UNIVERSIDADE FEDERAL DA PARAÍBA

CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA

DEPARTAMENTO DE QUÍMICA

PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

TESE DE DOUTORADO

ALGORITMO DAS PROJEÇÕES SUCESSIVAS PARA SELEÇÃO DE VARIÁVEIS EM CALIBRAÇÃO DE SEGUNDA ORDEM

Adriano de Araújo Gomes

João Pessoa-PB; Junho/2015

Adriano de Araújo Gomes

ALGORITMO DAS PROJEÇÕES SUCESSIVAS PARA SELEÇÃO DE VARIÁVEIS EM CALIBRAÇÃO DE SEGUNDA ORDEM

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

1º Orientador: Prof. Dr. Mário César Ugulino de Araújo (UFPB)

2º Orientador: Prof. Dr. Héctor Casimiro Goicoechea (UNL)

Bolsista: CