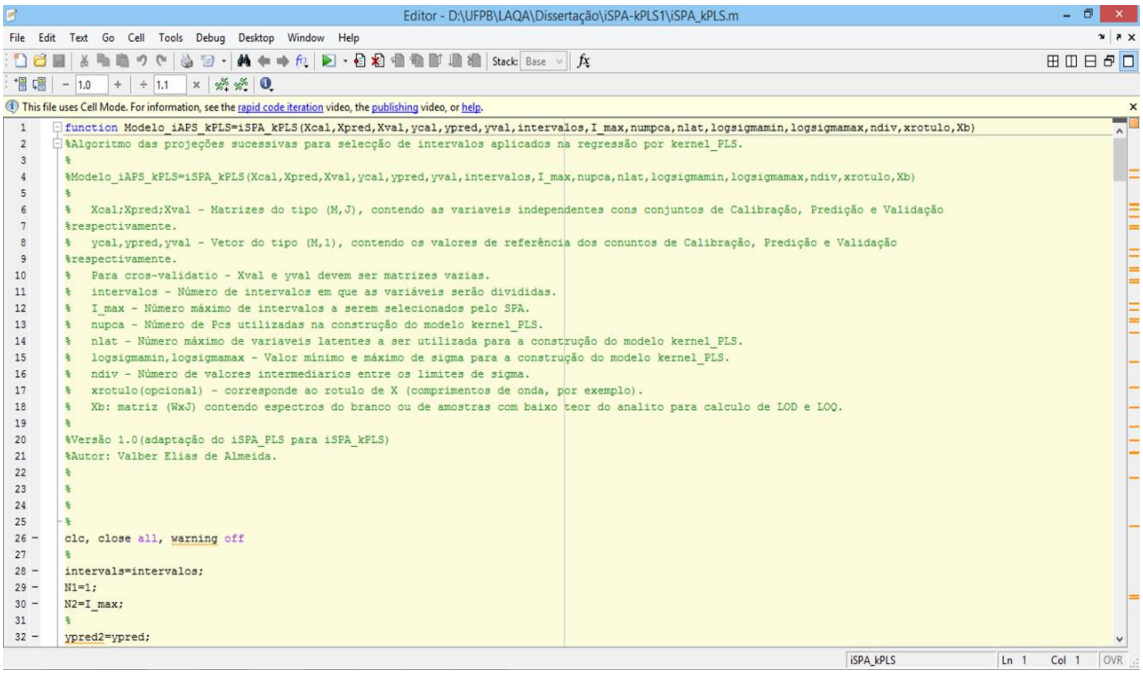

Figura 3.3 - Script Matlab contendo a rotina do algoritmo iSPA-Kernel-PLS.

Arquivos de entrada são necessários para a utilização do algoritmo, conforme descritos a seguir:

- $\mathbf{X_{cal}}$; $\mathbf{X_{pred}}$; $\mathbf{X_{val}}$: matrizes contendo as variáveis independentes dos conjuntos de calibração, predição e validação respectivamente.

- $\mathbf{y_{cal}}$; $\mathbf{y_{pred}}$; $\mathbf{y_{val}}$: vetores contendo os valores observados referentes aos parâmetros a serem determinados dos conjuntos de calibração, predição e validação respectivamente.
- Intervalos: número de intervalos em que as variáveis serão divididas.
- I_max: número máximo de intervalos a serem selecionados pelo SPA.
- nupca: número de componentes na qual os dados serão truncados.
- nlat: número máximo de VLs a serem utilizadas para a construção do modelo kernel PLS.
- logsigmamin; logsigmamax: valor mínimo e máximo de σ para a construção do modelo Kernel PLS.
- Ndiv: número de valores intermediários entre os limites de sigma.
- xrotulo (opcional): corresponde ao rótulo de X (comprimentos de onda, por exemplo).
- Xb (opcional): matriz contendo espectros do branco ou de amostras com baixo teor do analito para cálculo de LOD (do inglês, *Limit of Detection*) e LOQ (do inglês, *Limit of Quantification*).

O algoritmo comporta a construção de modelos utilizando validação cruzada do tipo *full cross validation* ou validação externa. No primeiro caso a matriz **Xval** e o vetor **yval** devem ser matrizes vazias. Inicialmente, o algoritmo realiza uma varredura com objetivo de verificar a quantidade de VLs e o valor de σ, computando um modelo *full* Kernel-PLS para cada par de valores de VL e σ indicado pelo usuário. O par correspondente ao modelo que apresentar o menor PRESS será utilizado na construção dos modelos a serem testados na fase 2 do iSPA.

Após a avaliação dos valores de VL e σ, a saída gráfica como a apresentada na Figura 3.4 é exibida.

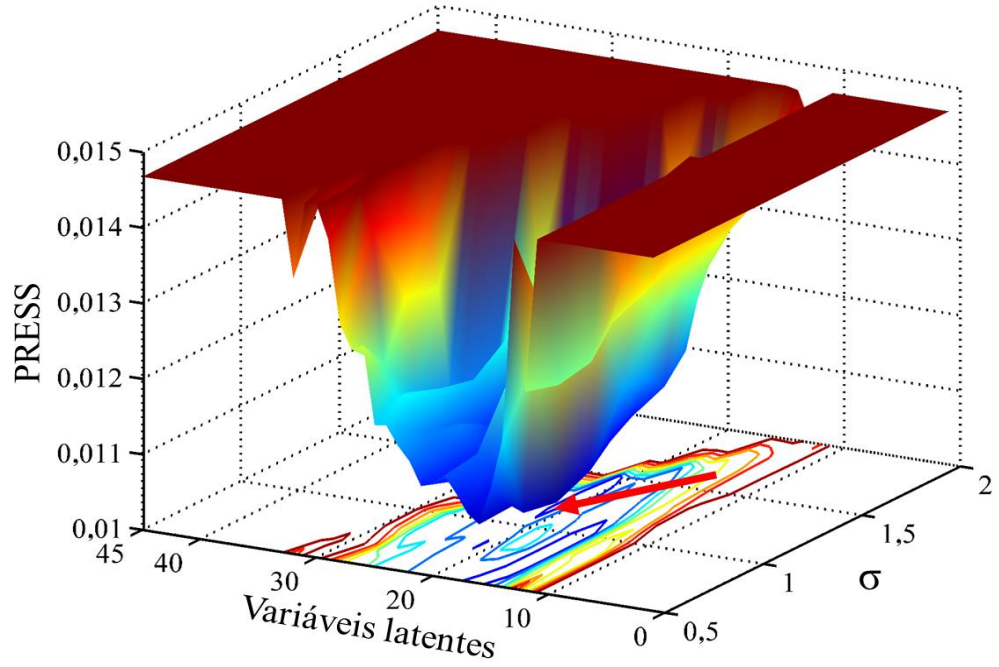

Figura 3.4 - Superfície de resposta correspondente a avaliação dos pares de valores de VL e σ (→ ponto escolhido PRESS =0,03; VL = 20; σ = 1,0).

O critério de escolha dos valores de VL e σ são baseados no princípio da parcimônia, ou seja, a menor quantidade de VLs necessárias para minimizar o erro associado. Ao se encontrar o mínimo, é verificado por meio de um teste F a 75% de confiança estatística se a redução do número VLs ocasiona um aumento significativo do PRESS, caso a afirmação não se confirme o processo é repetido até que o aumento do PRESS seja significativo.

Uma vez determinados estes valores iniciais, a fase 2 do SPA é então realizada. Após a avaliação das cadeias, é apresentado na janela de comandos do Matlab o relatório de saída do algoritmo (Figura 3.5) contendo os parâmetros e métricas de desempenho do modelo.

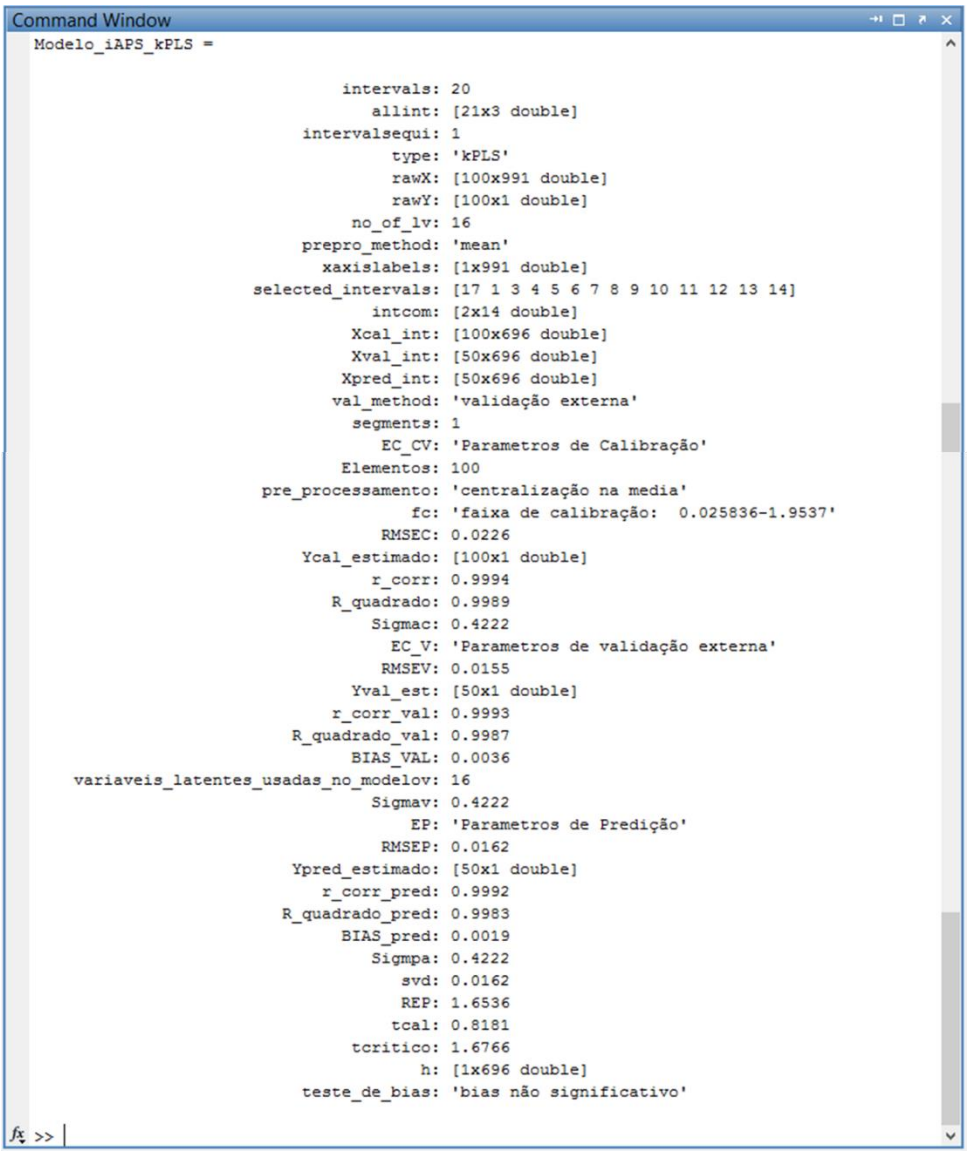

Figura 3.5 - Relatório de saída do algoritmo iSPA-Kernel-PLS

Após a apresentação do relatório, saídas gráficas também são apresentadas (figura 3.6) contendo os gráficos de valores preditos *versus* valores observados, para os conjuntos de calibração, validação (cruzada ou externa) e predição, os intervalos selecionados e as elipses EJCR para cada subconjunto de amostras. Os parâmetros do modelo apresentados no relatório são salvos no espaço de trabalho do Matlab com o nome "Modelo_iAPS_kPLS". De acordo com a necessidade do operador, esses dados podem ser facilmente acessados ou armazenados em algum diretório do computador. Na Figura 3.7 é apresentado o fluxograma representativo do algoritmo.

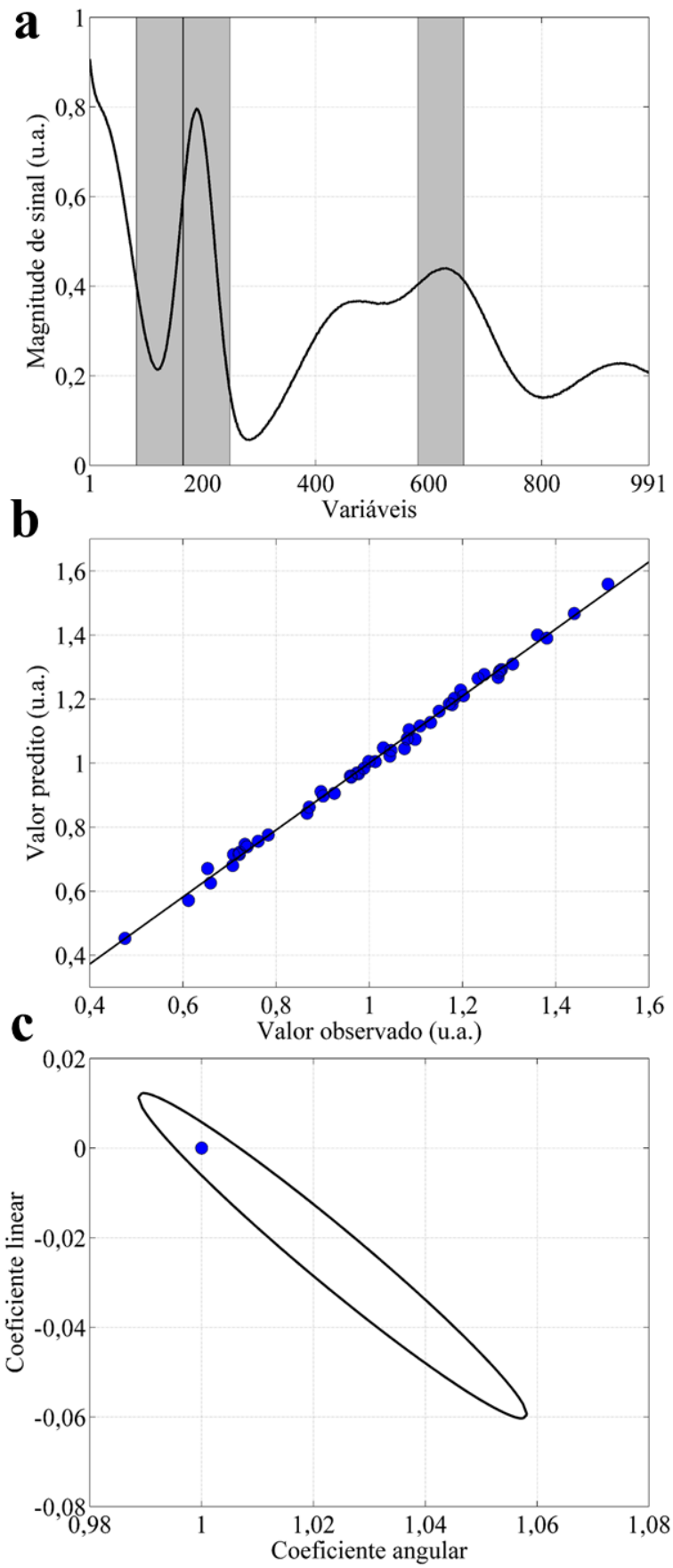

Figura 3.6 - Saídas gráficas do algoritmo iSPA-Kernel-PLS ((a) Intervalos selecionados; (b) Valores observados *versus* valores preditos; (c) elipse de confiança EJCR)

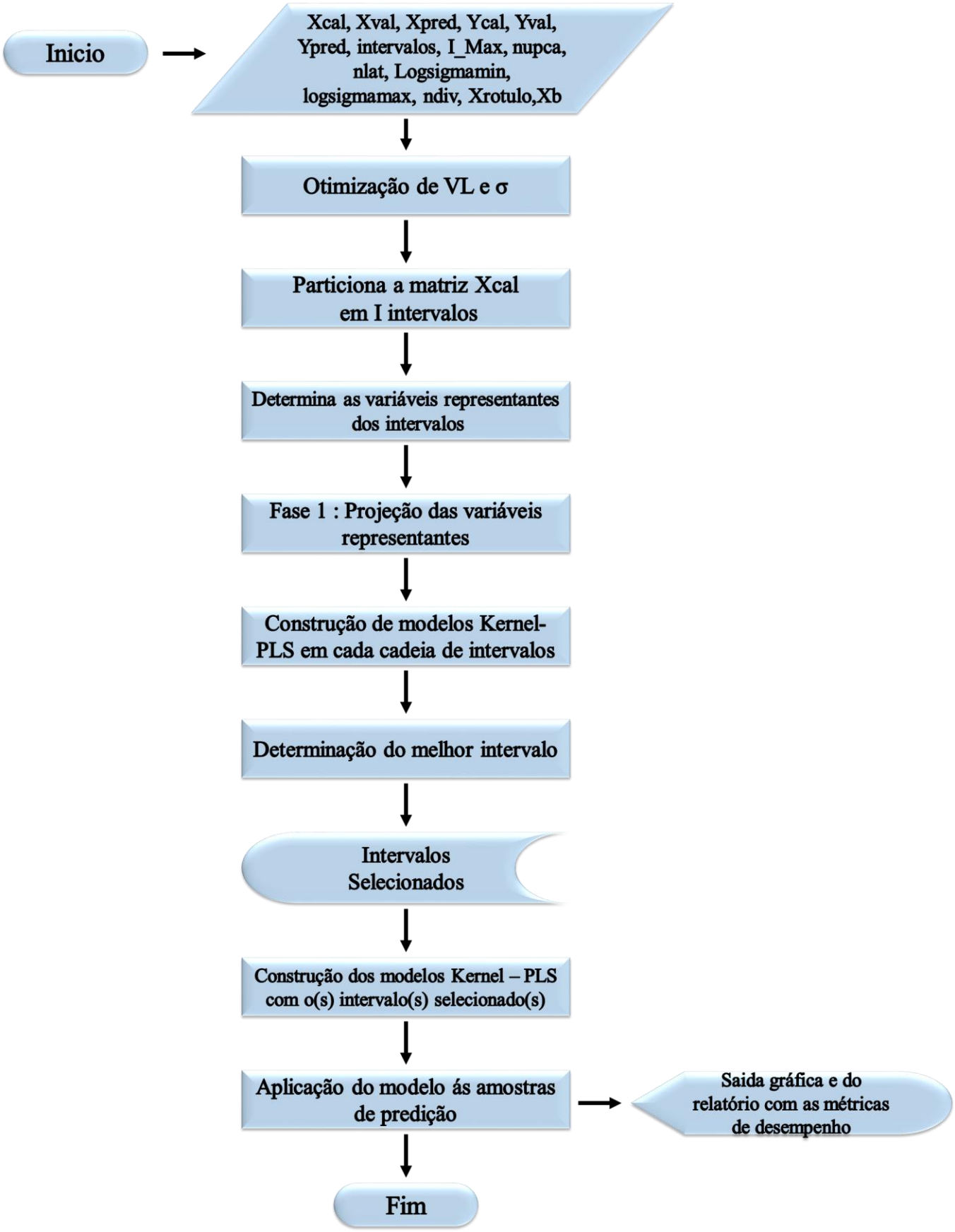

Figura 3.7 – Fluxograma do algoritmo iSPA-Kernel-PLS.

Capítulo 4

Resultados e Discussão

4.1 *INTERVAL* ALGORITMO DAS PROJEÇÕES SUCESSIVAS – KERNEL - MÍNIMOS QUADRADOS PARCIAIS

O SPA é uma técnica de seleção de variáveis primeiramente proposto para seleção de variáveis minimamente correlacionadas na calibração por Regressão Linear Múltipla (MLR, do inglês, *Multiple Linear Regression*) [16]. A principal característica do SPA está na forma com que as cadeias de variáveis são geradas. Partindo de uma variável de referência, calcula-se a projeção das demais variáveis em um plano a 90° desta variável e sucessivamente adicionadas à cadeia as variáveis de maior projeção. Este fato é também o responsável pela característica determinística do algoritmo SPA, de modo que a projeção de uma variável em relação a outra não se altera [16].

O SPA foi adaptado visando à seleção de intervalos para Kernel-PLS a partir de um algoritmo anteriormente proposto por Gomes e colaboradores [13]. O algoritmo para seleção de intervalos é dividido em duas fases, a primeira de geração das cadeias de intervalos e a segunda, de avaliação dos modelos Kernel-PLS construídos em cada um desses intervalos, resultando na seleção do de melhor desempenho. Inicialmente, a matriz de dados contendo V variáveis é dividida em W intervalos não sobrepostos, podendo ou não ser de tamanhos iguais, obedecendo à primeira afirmação quando V/W resultar em um valor inteiro. Em caso contrário, as variáveis restantes são distribuídas entre os intervalos até que o somatório das variáveis de todos os intervalos seja igual ao valor de V. Na fase 1, calcula-se a norma de cada vetor variável dentro dos W intervalos. Cada variável de maior norma dentro dos respectivos intervalos é definida como sua representante e armazenada numa matriz $\mathbf{W}_{cal}$ (N_{cal} x W).

Na etapa de geração, as cadeias são inicializadas com uma coluna de $\mathbf{W}_{cal}$, chamada $\mathbf{w}_v$ e cada uma é aumentada em relação à anterior num processo aditivo de variáveis chamado *forward* até um número M de colunas adicionadas à cada cadeia, correspondente

ao número máximo de intervalos a serem selecionados, definido inicialmente pelo usuário. Esse processo é repetido até que todas as colunas de $\mathbf{W}_{cal}$ tenham sido utilizadas como variáveis de referência na inicialização das cadeias. Cada incremento dentro da cadeia é realizado de acordo com o critério de maior projeção, adicionando sempre variáveis minimamente correlacionadas. Ainda na fase 1, as cadeias geradas são armazenadas numa matriz **SEL** (M x V) contendo os índices correspondentes aos *W* intervalos. O processo da fase 1 do SPA é descrito a seguir:

1. Inicialização

   $\mathbf{z}^{\mathbf{1}} = \boldsymbol{w}_v$ (vector que define as operações iniciais de projeção)

   $\mathbf{w}_j^{\mathbf{1}} = \mathbf{w}_j$, $j = 1, ..., V$

   SEL(1 , v) = v

   $i = 1$, (Contador de interações).

2. Cálculo da matriz $\mathbf{P}^i$ da projeção no subespaço ortogonal a $\mathbf{z}^i$ como:

$$\mathbf{P}^i = \mathbf{I} - \frac{\mathbf{z}^i(\mathbf{z}^i)'}{(\mathbf{z}^i)'\mathbf{z}^i} \qquad \text{Equação 22}$$

   Onde **I** é a matriz de identidade de tamanho (N_{cal} x N_{cal}), z é o vetor que define as operações iniciais de projeção.

3. Cálculo dos vetores $\boldsymbol{w}_j^{i+1}$ projetados como:

$$\boldsymbol{w}_j^{i+1} = \mathbf{P}^i \mathbf{w}_j^i, \text{ para todo } j = 1, \ldots, V. \qquad \text{Equação 23}$$

   Onde **P** é a matriz projeção no subespaço ortogonal.

4. Determinação do índice j^* do maior vetor projetado no plano ortogonal a $\mathbf{z}^i$, e armazenamento desse índice no elemento (i + 1, V) da matriz **SEL**:

$$j^* = \underset{j=1,\ldots,V}{\arg\max} \left\| \boldsymbol{w}_j^{i+1} \right\| \qquad \text{Equação 24}$$

$$\mathbf{SEL}(i + 1, V) = j^*. \qquad \text{Equação 25}$$

   Onde $\boldsymbol{w}_j^{i+1}$ é o vetor projetado para cada elemento i da matriz **W**.

5. Faz-se $\mathbf{z}^{i+1} = \boldsymbol{w}_{j*}^{i+1}$ **(**vetor que define as operações de projeções das interações subsequentes)
6. Se $i < M$, faz-se $i = i + 1$ e retorna-se a etapa 2.

Na fase 2, é construído um modelo Kernel–PLS para cada cadeia de intervalos indicados pelos índices localizados na matriz **SEL**. Os subconjuntos de M intervalos candidatos obtidos a partir de $\boldsymbol{w}_v$ são definidos pelo conjunto de índices *{SEL(1, v), SEL(2, v), ..., SEL(m, v)}*, onde *v* varia de 1 a *W* e *m* varia de 1 a (*W*-1) uma vez que ao admitir *m*=*W*, admite-se também a utilização de todos os intervalos para construção do modelo, retornando a um *full* Kernel-PLS ao invés de um modelo iSPA-Kernel-PLS. A combinação ótima de intervalos é determinada de acordo com o menor valor de RMSECV ou de validação externa RMSEV.

4.2 ANÁLISE DOS DADOS SIMULADOS

4.2.1 Banco de dados 1

4.2.1.1 Kernel - Mínimos Quadrados Parciais

Os espectros obtidos para cada amostra são demonstrados na Figura 4.1, referentes a 100 amostras de calibração e a 50 de predição, perfazendo uma quantidade total de 150 amostras. Cada amostra foi obtida por meio do somatório dos constituintes descritos na seção 3.1.1, nas diferentes concentrações.

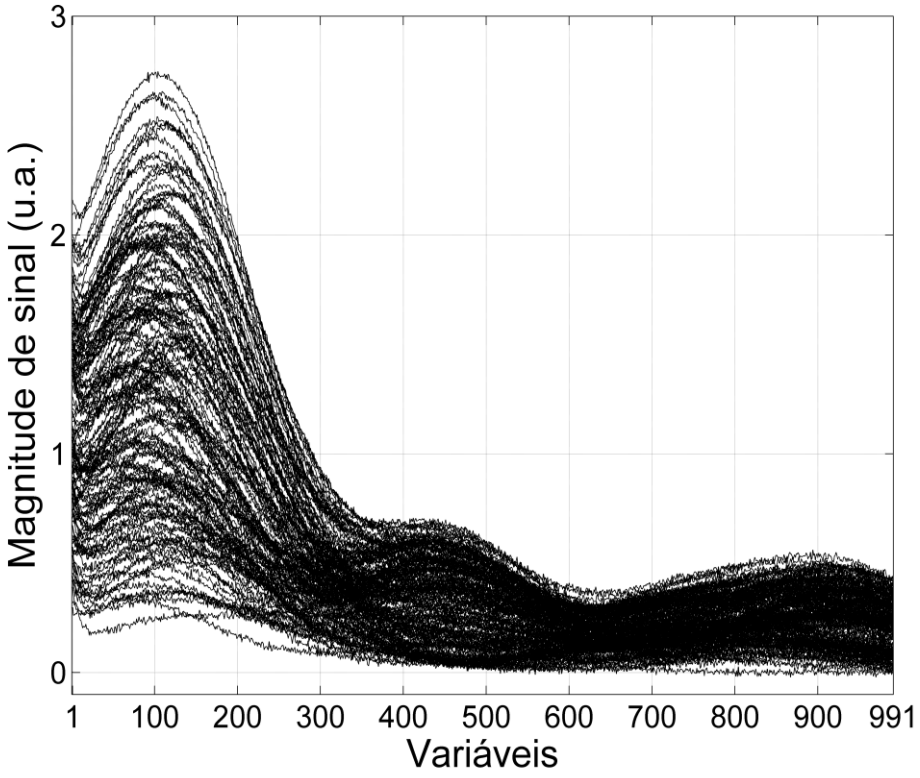

Figura 4.1 - Resposta dos dados simulados 1 das amostras de calibração e predição.

Utilizando o modelo *full spetrum*, foi realizada uma avaliação por validação cruzada *leave one out* do número ideal de VLs e σ. Na Figura 4.2 está apresentada a flutuação dos valores de PRESS representada pela superfície de resposta, em relação a estes parâmetros.

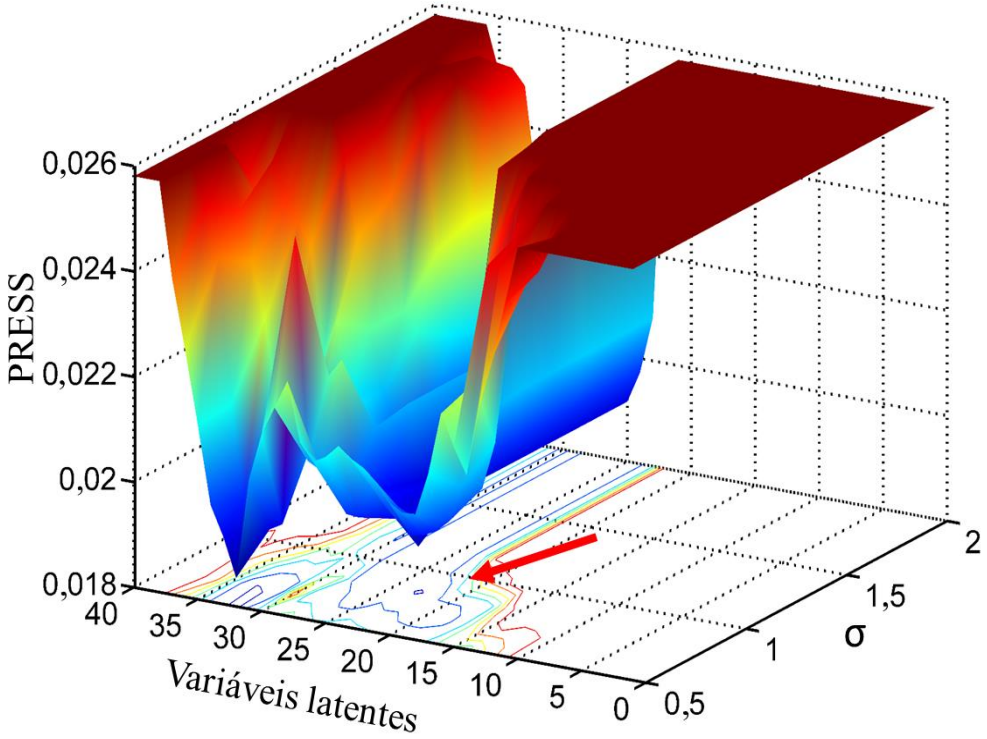

Figura 4.2 - Otimização de VLs e σ para o *full* Kernel-PLS (→ ponto escolhido PRESS =0,02; VL = 22; σ = 1,0).

São verificados pontos de mínimo de acordo com a superfície de resposta em torno de 32 VLs e σ =0,5. Entretanto, como é realizada uma avaliação da possibilidade de se reduzir a quantidade de VLs, visando uma maior parcimônia, outros pontos são avaliados, sem provocar aumento significativo do valor de PRESS. Um modelo com um menor número de fatores necessita de uma capacidade computacional também menor, acelerando o processamento dos dados. O ponto correspondente aos valores escolhidos é indicado pela seta vermelha (σ = 1 e VLs = 22). Provavelmente, o ponto de menor valor de PRESS está associada a um sobreajuste do modelo, devido a uma maior quantidade de VLs utilizadas [45].

A partir dos valores obtidos na otimização, foram construídos modelos Kernel-PLS para os dados *full spectrum*, utilizando as amostras do conjunto de calibração. Em seguida, por meio da validação *leave on out* o modelo construído e na Figura 4.3 são

apresentados os gráficos de valores preditos *versus* valores observados e os gráficos das elipses de confiança.

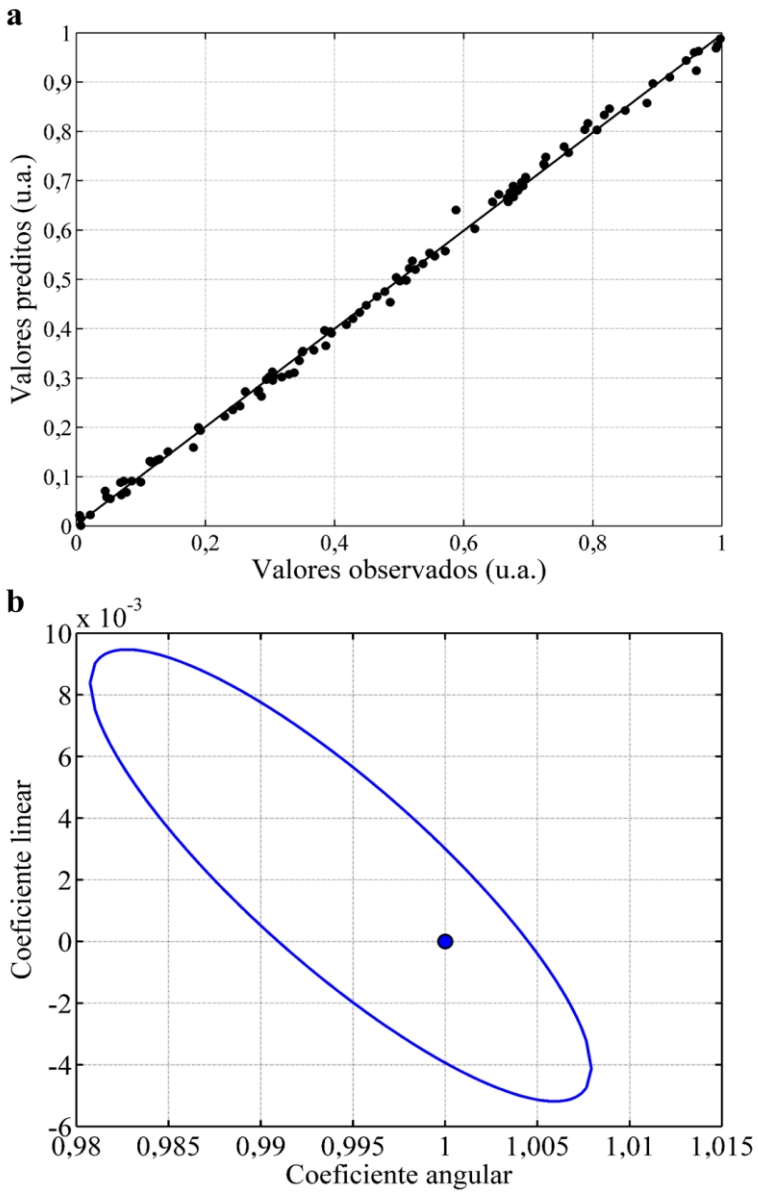

Figura 4.3 - Parâmetros de validação do modelo *full* Kernel-PLS: (a) reta de ajuste dos valores preditos *versus* observados por OLS; (b) elipse EJCR.

O modelo obtido utilizando todo o espectro de dados, demonstra indícios de que pode ser utilizado para a predição da concentração do analito simulado neste estudo, de acordo com o gráfico de valores preditos *versus* valores observados, uma vez que a maioria das amostras apresentaram-se bastante próximas à reta de ajuste, que tem valores de R^2 = 0,997 e RMSECV = 0,0138, valor que está pouco acima do ruído arbitrário adicionado ás

medidas (0,0100), sendo esse erro em termos percentuais no valor de 2,99%. O método EJCR também confirma que o modelo é valido uma vez que a elipse contém dentro do intervalo de confiança o ponto ideal.

Uma consideração a ser feita, é a quantidade de VLs que foram utilizadas no modelo (22 VLs). Além disso, apesar de ter sido validado, ainda aparecem amostras deslocadas da reta de ajuste. O modelo foi então utilizado para a predição de amostras externas às que foram utilizadas para sua construção. Na Figura 4.4 são apresentados os gráficos de desempenho do modelo para o conjunto de predição.

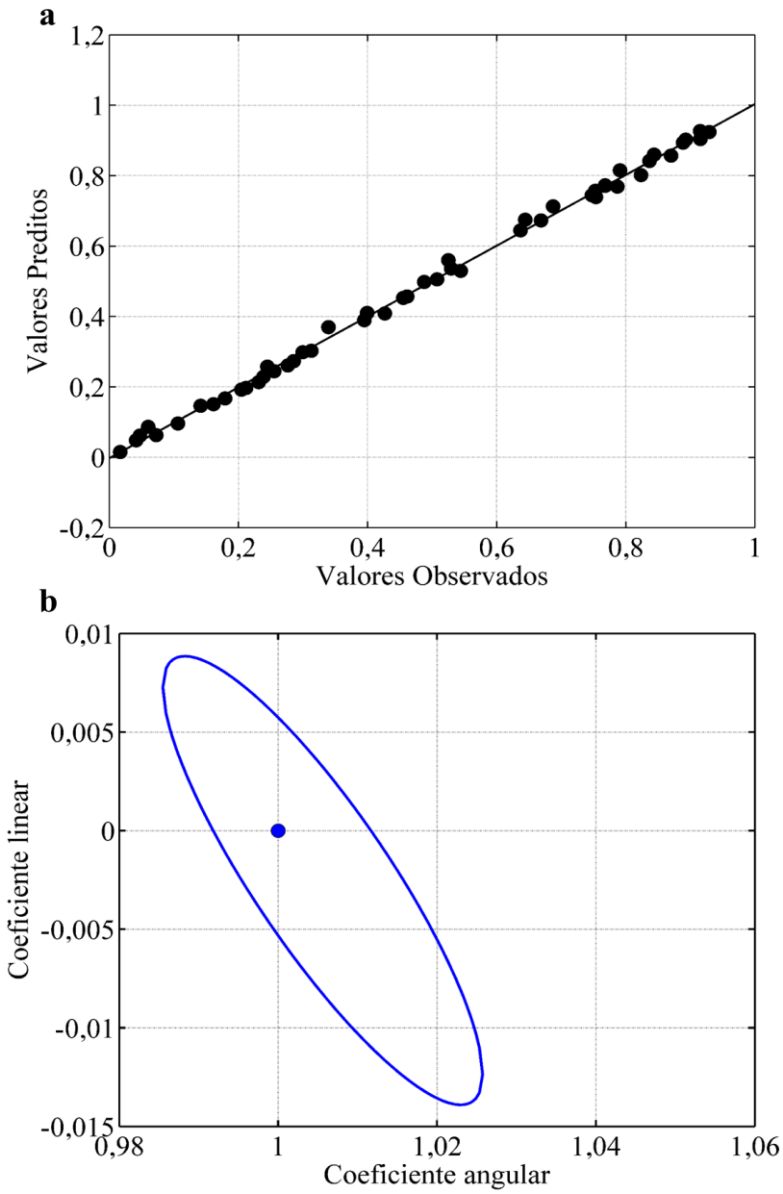

Figura 4.4 - Parâmetros de predição do modelo *full* Kernel-PLS: (a) reta de ajuste dos valores preditos *versus* observados por OLS; (b) elipse EJCR.

Os resultados obtidos para determinação da concentração nas amostras de predição mostraram-se satisfatórios, uma vez observados os valores de R^2 = 0,997 e RMSEP = 0,0143. O erro relativo percentual foi de 2,92%, levemente inferior ao erro obtido para o conjunto de validação. O método EJCR também confirma que o modelo é adequado para a predição da concentração de amostras externas, uma vez que a elipse contém o ponto ideal.

4.2.1.2 *Interval* Algoritmo das Projeções Sucessivas – Kernel - Mínimos Quadrados Parciais

Foi construído um modelo para cada quantidade W de intervalos a serem selecionados perfazendo uma quantidade de 10 modelos. Os seus desempenhos na determinação da concentração das amostras do conjunto de calibração por validação cruzada são apresentados na Tabela 4.1.

Tabela 4.1 - Resultados da validação dos modelos iSPA-Kernel-PLS

Modelo	R^2	r	RMSECV	REP %	*Bias* (tcal) tcrít = 1,6604	N° variáveis	VL
Kernel-PLS	0,997	0,998	0,0138	2,99	4,62e-4 (0,3337)	991	22
2^a-iSPA-k-PLS(1)^b	0,996	0,998	0,0162	3,52	9,74e-4 (0,5980)	496	35
3^a-iSPA-k-PLS(1)^b	0,998	0,999	0,0119	2,57	7,24e-5 (0,0608)	330	22
4^a-iSPA-k-PLS(1)^b	0,998	0,999	0,0122	2,65	1,31e-4 (0,1076)	248	21
5^a-iSPA-k-PLS(1)^b	0,997	0,999	0,0132	2,86	8,48e-5 (0,0641)	198	17
6^a-iSPA-k-PLS(3)^b	0,996	0,998	0,0162	3,52	9,74e-4 (0,5980)	496	35
7^a-iSPA-k-PLS(2)^b	0,998	0,999	0,0116	2,52	4,18e-4 (0,3589)	283	16
8^a-iSPA-k-PLS(2)^b	0,998	0,999	0,0122	2,65	1,32e-5 (0,1076)	248	21
9^a-iSPA-k-PLS(2)^b	0,998	0,999	0,0113	2,45	2,31e-4 (0,2039)	220	15
10^a-iSPA-k-PLS(5)^b	0,996	0,998	0,0162	3,52	9,74e-4 (0,5980)	496	35

a Quantidade de intervalos divididos; *b* Quantidade de intervalos selecionados; Faixa de concentração 0,0053 a 0.9976 unidades.

Destacam-se nos modelos construídos utilizando a seleção de intervalos pelo SPA os dois modelos de menor RMSECV quando W = 7 e sendo dois intervalos selecionados (7(2)) e W = 7 e sendo dois intervalos selecionados (9(2)), ambos são construídos eliminando mais de 70% das variáveis originais, reduzindo consideravelmente a

quantidade de VLs, 22 para 15 e 16 respectivamente. Os modelos 4(1), 8(2) e 3(1) apesar de também apresentarem um erro de validação próximo aos dos modelos 7(2) e 9(2), não apresentaram uma redução significativa na quantidade de VLs. O modelo 5(1) não foi também considerado, uma vez que apresenta piora significativa no valor do RMSECV (p=0,0200 contra p crítico 0,0500. De acordo com o teste *t* randômico [52]) em relação ao modelo de menor erro 9(2). Na Figura 4.5 é apresentada a variação do RMSECV em relação à quantidade de variáveis originais, reforçando o destaque para os modelos 7(2) e 9(2). Avaliando estes dois modelos verifica-se que ao menos em termos do erro de validação, a eliminação das variáveis possivelmente não informativas tende a reduzir o erro.

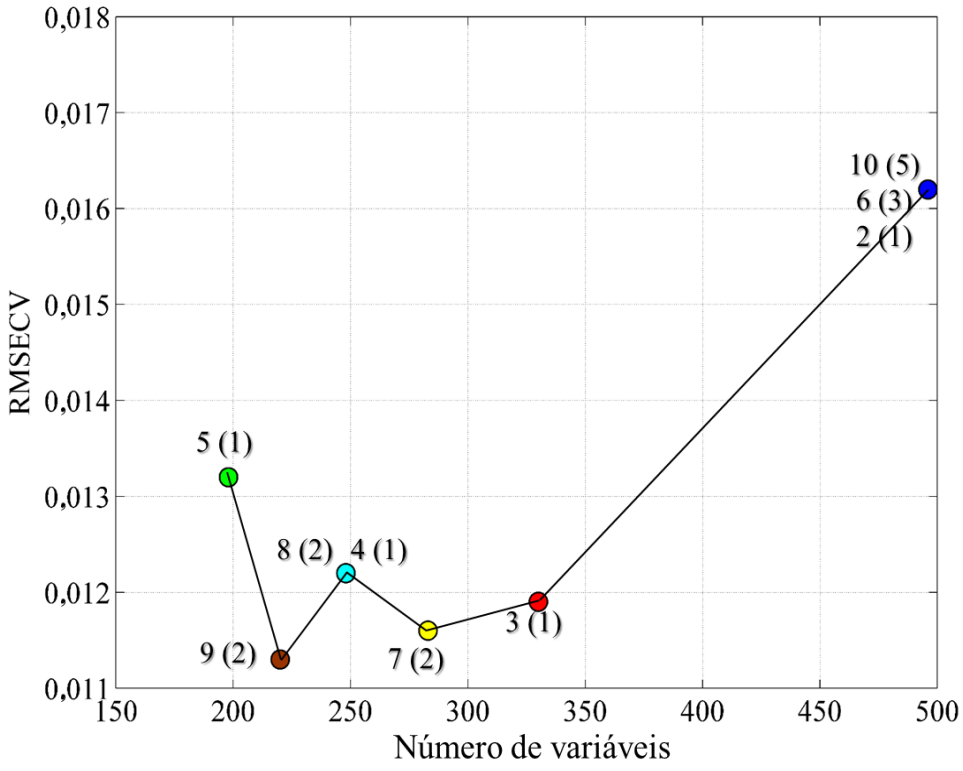

Figura 4.5 - RMSECV em função do número de variáveis dos dados simulados 1.

Os intervalos selecionados pelo SPA por validação cruzada para a construção dos modelos 7(2) e 9(2) são apresentados na Figura 4.6. Apesar de terem selecionado quantidades iguais de intervalos, a quantidade de varáveis em cada intervalo foi diferente.

Ambos os modelos selecionaram os intervalos na região onde a resposta referente ao analito foi adicionada. Além disso, nenhum outro intervalo foi selecionado na região onde apenas os constituintes coexistiam (500 a 991), fato que demonstra, para esse caso, que o algoritmo foi capaz de identificar a região que continha informação útil, descartando as informações não interessantes para a solução do problema.

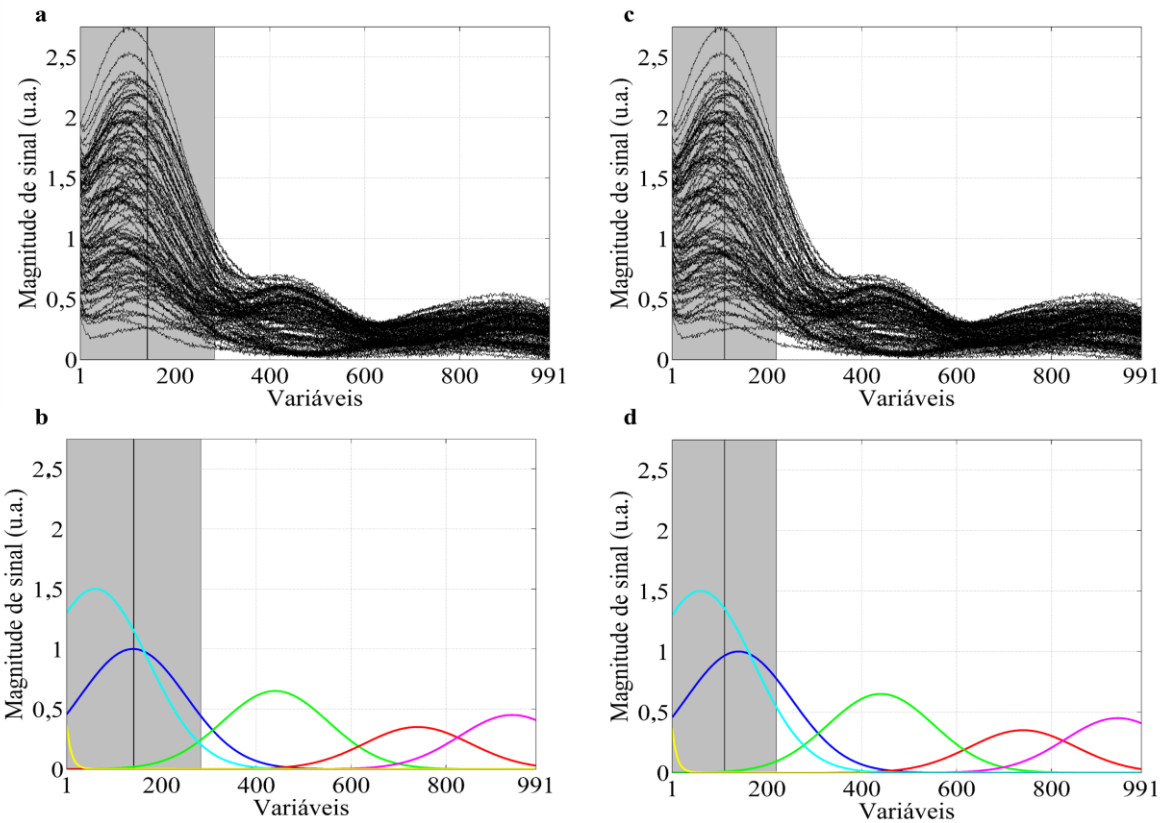

Figura 4.6 - Intervalos selecionados pelo iSPA-Kernel-PLS: para o modelo 7(2) sobre as respostas das amostras (a) e os perfis puros (b), e para o modelo 9(2) sobre as respostas das amostras(c) e os perfis puros (d) para o banco de dados 1.

A região compreendida entre 285 e 500 não foi incluída em nenhum dos dois modelos. Apesar desta região conter ainda informação do analito, apresenta um sinal de mínimo para o analito e máximo para o constituinte 1. A região onde os dois modelos não coincidiram na seleção foi entre 221 e 284. Essa região engloba uma intercessão entre os constituintes verde e ciano, que se sobrepõem, respectivamente, total e parcialmente com o analito. A otimização da quantidade de VLs e do σ para os modelos 7(2) e 9(2) é

demonstrada na Figura 4.7. No caso do modelo 7(2), assim como ocorreu para o modelo *full spetrum*, a quantidade de VLs não compreende a região de mínimo, que ocorre utilizando 26 VLs. Novamente nesse caso, pela diferença entre o mínimo e o ponto escolhido não ser significativa, e esta condição garante maior parcimônia do modelo. A quantidade de 16 VLs foi a escolhida. Para o modelo 9(2), houve a coincidência da quantidade de VLs escolhida estar no mínimo da superfície de resposta. Os dois modelos obtiveram o valor de σ = 0,500 como ideal.

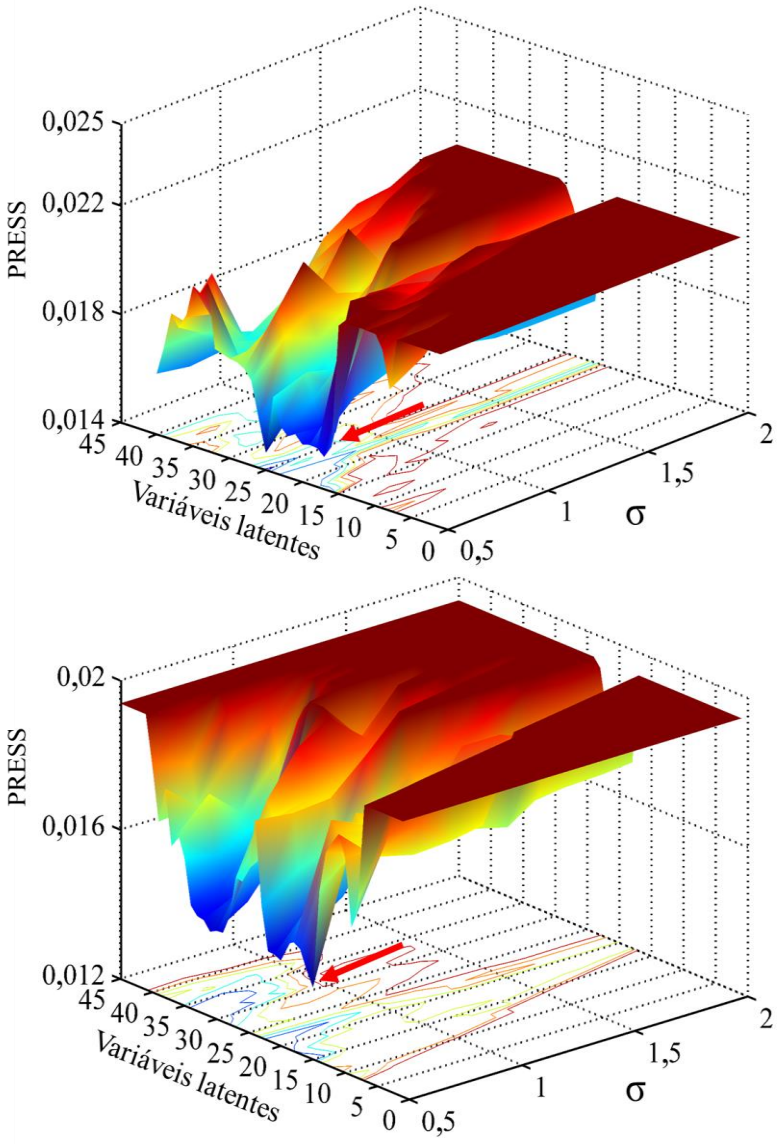

Figura 4.7 - Superfícies de resposta para otimização dos parâmetros de construção dos modelos: (a) → ponto escolhido para o modelo 7(2) PRESS =0,017; VL = 16; σ = 0,5; (b) → ponto escolhido para o modelo 9(2) PRESS =0,014; VL = 15; σ = 0,5.

Os dois modelos apresentaram erros de validação e parâmetros de desempenho melhores que o método *full*, tais como RMSECV significativamente menor (p=0,0250 7(2) e 0,0100 9(2), contra p critico 0,0500), R^2, e uma melhora substancial na parcimônia, tanto em termos de variáveis originais, como de VLs. Na figura 4.8 estão apresentados os gráficos de valores observados *versus* valores preditos, e elipses de confiança para os modelos 7(2) e 9(2).

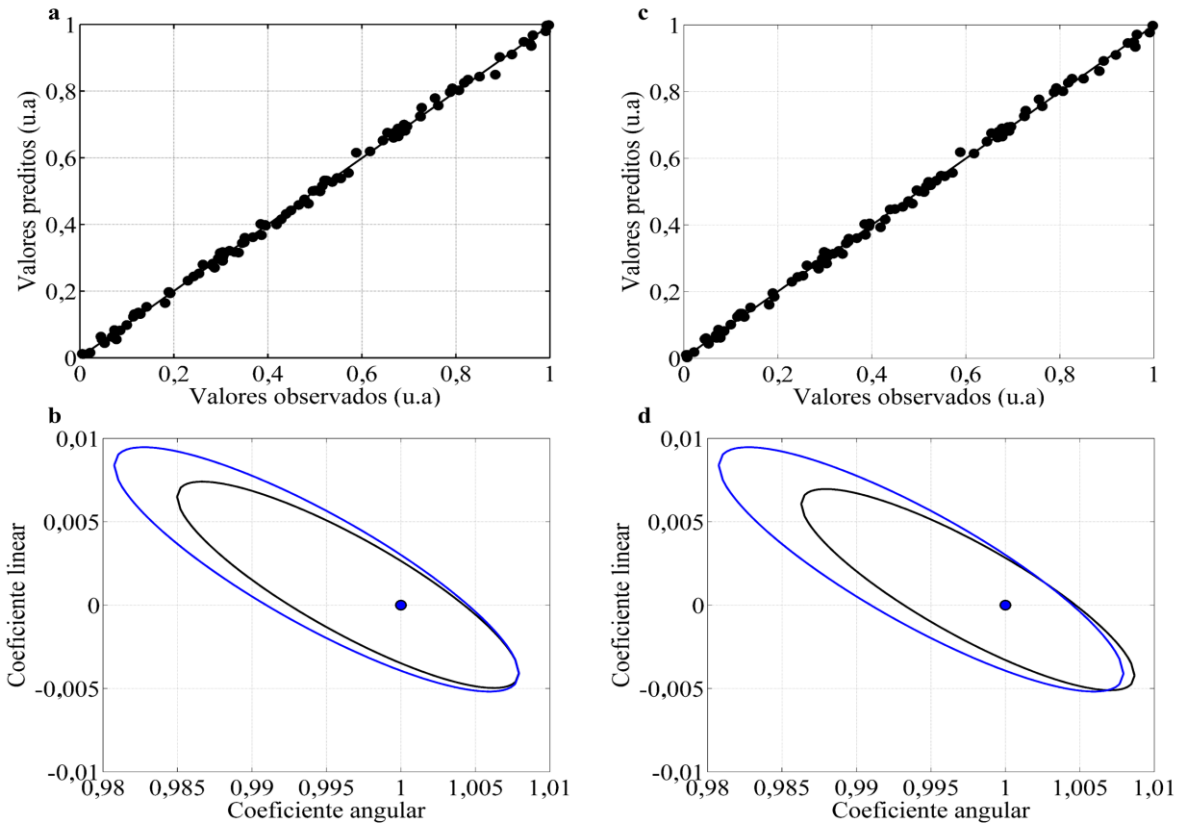

Figura 4.8 - Parâmetros de predição do modelo iSPA-Kernel-PLS para as amostras de calibração por validação cruzada (reta de ajuste dos valores preditos *versus* observados por OLS para os modelos 7(2) (a) e 9(2) (c), elipse EJCR dos modelos 7(2) (b) e 9(2) (d) (Kernel-PLS; iSPA-Kernel-PLS).

Assim como nos resultados numéricos os métodos gráficos demonstram que os modelos são bastante parecidos em termos de validação, e ambos melhores ajustados que o método *full*, tanto em relação ao método EJCR como pela reta de ajuste dos valores preditos *versus* valores observados. Estando os modelos bem validados, estes foram utilizados para a determinação da concentração do analito nas amostras externas de

predição. Com isso esperou-se avaliar a capacidade preditiva dos dois modelos construídos com os respectivos intervalos selecionados. Na Tabela 4.2 estão contidos os resultados referentes ao desempenho no modelo para as amostras de predição.

Tabela 4.2 - Resultado da determinação da concentração por iSPA-Kernel-PLS para as amostras de predição

Modelo	r	R^2	RMSEP	REP %	*Bias* (tcal) tcrit = 1,6766
7(2)	0,998	0,999	0,0126	2,58	1,80e-3 (1,0242)
9(2)	0,979	0,989	0,0448	9,18	1,64e-2 (2,7539)

Com base nos resultados obtidos verifica-se que o modelo 7(2) conseguiu prever as concentrações das amostras de predição de forma bastante satisfatória, obtendo uma leve melhora na predição em relação ao método *full*, não apresentando diferença significativa em comparação dos valores de RMSEP, ou seja, com menos de 30% das variáveis, consegue-se um modelo melhor ajustado e uma predição equivalente ao método *full*. O modelo 9(2) apesar de validado não foi adequado para a predição da concentração de amostras externas. Provavelmente as variáveis adicionadas pelo modelo 7(2) possuem informações úteis, que tornam esse modelo mais generalista que o modelo 9(2).

Para que os modelos de calibração multivariada de primeira ordem consigam predizer as concentrações dos analitos adequadamente frente os constituintes, a informação destes precisam estar variando de forma evidenciada nos dados. Aparentemente, no modelo 9(2) ocorreu este fato, de modo que o modelo conseguiu ajustar bem as amostras de calibração, porém para as amostras de predição não ocorreu o mesmo. Na Figura 4.9 são apresentados os gráficos de desempenho dos modelos para as amostras de predição.

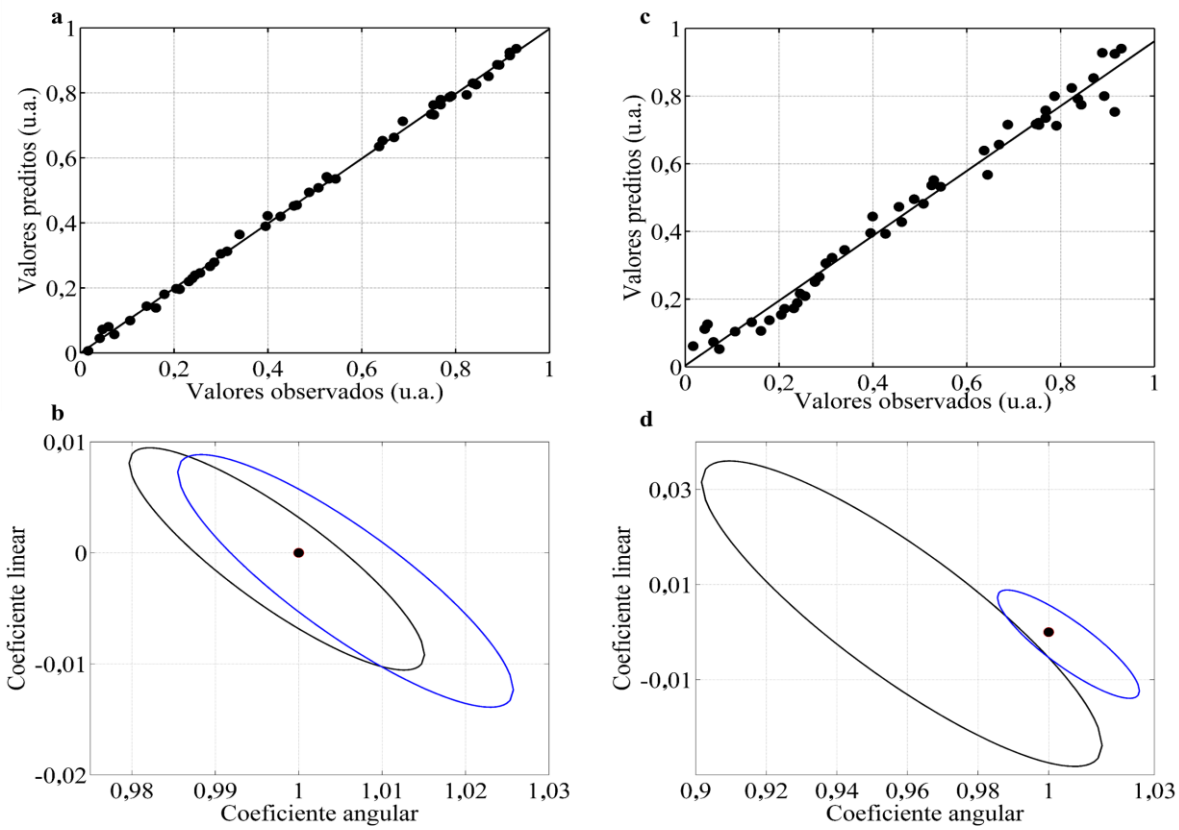

Figura 4.9 - Parâmetros de predição do modelo iSPA-Kernel-PLS para as amostras de calibração por validação cruzada (reta de ajuste dos valores preditos *versus* observados por OLS para os modelos 7(2) (a) e 9(2) (c), elipse EJCR dos modelos 7(2) (b) e 9(2) (d) (▬ Kernel-PLS; ▬ iSPA-Kernel-PLS).

Baseado na inspeção dos parâmetros estatísticos de predição, concomitantemente com as ferramentas gráficas exibidas na Figura 4.9, pode-se inferir que o modelo 7(2) foi adequado para predizer a concentração de todas as amostras do conjunto de predição, em contraste ao modelo 9(2).

4.2.2 Banco de dados 2

4.2.2.1 Kernel- Mínimos Quadrados Parciais

Neste estudo de caso, diferentemente do primeiro banco de dados simulados, foi utilizada para validação de amostras externas. Os espectros gerados para cada amostra são apresentados na Figura 4.10, nela estão contidas as 100 amostras de calibração, as 50

de validação e as 50 de predição perfazendo uma quantidade de 200 amostras obtidas semelhantemente às amostras do primeiro banco de dados.

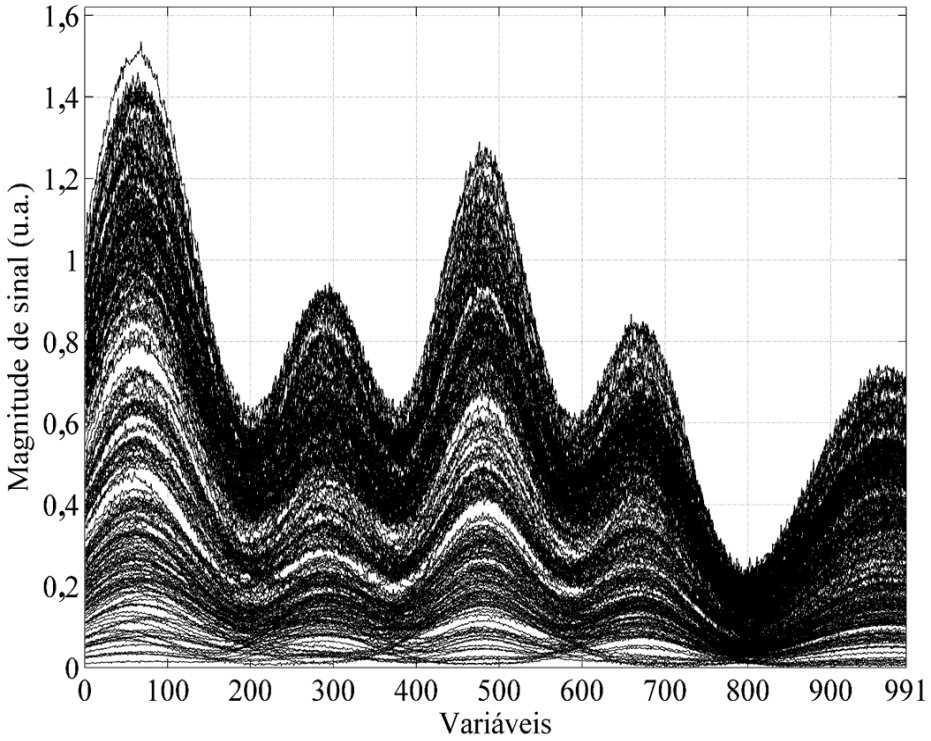

Figura 4.10 - Respostas simuladas para o banco de dados 2.

A busca pelos valores ideais de VLs e σ para a construção do modelo Kernel-PLS utilizando todo o espectro foi realizada e são apresentadas na Figura 4.11 por meio da superfície de resposta.

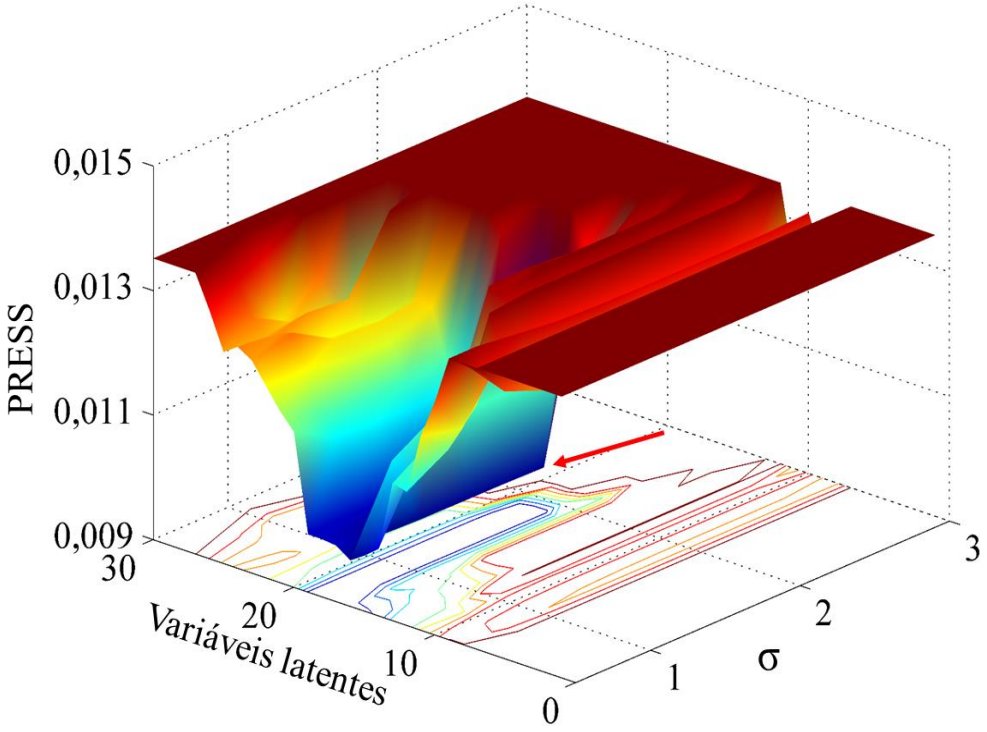

Figura 4.11 - Otimização de VLS e σ para o *full* Kernel-PLS (→ ponto escolhido PRESS =0,010; VL = 17; σ = 1,8889)

O ponto de mínimo PRESS demonstrado pela seta vermelha na superfície de resposta, faz referência ao ponto selecionado para a utilização na construção do modelo, sendo a quantidade de 17 VLs e σ = 1.8889. O modelo Kernel-PLS foi então construído tomando os valores obtidos na otimização, a partir das amostras do conjunto de calibração, e em seguida o modelo foi utilizado para a predição da concentração das amostras externas utilizadas para validação. Na Figura 4.12 são apresentados os gráficos de valores preditos *versus* valores observados e os gráficos das elipses de confiança para o conjunto de validação externa.

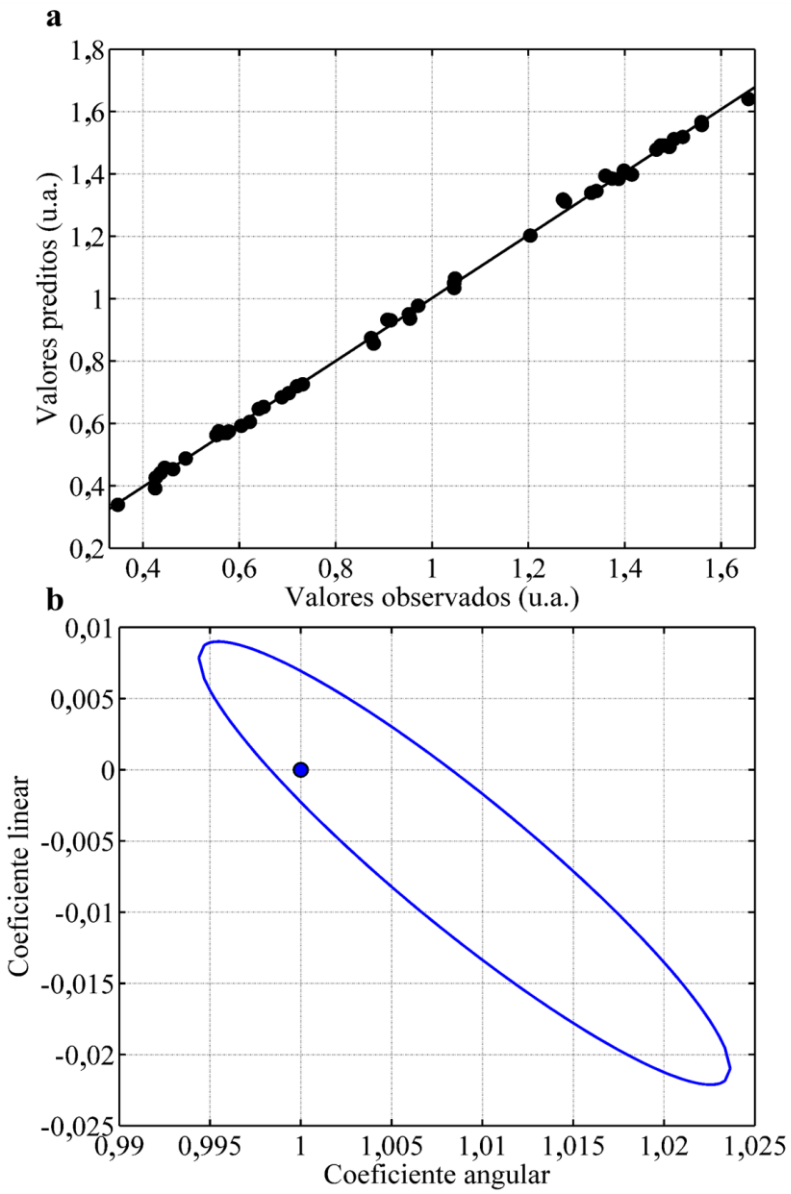

Figura 4.12 - Parâmetros de validação do modelo *full* Kernel-PLS ((a)reta de ajuste dos valores preditos *versus* observados por OLS, (b) elipse EJCR) para o segundo banco de dados.

O modelo obtido utilizando todas as variáveis do espectro de dados, apresenta métricas de desempenho aceitáveis para a utilização na determinação da concentração de amostras em termos da validação externa. As amostras se apresentaram próximas da reta de ajuste feita pelo método OLS, os valores de $R^2 = 0,998$ e RMSEV = 0,0150 sendo esse erro em termos percentuais no valor de 1,52%. O método EJCR corrobora com esses parâmetros confirmando que o modelo pode ser utilizado, uma vez que a elipse contém dentro do intervalo de confiança o ponto ideal.

O modelo foi então utilizado para a predição de amostras para testar o modelo construído, estas amostras são totalmente externas às que foram utilizadas para construção e validação do modelo. Na Figura 4.13 são apresentados os gráficos valores preditos *versus* referência e elipse de confiança para o conjunto de amostras de predição do modelo Kernel-PLS.

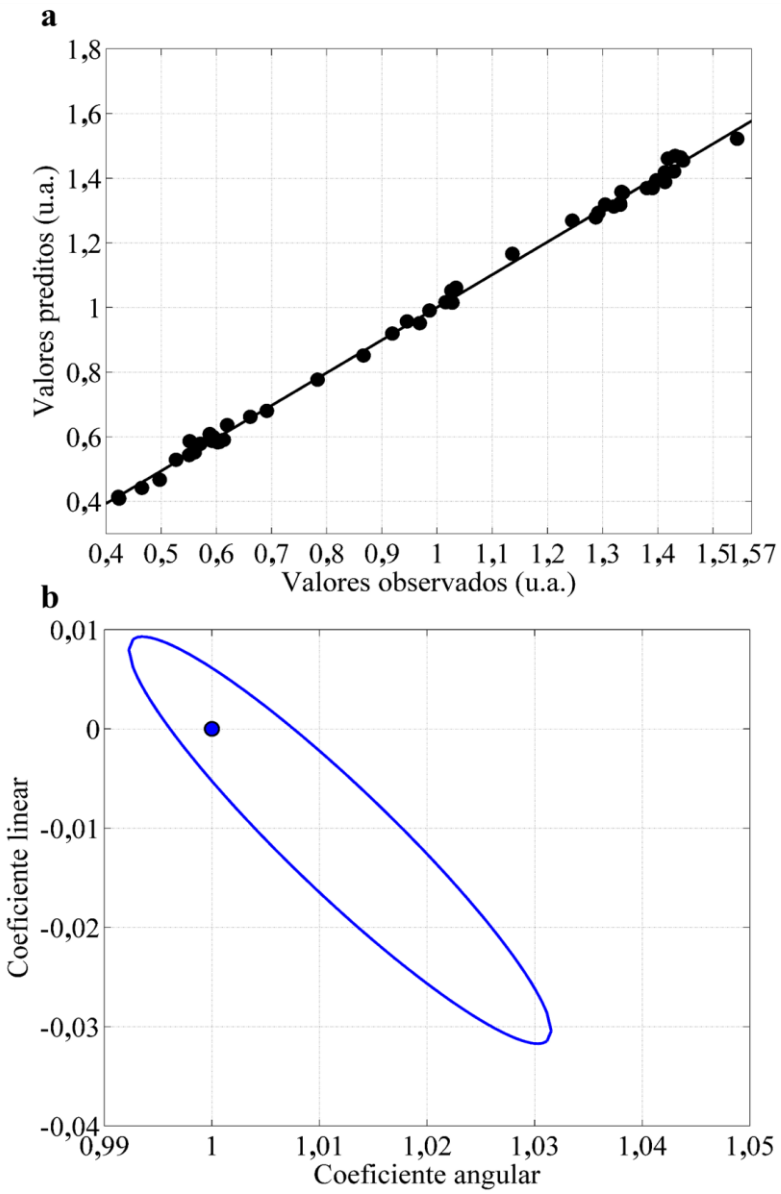

Figura 4.13 - Parâmetros de predição do modelo *full* Kernel-PLS (a)reta de ajuste dos valores preditos *versus* observados por OLS, (b) elipse EJCR) para o segundo banco de dados.

O erro obtido para o conjunto de predição mostrou-se um pouco maior do que o erro obtido para o conjunto de validação, RMSEP = 0,0182, além dos valores de R^2 = 0,997 e

REP = 1,85%. Uma vez que a elipse de confiança obtida pelo método EJCR contém em seu interior o ponto ideal, pode-se afirmar que a predição da concentração das amostras obteve resultados aceitáveis.

4.2.2.2 *Interval* Algoritmo das Projeções Sucessivas – Kernel - Mínimos Quadrados Parciais

Utilizando agora como ponto de partida os valores de VLs e σ obtidos como ideais para o modelo Kernel-PLS, foram realizadas novas otimizações a cada avaliação das cadeias na fase 2 do SPA. Foi construído um modelo para cada *W* perfazendo 10 modelos. Os desempenhos dos modelos na determinação da concentração das amostras do conjunto de validação externa são apresentados na Tabela 4.3.

Tabela 4.3 - Resultados da validação dos modelos iSPA-Kernel-PLS para o banco de dados 2

Modelo	r	R^2	RMSEV	REP %	*Bias* (tcal) tcrit = 1,6766	N° de variáveis	VL
k-PLS	0,998	0,999	0,0150	1,52	2,30 e-3 (1,0515)	991	17
2^a-iSPA-k-PLS(1)^b	0,967	0,983	0,0818	8,29	2,27 e-2 (1,7496)	495	4
3^a-iSPA-k-PLS(2)^b	0,998	0,999	0,0148	1,50	3,50 e-3 (1,7156)	661	17
4^a-iSPA-k-PLS(2)^b	0,998	0,999	0,0133	1,35	2,18 e-4 (0,1147)	496	13
5^a-iSPA-k-PLS(4)^b	0,999	0,999	0,0138	1,39	4,50 e-3 (2,3870)	793	16
6^a-iSPA-k-PLS(5)^b	0,999	0,999	0,0134	1,36	2,00 e-3 (1,0644)	826	17
7^a-iSPA-k-PLS(3)^b	0,999	0,999	0,0131	1,32	2,30 e-3 (1,2268)	425	16
8^a-iSPA-k-PLS(7)^b	0,999	0,999	0,0133	1,35	1,50 e-3 (0,7855)	868	17
9^a-iSPA-k-PLS(8)^b	0,999	0,999	0,0133	1,35	8,29 e-4 (0,4374)	881	17
10^a-iSPA-k-PLS(8)^b	0,999	0,999	0,0128	1,30	1,20 e-3 (0,6715)	792	17

a Quantidade de intervalos divididos; *b* Quantidade de intervalos selecionados; Faixa de concentração 0,3478 a 1.6570 unidades.

O melhor modelo iSPA-Kernel-PLS para este estudo de caso é o 7(3), uma que esse possui o melhor compromisso entre erro de validação e ganho em parcimônia, tendo apresentado o segundo menor erro de validação, e não possuindo diferença significativa entre ele e o modelo de mínimo RMSEV 10(8). O modelo 2(1) como esperado não obteve uma boa predição, uma vez que não foi possível incluir as duas regiões informativas

referentes ao analito, provocando assim um subajuste. É importante ressaltar a importância da escolha da quantidade de intervalos que o espectro será dividido, pois, assim como no caso do modelo 2(1) as informações referentes ao analito podem não ser incluídas simultaneamente nos intervalos aptos a serem selecionados, gerando modelos deficientes de informação. Na Figura 4.14 é apresentada a representação gráfica do comportamento do RMSEV em relação ao número de variáveis originais. De modo geral os modelos apresentaram erros de validação relativamente próximos, sendo determinante para a escolha do modelo 7(3) a redução de 57% das variáveis originais e uma variável latente.

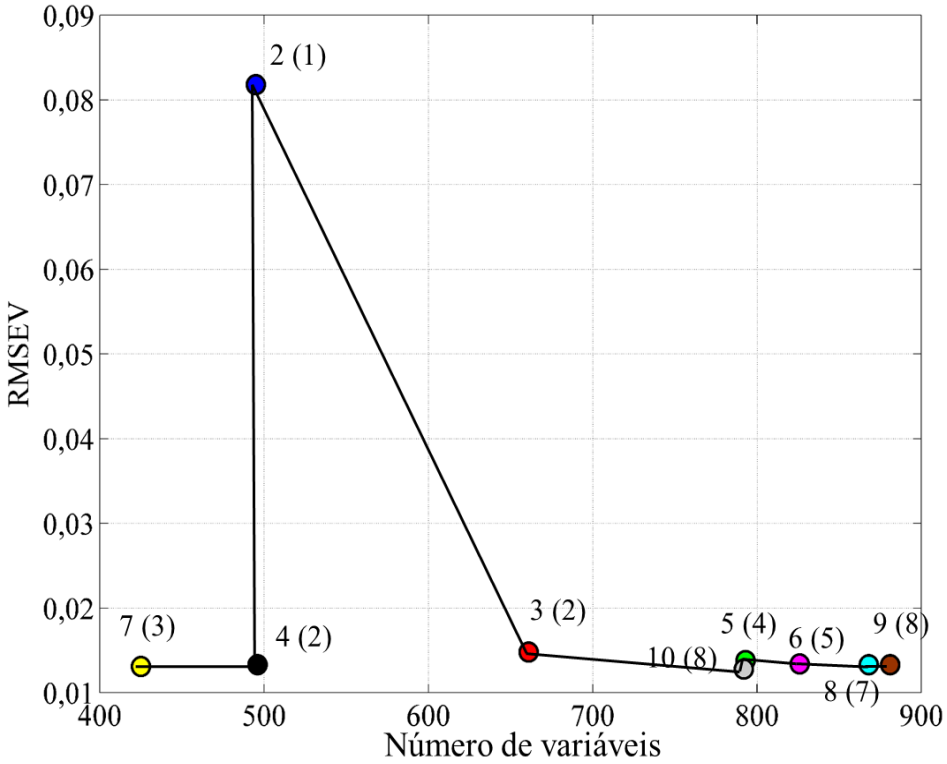

Figura 4.14 - RMSEV em função do número de variáveis para o segundo banco de dados.

Os intervalos selecionados para a construção do modelo iSPA-Kernel-PLS por validação externa são apresentados na Figura 4.15. Os intervalos selecionados condizem com a expectativa do momento da geração dos dados, os três intervalos compreendem a

região de resposta do analito, além disso contém cinco intercessões entre os constituintes e o sinal do analito, e duas intercessões entre dois constituintes. Regiões onde a contribuição dos constituintes é máxima foram evitadas, são essas: entre a variável 1 e 142, a região entre 427 e 568 e a região entre 710 e 991. As regiões de máximo do sinal do analito também incluídas, estas todas em mínimo dos outros constituintes.

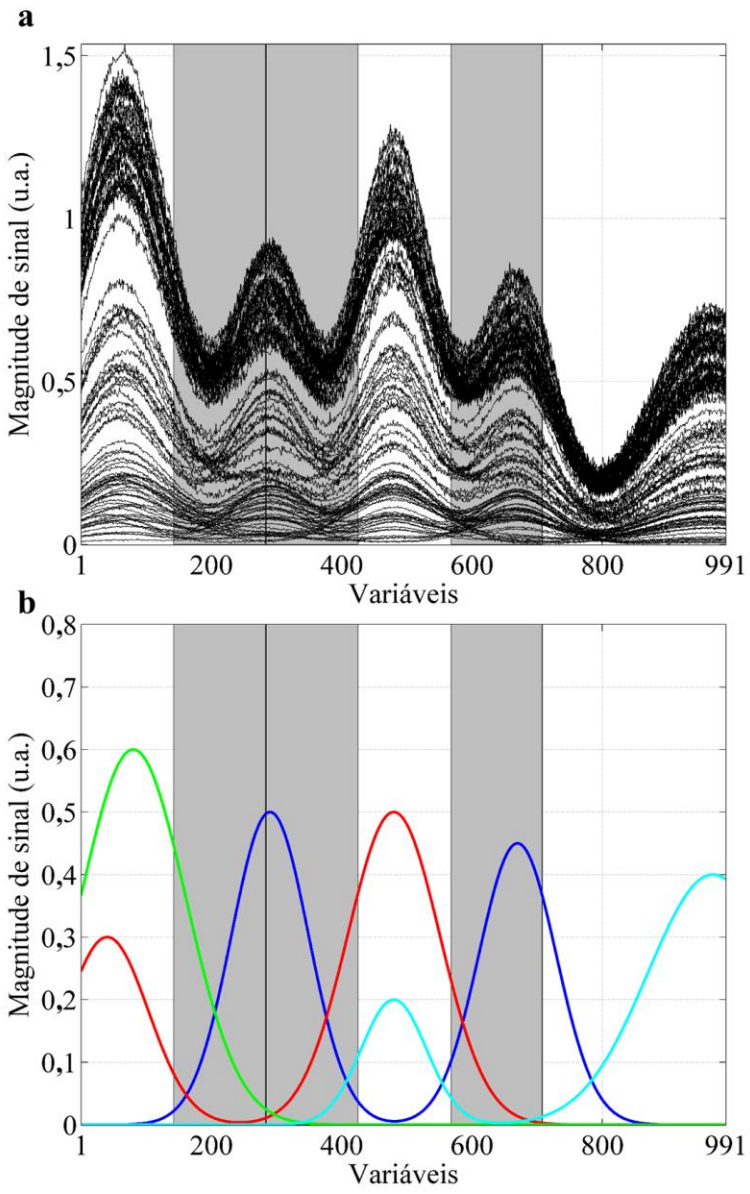

Figura 4.15 - Intervalos selecionados pelo iSPA-Kernel-PLS para o modelo 7(3) sobre as respostas das amostras (a) e os perfis puros (b).

A otimização da quantidade de VLs e do σ para o modelo 7(3) é demonstrada na Figura 4.16. Para o modelo iSPA-Kernel-PLS foram obtidos após a otimização dos valores de VLs 16 e o parâmetro da largura da transformação gaussiana σ = 2,1667.

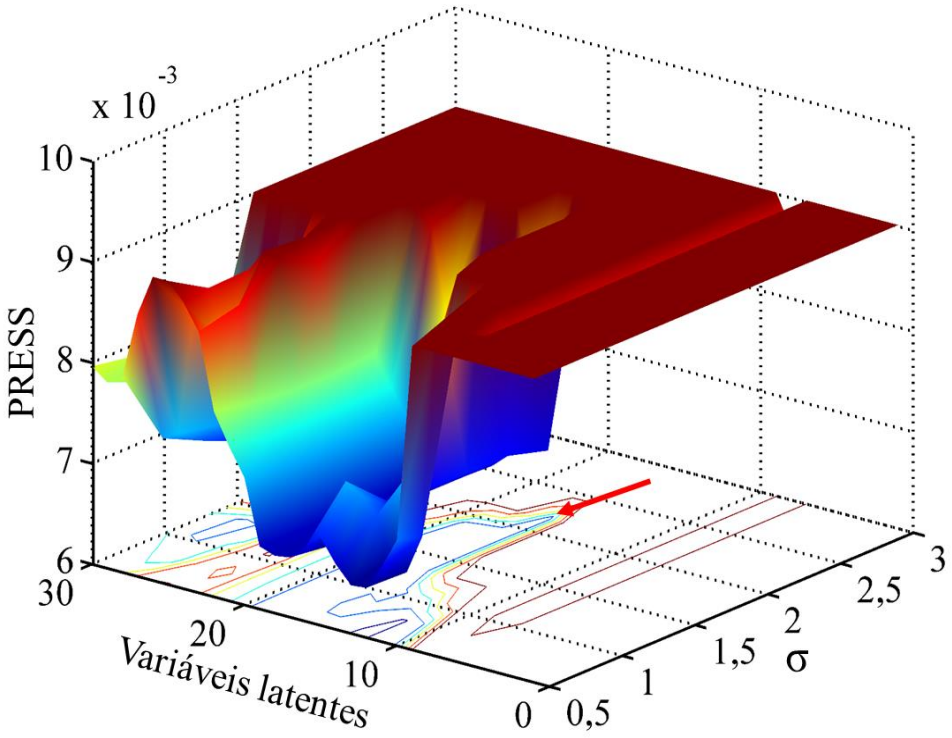

Figura 4.16 - Superfícies de Resposta para otimização dos parâmetros de construção do modelo iSPA-Kernel-PLS para o segundo banco de dados.

O modelo iSPA-Kernel-PLS apresentou erros de validação e métricas de desempenho equivalentes ao modelo Kernel-PLS mesmo utilizando menos da metade das variáveis originais.

Na Figura 4.17 estão apresentados os gráficos de valores observados *versus* valores preditos, e elipses de confiança. A remoção das variáveis não informativas neste estudo de caso não causou uma diminuição significativa no erro de validação, entretanto, pôde se mostrar que a presença dessas variáveis também não trazia nenhum benefício para o modelo. Dessa forma o benefício da seleção dos intervalos nestes casos está na possibilidade de restringir a medição da informação instrumental apenas para essa região

especifica, além de reduzir o esforço computacional no processamento dos dados e consequentemente no tempo da análise. O RMSEV para o modelo 7(3) foi de 0,0131, R^2, além da redução de uma VL.

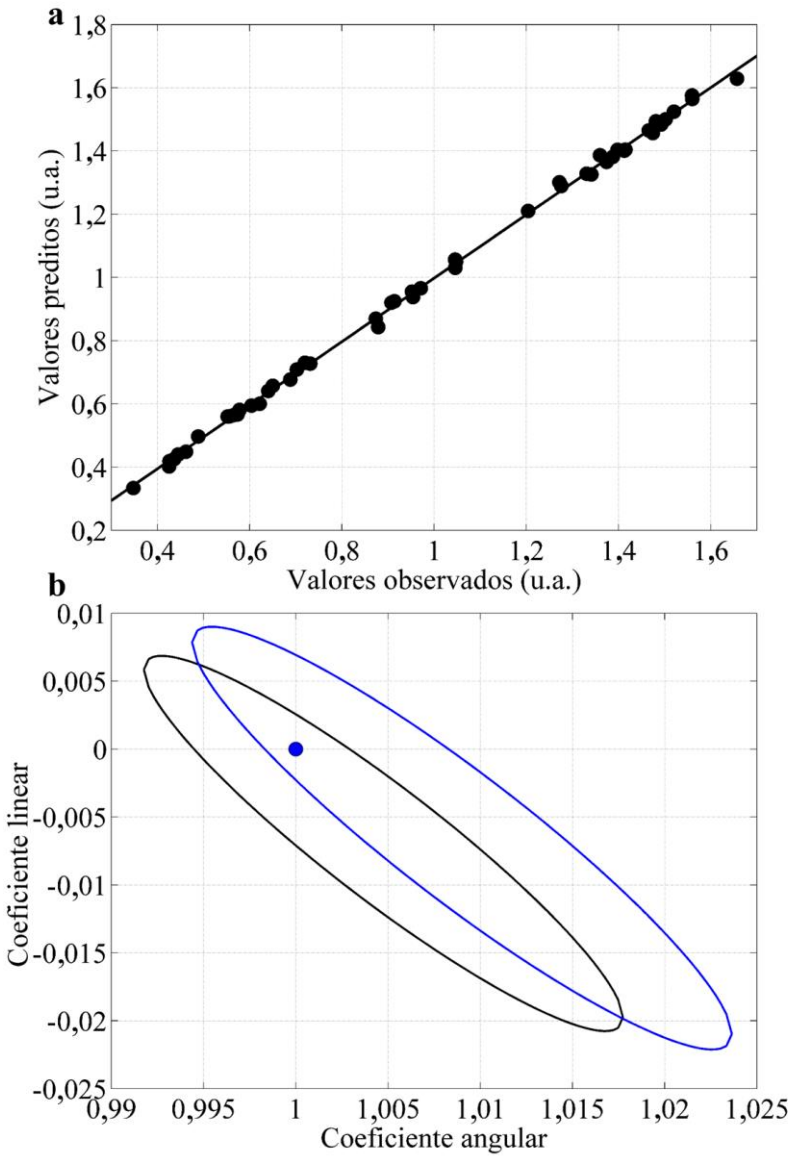

Figura 4.17 - Parâmetros de predição do modelo iSPA-Kernel-PLS para as amostras de validação externa (reta de ajuste dos valores preditos *versus* observados por OLS (a) e elipse EJCR (b) (Kernel-PLS; iSPA-Kernel-PLS)).

Analisando os tamanhos das elipses, verifica-se uma leve diferença entre o modelo 7(3) e o modelo *full* assim como verificado nas outras métricas de desempenho, mostrando que os modelos possuem validação bem próximas, com o iSPA-Kernel-PLS

um pouco melhor em termos de erro de validação. No gráfico de valores observados *versus* valores preditos verifica-se que as amostras estão bem comportadas e bastante próximas da reta, o que representa que os valores preditos se aproximaram bastante do valor real.

O modelo construído foi então utilizado para a determinação da concentração do analito nas amostras externas de predição. Com base nos resultados obtidos verifica-se que o modelo 7(3) obteve êxito na previsão das concentrações das amostras do conjunto de predição, apresentando desempenho equivalente ao observado para o método *full*. Na Figura 4.18 são apresentados os gráficos de desempenho do modelo para as amostras de predição.

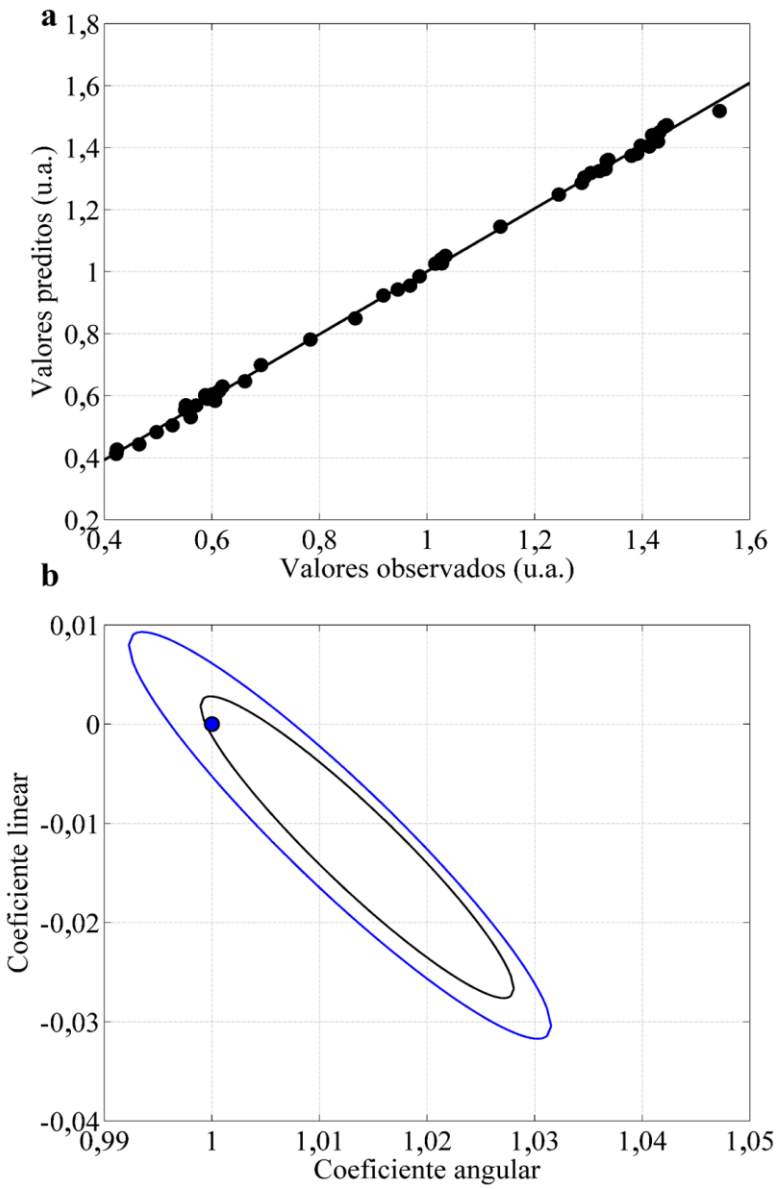

Figura 4.18 - Parâmetros de predição do modelo iSPA-Kernel-PLS para as amostras de predição (reta de ajuste dos valores preditos *versus* observados por OLS (a) e elipse EJCR (b) (Kernel-PLS; iSPA-Kernel-PLS)).

Assim como os resultados de validação, na predição o comportamento do modelo foi similar, verifica-se uma leve melhora no desempenho em termos de RMSEP = 0,0140, R^2 = 0,998 e REP = 1,00%.

Com base nos resultados obtidos nos dois bancos de dados simulados, verifica-se a potencialidade da seleção de intervalos na calibração multivariada não linear, em sistemas controlados. O algoritmo mostrou-se consistente no objetivo de buscar as variáveis mais

informativas, e eliminar as não informativas, com o objetivo de otimizar a utilização dos dados para a calibração.

Baseado no desempenho frente aos dados simulados, o algoritmo foi empregado em um estudo de caso real, para a determinação de dois parâmetros em amostras oriundas da produção de açúcar.

4.3 ANÁLISE DOS DADOS EXPERIMENTAIS

Os espectros NIR na faixa de 400 a 1.888 nm para as 1.797 amostras são exibidos na Figura 4.19. As amostras foram divididas em calibração (1.000 amostras), validação (397 amostras) e predição (400 amostras).

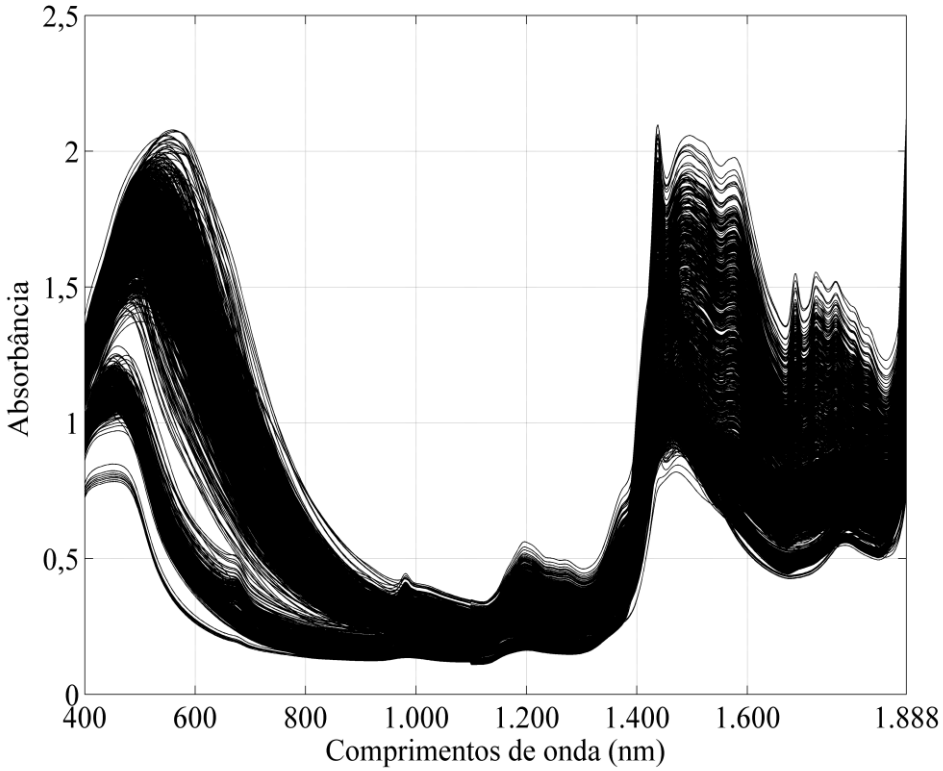

Figura 4.19 - Espectros NIR para amostras obtidas na produção de açúcar.

O espectro obtido no estudo referente aos dados disponíveis na internet apresenta uma banda na região visível entre 400 e 800nm. Na região entre 800 e 1.000nm está a região

característica do terceiro sobretom de CH, CH_2 e CH_3, R-OH, segundo sobretom O-H e contribuição da absorção da molécula de água. Em torno do comprimento de onda 1200nm está a região de segundo sobretom de CH, CH_2 e CH_3, absorção da molécula de água, e primeiro sobretom O-H. A região compreendida entre os comprimentos de onda 1400 a 1600nm é referente ao primeiro sobretom de CH, CH_2 e CH_3, combinações de O-H, e contribuição da absorção da molécula de água. Principalmente a absorção referente aos sobretons de CH, CH_2 e CH_3 e O-H, podem ser atribuídas aos açúcares e como representante majoritário a sacarose[53].

4.3.1 Kernel-PLS

Assim como feito para os dados simulados, inicialmente uma otimização dos valores de VL e σ foi realizada para grau brix e açúcares totais. Sendo escolhidas respectivamente 39 e 49 VLs e valor de σ = 0,500 para ambos. Os modelos *full* Kernel-PLS foram então construídos utilizando os valores obtidos na otimização, a partir das amostras do conjunto de calibração, e em seguida o modelo foi utilizado para a predição da concentração das amostras de validação. São apresentados na Figura 4.20 os gráficos da EJCR e de valores preditos *versus* valores observados para a validação externa dos modelos *full*.

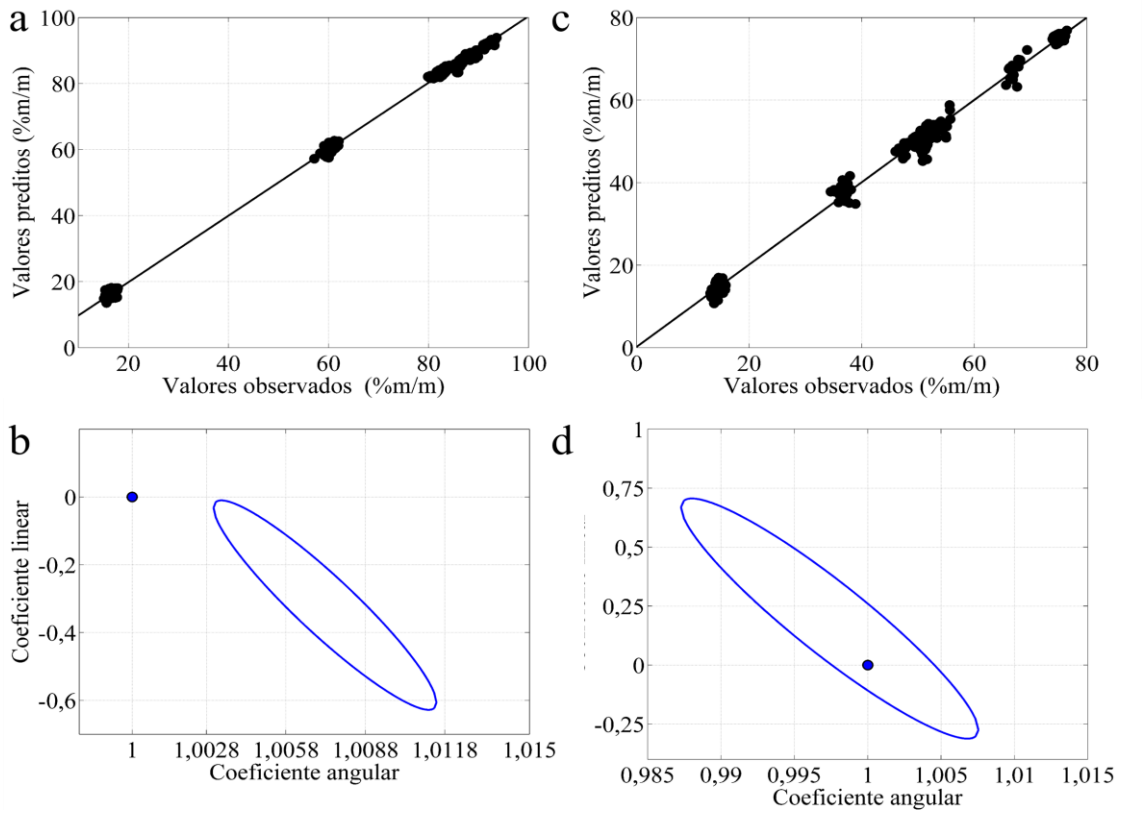

Figura 4.20 - Parâmetros de validação do modelo Kernel-PLS para as amostras de validação externa (reta de ajuste dos valores preditos *versus* observados por OLS para grau brix (a) e açúcares totais (c), elipse EJCR para grau brix (b) e açúcares totais (d)).

O modelo construído para grau brix apesar de ter apresentado RMSEV no valor de 0,8704 não contem, de acordo com a elipse de confiança, o ponto ideal, ou seja, o modelo contém *bias* significativo, enquanto que para açúcares totais a elipse contém o ponto ideal e o valor de RMSEV do modelo é de 1,3942. Em relação ao gráfico de valores observados *versus* valores preditos, os dois modelos apresentam, de modo geral, amostras próximas a reta de ajuste, sendo o modelo para grau brix visualmente mais próximos da reta, o que reflete no menor valor de RMSEV em relação ao de açúcares totais. Os valores de R^2 são de, respectivamente, 0,999 e 0,994 para grau brix e açúcares totais.

Os modelos foram em seguida utilizados para prever o grau brix e o de açúcares totais nas amostras do conjunto de predição. Na Figura 4.21 são apresentados os gráficos dos

valores preditos *versus* referência e elipse de confiança para o conjunto de amostras de predição para grau brix e açúcares totais.

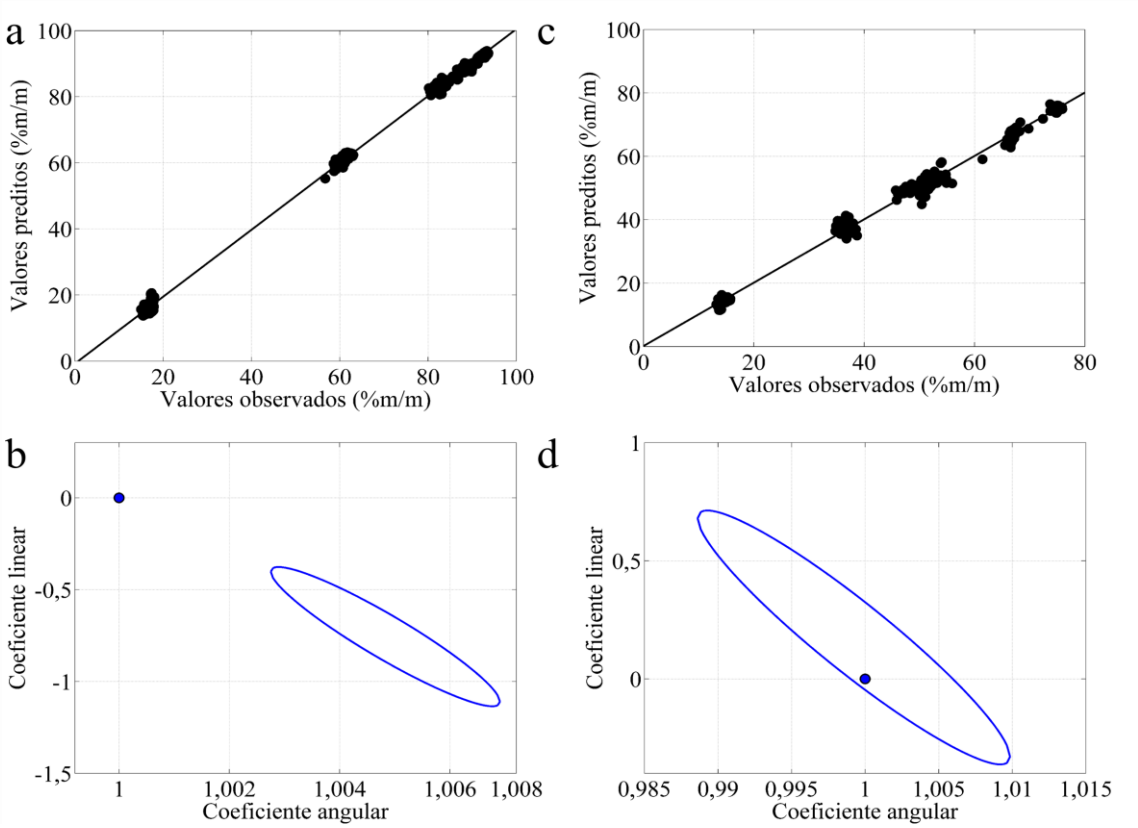

Figura 4.21 - Parâmetros de predição do modelo Kernel-PLS (reta de ajuste dos valores preditos *versus* observados por OLS para grau brix (a) e açúcares totais (c), elipse EJCR para grau brix (b) e açúcares totais (d)).

O erro na predição do grau brix obtido para o conjunto de predição mostrou-se maior que o erro observado para conjunto de validação, RMSEP = 1,0505, além dos valores de R^2 = 0,998 e REP = 1,54%. E para açúcares totais o erro do conjunto de predição se mostrou menor que o erro do conjunto de validação, RMSEP = 1,3712, além dos valores de R^2 = 0,994 e REP = 2,88%. Assim como ocorreu para o conjunto de validação, no conjunto de predição a elipse construída para a determinação do grau brix não contém o ponto ideal, e a elipse de açúcares totais contém. Da mesma forma, no gráfico valores preditos *versus* valores observados as amostras parecem mais bem modeladas no modelo para grau brix que no modelo para açúcares totais.

4.3.2 *Interval* Algoritmo das Projeções Sucessivas – Kernel - Mínimos Quadrados Parciais

De modo similar ao que foi feito para a construção dos modelos iSPA-Kernel-PLS, para a construção dos modelos para grau brix e açúcares totais foram realizadas otimizações a cada avaliação das cadeias na fase 2 do SPA, para otimizar os valores de σ e VL. Foi construído um modelo para cada *W* perfazendo 10 modelos. Em relação aos modelos para grau brix, o que apresentou melhor desempenho foi dividindo o espectro em 4 intervalos e sendo 2 selecionados, perfazendo uma quantidade de 375 variáveis originais, e diminuindo uma variável latente em relação do modelo Kernel-PLS. Para açúcares totais o modelo em que o espectro foi dividido em 7 intervalos e 2 foram selecionados. Este modelo contém uma quantidade de 214 variáveis originais e 45 VLs. Ambos os modelos obtiveram como ideal o valor de σ = 0,500.

Os intervalos selecionados para a construção do modelo iSPA-Kernel-PLS por validação externa para grau brix e açúcares totais são apresentados na Figura 4.22.

Os intervalos selecionados para construção do modelo para quantificação do grau brix estão compreendidos, nas regiões do segundo sobretom de CH e CH_2, e de primeiro sobretom de combinações C-H e R-OH presentes nas estruturas moleculares dos açúcares. Os intervalos selecionados para construção do modelo para quantificação do teor de açúcares totais em termos de sacarose, estão compreendidos na região de terceiro sobretom de CH, CH_2 e CH_3 e de segundo sobretom de CH, CH_2 e CH_3, além de primeiro sobretom da combinação C-H [53] correspondentes à própria estrutura molecular da sacarose. Outro fato a ser considerado é o fato de que a contribuição das bandas características de absorção da água apresenta primeiro e segundo sobretom ao longo da região selecionada, tanto para grau brix quanto para teor de açúcares totais em termos de sacarose [53].

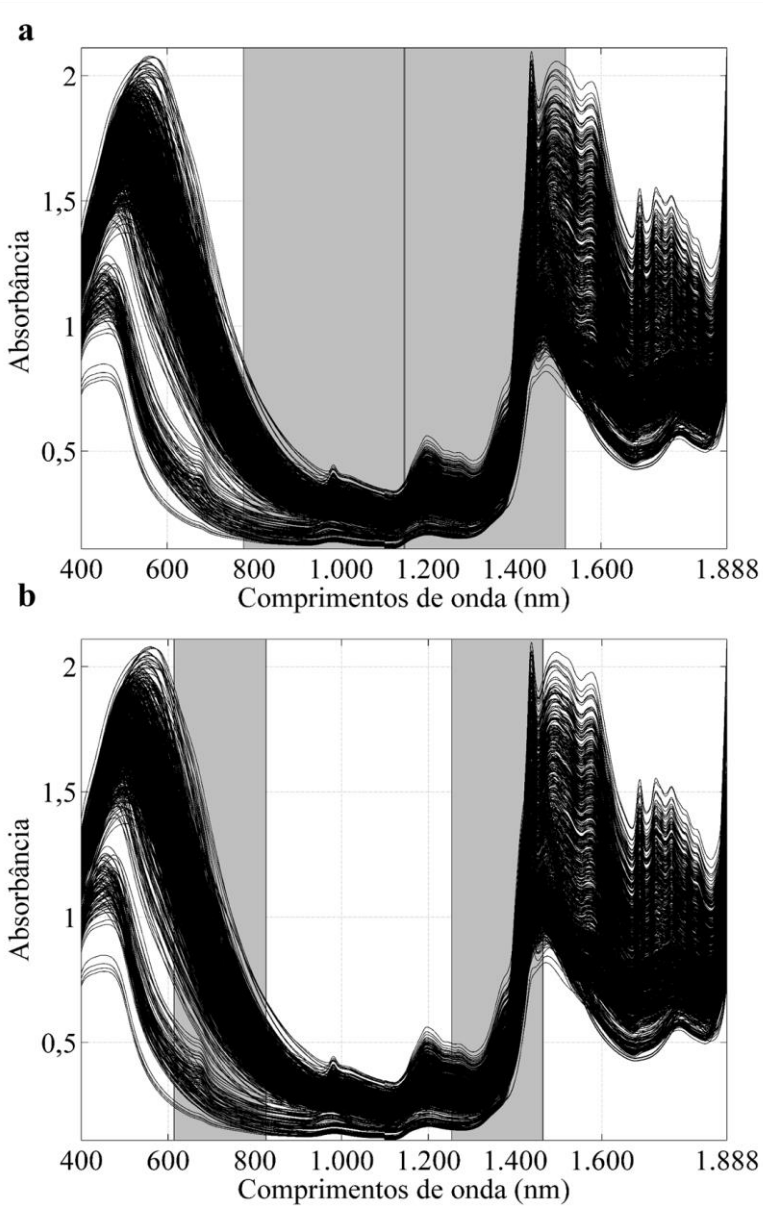

Figura 4.22 - Intervalos selecionados pelo iSPA-Kernel-PLS para o modelo 4(2) grau brix (a) e o modelo 7(2) açúcares totais (b).

Na Figura 4.23 estão apresentados os gráficos de valores observados *versus* valores preditos, e elipses de confiança para os modelos 4(2) e 7(2) para os conjuntos de validação de grau brix e açúcares totais respectivamente.

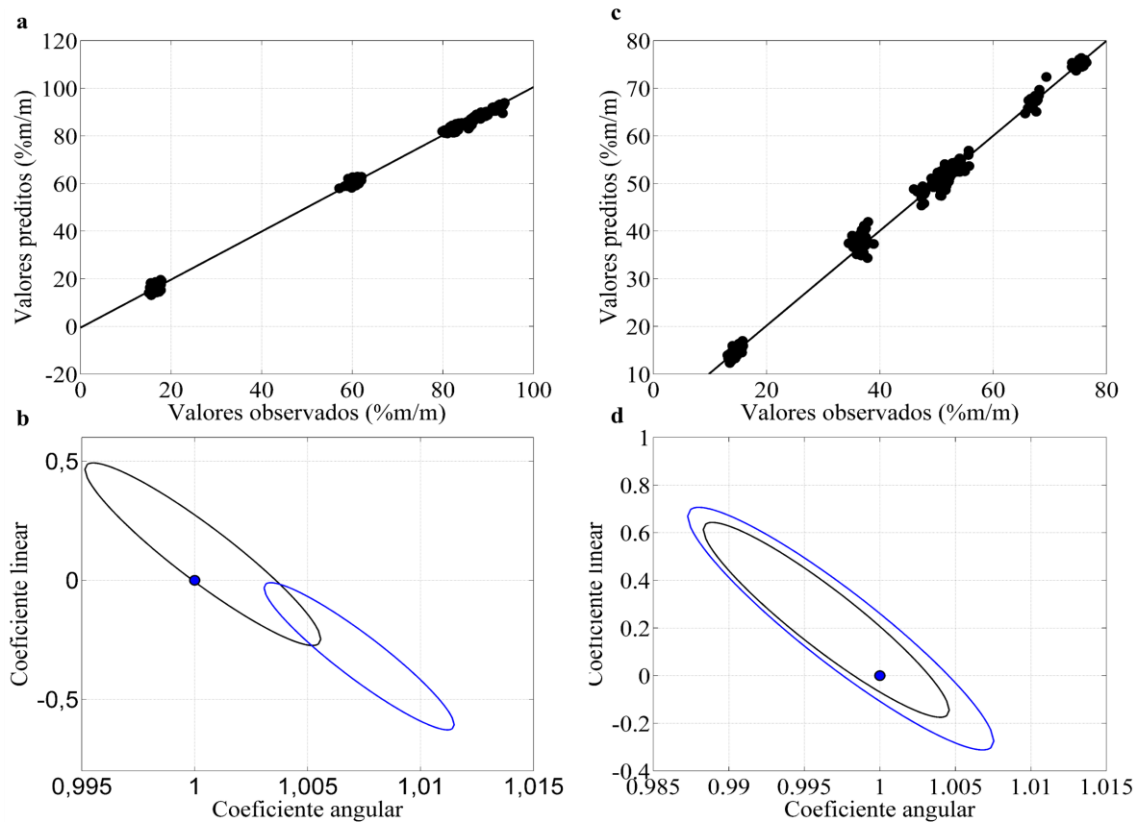

Figura 4.23 - Parâmetros de validação do modelo iSPA-Kernel-PLS (reta de ajuste dos valores preditos *versus* observados por OLS para grau brix (a) e açúcares totais (c), elipse EJCR para grau brix (b) e açúcares totais (d) (▬ Kernel-PLS; ▬ iSPA-Kernel-PLS)).

A validação dos modelos iSPA-Kernel-PLS tanto para grau brix quanto para açúcares totais se mostraram melhor ajustados que o modelo *full*, apesar do valor de RMSEV para grau brix ser levemente maior 1,0371, este em relação ao método EJCR contém o ponto ideal no interior de sua elipse, diferentemente do que ocorreu com o Kernel-PLS. Já em relação a açúcares totais o seu valor de RMSEV=1,0804 mostrou-se menor que o *full*. Em termos de R^2 o modelo para grau brix obteve o valor de 0,998 e o de açúcares totais 0,996.

Com base nos resultados de validação apresentados tanto para grau brix quanto para açúcares totais, pode-se afirmar que a estratégia de seleção de intervalos mostrou-se uma ferramenta bastante importante para a calibração não linear, uma vez que ao restringir a construção dos modelos utilizando apenas as variáveis realmente importantes para a explicação do objeto de estudo, o modelo foi otimizado e melhorado, uma vez que deixou

de apresentar tendências significativas, dentro do intervalo de confiança delimitado pela elipse para o conjunto de validação, no caso do grau brix.

Uma vez que os modelos iSPA-Kernel-PLS mostraram-se adequados estes foram utilizados para prever o grau brix e açúcares totais para as amostras de predição. Na Figura 4.24 são apresentados os gráficos EJCR e o de valor observado *versus* valores preditos.

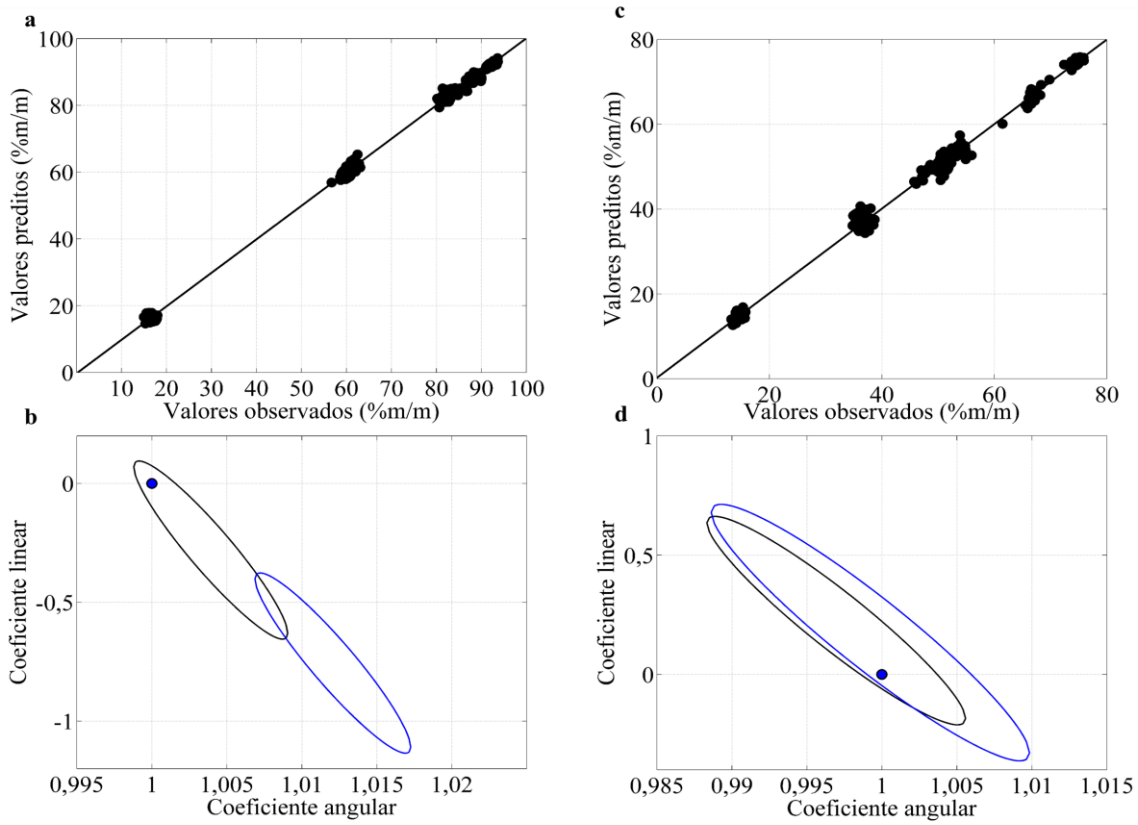

Figura 4.24 - parâmetros de predição do modelo iSPA-Kernel-PLS (reta de ajuste dos valores preditos *versus* observados por OLS para grau brix (a) e açúcares totais (c), elipse EJCR para grau brix (b) e açúcares totais (d).

Como havia sido verificado na avaliação do desempenho do modelo para o conjunto de validação, na predição os modelos baseados na seleção de intervalos se mostraram mais adequados para a aplicação que o método baseado na utilização do espectro completo. O RMSEP obtido por meio do iSPA-Kernel-PLS para a predição do grau brix e de açúcares totais foram respectivamente 0,9932 e 0,9403, contra 1,0505 e 1,3712 para

o método Kernel-PLS. Além disso, valores de R^2 para grau brix e açúcares totais foram respectivamente 0,998 e 0,997. Em termos percentuais os valores de REP foram 1,45% e 1,97% para grau brix e açúcares totais, respectivamente.

Capítulo 5

Conclusões

5 CONCLUSÕES

O iSPA-Kernel-PLS proposto neste trabalho é uma generalização do SPA para seleção de intervalos na calibração linear. Os benefícios verificados na seleção de intervalos na calibração linear já consolidados na literatura, também foram verificados na prática, para a calibração não linear, mais precisamente em Kernel–PLS.

Para os três casos em questão, variáveis pouco informativas não foram selecionadas e houve aumento expressivo na parcimônia dos modelos quando comparados à técnica Kernel-PLS. De modo geral, nos estudos de caso, a seleção de intervalos foi benéfica para a construção dos modelos não lineares. Mesmo em casos onde o Kernel-PLS já apresenta resultados muito bem ajustados.

Nos casos em que os resultados foram equivalentes é mais sensato optar pelos modelos com seleção de variáveis, uma vez que o esforço computacional e o tempo de análise são evidentemente reduzidos. Após a seleção dos intervalos, equipamentos dedicados baseados por exemplo em LEDs, podem ser construídos para a faixa reduzida pelo algoritmo, ou medidas apenas na faixa selecionada em equipamentos comerciais, ganhando em tempo de análise e economia na utilização dos equipamentos, além de menor esforço computacional para o tratamento dos dados.

6 REFERÊNCIAS

[1] A. Dankowska, A. Domagala, W. Kowalewski. Quantification of *Coffea arabica* and Coffea canephora var. robusta concentration in blends by means of synchronous fluorescence and UV-Vis spectroscopies, Talanta. 172 (2017) 215-220.

[2] G. A. Petrucelli, R. J. Poppi, R. L. Mincato, E. R. P. Filho, TS-FF-AAS and multivariate calibration: A proposition for sewage sludge slurry sample analyses, Talanta, 71 (2007) 620 – 626.

[3] S. Zhu, B. Chen, M. He, T. Huang, B. Hu, Speciation of mercury in water and fish samples by HPLC-ICP-MS after magnetic solid phase extraction, Talanta, 171 (2017) 213 – 219.

[4] M. Sirajuddin, S. Ali, A. Badshah, Drug–DNA interactions and their study by UV–Visible, fluorescence spectroscopies and cyclic voltammetry, Journal of Photochemistry and Photobiology B: Biology, 124 (2013) 1-19.

[5] N. Boaz, R. R Coifman, the Prediction Error in CLS and PLS: The Importance of Feature Selection Prior to Multivariate Calibration. Journal of Chemometrics. 19 (2005) 107 – 118.

[6] M. Forina, S. Lanteri, M. C. C. Oliveros, C. P. Millan, Selection of useful predictors in multivariate calibration. Analytical and Bioanalytical Chemistry380 (2004) 397 – 418.

[7] R. Leardi, L. Nørgaard, Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions. Journal of Chemometrics 18 (2004) 486 – 497.

[8] S. Wold, M. Sjöström, L. Eriksson, PLS-Regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 58 (2001) 109 – 130.

[9] C. H. Spiegelman, M. J. McShane, M. J. Goetz, M. Motamedi, Q. L. Yue. G. L. Coté, Theoretical justification of wavelength selection in PLS calibration: development of a new algorithm. Analytical Chemistry 70 (1998) 35 – 44.

[10] J. A. F. Pierna, O. Abbas, V. Beaten, P. Dardenne, A backward variable selection method for PLS regression (BVSPLS). Analytica Chimica Acta 642 (2009) 89 – 93.

[11] F. Allegrini, A. C. Olivieri, a new and efficient variable selection algorithm based on ant colony optimization. applications to near infrared spectroscopy/partial least-squares analysis. Analytica Chimica Acta 699 (2011) 18 - 25.

[12] L. Nørgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, S. B. Engelsen, Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. Applied Spectroscopy 54 (2000) 413 – 419.

[13] A. A. Gomes, R. K. H. Galvão, M. C. U. Araújo, G. Véras, E. C. Silva, The successive projections algorithm for interval selection in PLS, Microchemical Journal 110 (2013) 202 – 208.

[14] A. A. Gomes, M. R. Alcaraz, H. C. Goicoechea, M. C. U. Araújo, The successive projections algorithm for interval selection in trilinear partial least-squares with residual bilinearization. Analytica Chimica Acta 811 (2014) 13 – 22.

[15] R.K.H. Galvão, M.C.U. Araújo, in: B. Walczak, R. Tauler, S. Brown (Eds.), Comprehensive Chemometrics, vol. 3, Elsevier, Oxford, 2009, pp. 233–283.

[16] M. C. U. Araújo, T. C. B. Saldanha, R. K. H. Galvão, T. Yoneyama, H. C. Chame, V. Visani, The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemometrics and Intelligent Laboratory Systems 57 (2001) 65 – 73.

[17] S. F. C. Soares, A. A. Gomes, A. R. G. Filho, M. C. U. Araújo, R. K. H. Galvão, The successive projections algorithm. trends analytical chemistry 42 (2013) 84 – 98.

[18] F. A. Honorato, R. K. H. Galvão, M. F. Pimentel, B. B. Neto, M. C. U. Araújo, F. R. Carvalho, Robust modeling for multivariate calibration transfer by the successive projections algorithm. Chemometrics and Intelligent Laboratory Systems 76 (2005) 65 – 72.

[19] H. A. D. Filho, R. K. H. Galvão, M. C. U. Araújo, E. C. Silva, T. C. B. Saldanha, G. E. José, C. Pasquini, I. M. Raimundo Jr, J. J. R. Rohwedder, A strategy for selecting calibration samples for multivariate modelling. Chemometrics and Intelligent Laboratory Systems 72 (2004) 83 – 91.

[20] M. Ghasemi-Varnamkhasti, S.S. Mohtasebi, M. L. Rodriguez-Mendez, A. A. Gomes, M. C. U. Araújo, R. K. H. Galvão, Screening analysis of beer ageing using near infrared spectroscopy and the successive projections algorithm for variable selection. Talanta. 89 (2012) 286 – 291.

[21] K. Wang, T. Chen, R. Lau, bagging for robust non-linear multivariate calibration of spectroscopy. Chemometrics and Intelligent Laboratory Systems 105 (2011) 1 – 6.

[22] J. C. L. Alves, R. J. Poppi, Quantification of conventional and advanced biofuels contents in diesel fuel blends using near-infrared spectroscopy and multivariate calibration. Fuel. 165 (2016) 379 – 388.

[23] W. Ni, L. Nørgaard, M. Mørup, Non-Linear calibration models for near infrared spectroscopy. Analytica Chimica Acta 813 (2014) 1-14.

[24] M. Wang, G. Yan, Z. fei, Kernel PLS based prediction model construction and simulation on theoretical cases. Neurocomputing. 165 (2015) 389 – 394.

[25] H. Shinzawa, P. Ritthiruangdej, Y. Ozaki, Kernel analysis of partial least squares (PLS) regression models. Applied Spectroscopy 65 (2011) 549 – 556.

[26] D. S. Cao, Y. Z. Liang, Q. S. Xu, Q. N. Hu, L. X. Zhang, G. H. Fu, Exploring nonlinear relationships in chemical data using kernel-based methods. Chemometrics and Intelligent Laboratory Systems 107 (2011) 106 – 115.

[27] N. Benoudjit, E. Cools, M. Meurens, M. verleysen, Chemometric Calibration of infrared spectrometers: selection and validation of variables by non-linear models. Chemometrics and Intelligent Laboratory Systems 70 (2004) 47 – 53.

[28] Q. Chen, J. Ding, J. Cai, J. Zhao, Rapid measurement of total acid content (TAC) in vinegar using near infrared spectroscopy based on efficient variables selection algorithm and nonlinear regression tools. Food Chemistry. 135 (2012) 590 – 595.

[29] C. Tan, M. Li, Mutual information-induced interval selection combined with kernel partial least squares for near-infrared spectral calibration. Spectrochimica Acta A. 71 (2008) 1266 – 1273.

[30] J. Lee, K. Chang, C. H. Jun, R. K. Cho, H. Chung, H. Lee, Kernel-Based calibration methods combined with multivariate feature selection to improve accuracy of near-

infrared spectroscopic analysis. Chemometrics and Intelligent Laboratory Systems 147 (2015) 139 – 146.

[31] D. A. E. Skoog, J. J. Leary, Principles of Instrumental Analysis. 6ed New York: Saunders College Publishing, 1992.

[32] G. M. Escandar, A. C. Olivieri, Pratical three-way calibration. Elsevier, 2014.

[33] E. Besalú, The connection between inverse and classical calibration, Talanta, 116 (2013) 45 – 49.

[34] K. R. Beebe, R. J. Peel, M. B. Seasholtz, Chemometrics: A Pratical Guide, New York, John Wiley & Sons, 1998.

[35] A. A. Gomes, Algoritmo das projeções sucessivas aplicado à seleção de variáveis em regressão PLS. Dissertação de Mestrado, João Pessoa, UFPB, 2012.

[36] R. W. Kennard, L. A. Stone, Computer aided design of experiments, Technometrics, 11 (1969) 137 – 148.

[37] R. K. H. Galvão. M. C. U. Araújo, G. E. josé, M. J. C. Pontes, E. C. Silva, T. C. B. Saldanha, A method for calibration and validation subset partitioning, Talanta, 67 (2005) 736 – 740.

[38] S. Wold, J. Trygg, A. Berglund, H. Antti, Some recent developments in PLS modeling. Chemometrics and Intelligent Laboratory Systems 58 (2001) 131–150.

[39] S. Wold, M. Sjöström, L. Eriksson. PLS. regression: A basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 58 (2001), 109–130.

[40] R. G. Brereton, Chemometrics: data analysis for the laboratory and chemical plant. New York: John Wiley & Sons, 2003.

[41] H. Martens, T. Naes, Multivariate calibration, John Wiley: New York, 1989.

[42] L. Lathauwer, N. Moor, J. Vandewallet, Journal of Chemometrics 14 (2000) 123

[43] P. Geladi, B. R Kowalski, Partial least square: a tutorial, Analytica Chimica Acta, 185 (1986) 1-17.

[44] B. Walczaka, D. L. Massart, The Radial basis functions - partial least squares approach as a flexible non-linear regression technique. Analytica Chimica Acta 331 (1996) 177 – 185.

[45] A. García-Reiriz, P. C. Damiani, A. C. Olivieri, Residual bilinearization combined with kernel-unfolded partial least-squares: a new technique for processing non-linear second-order data achieving the second-order advantage. Chemometrics and Intelligent Laboratory Systems 100 (2010) 127 – 135.

[46] P. Alderrama, J. W. B. Braga, R. J. Poppi, Estado da arte das figuras de mérito em calibração multivariada. Quimica Nova, 32 (2009) 1278 – 1287.

[47] A. G. Gonzalez, M. A. Herrador, A. G. Asuero, Intra – Laboratory testing of method accuracy from recovery assays, Talanta, 48 (1999) 729 – 736.

[48] W. Li, F. S. C. Lee, X, Wang, Y. He, 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 (2006) 529 – 536.

[49] Annual Book of ASTM Standards; Standards pratices for infrared, multivariate, quantitative analisys, E1655, vol 03.06, ASTM international: West Conshohocken 2000.

[50] A. C. Olivieri, H. C, Goicoechea, F. A. Iñón, MVC1: an integrated MatLab toolbox for first-order multivariate calibration. Chemometrics and Intelligent Laboratory Systems 73 (2004) 189 – 197.

[51] Z. Xiaobo, Z. Jiewen, M. J. W. Povey, M. Holmes, M. Hanpin, Variables selection methods in near-infrared spectroscopy. Analytica Chimimica. Acta 667 (2010) 14 – 32.

[52] H. Van Der Voet, Comparing the predictive accuracy of models using a simple randomization test. Chemometrics and Intelligent Laboratory Systems 25 (1994) 313 – 323.

[53] Z. Xiaobo, Z. Jiewen, M. J. W. Povey, M. Holmes, M. Hanpin, Variables Selection Methods in Near-Infrared Spectroscopy. . Analytica Chimica Acta 667 (2010) 14 – 32.

[54] R. Tange, M. A. Rasmussen, E. Taira, R. Bro, Application of support vector regression for simultaneous modelling of near infrared spectra from multiple process steps. Journal of Near Infrared Spectroscopy. 23 (2015) 75 – 84.

Anexo

ANEXO 1 DADOS EXPERIMENTAIS DE DOMÍNIO PÚBLICO

Os dados de espectroscopia NIR foram obtidos a partir de quatro etapas na fabricação de açúcar: moagem (suco), evaporação (xarope), cristalização (massa cozida) e centrifugação (melaço), Figura a1, onde foram determinados o grau brix e o teor de açúcares totais, em percentagem de massa [54].

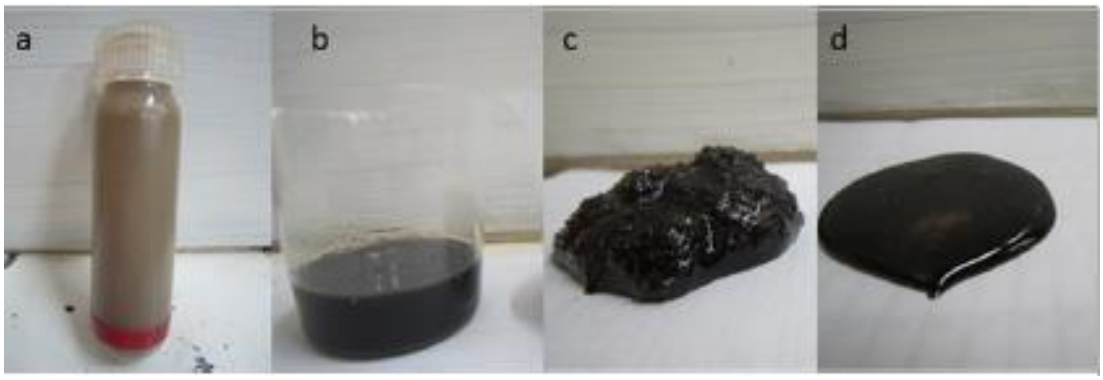

Figura a1 - Fotografias de amostras em cada umas das etapas (a) moagem (suco), (b) evaporação (xarope), (c) cristalização (massa cozida) e (d) centrifugação (melaço) [54].

3.3.1 Aquisição das amostras

As amostras são provenientes de uma fábrica de açúcar japonesa (Daito Tokyo Co.), coletadas pelos autores durante o processo de produção do açúcar, a partir da cana de açúcar, em cada uma das quatro etapas. As medidas foram feitas imediatamente após a coleta, na temperatura em que se encontravam durante o processo. O referido banco de dados contém efeitos de correlação não linear entre as propriedades de interesse e o sinal analítico medido, devido a alterações nas propriedades físicas e químicas das amostras, como: viscosidade, temperatura, pH, composição química e grau de cristalização, inerentes ao processo industrial de produção [54].

3.3.2 Obtenção dos espectros NIR

As medidas de transflectância das 1.797 amostras foram realizadas na região espectral Vis-NIR compreendida entre 400 e 1.888 nm, com resolução espectral de 2 nm, utilizando

um espectrômetro NIR (DS2500, FOSS AB, Hilleroed, Dinamarca), equipado com um suporte de amostra e uma placa refletora de 0,5 mm de espessura (refletor de ouro, Foss Co. LTd). Foram utilizados cerca de 5 mL de cada amostra para a aquisição dos espectros.

3.3.3 Obtenção dos parâmetros de referência

Os parâmetros de referência grau brix e açúcares totais foram medidos em termos de teor de sólidos dissolvidos e teor de sacarose respectivamente. As medidas referentes ao grau brix foram feitas utilizando um refratômetro (ABBEMAT-WR, Anton Paar GmbH, Germany) e açúcares totais foi medido utilizando um polarímetro (MCP500, Anton Paar GmbH, Germany). As medidas das amostras de xarope, massa cozida e melaço foram obtidas após uma diluição de 5 ou 6 vezes, com água.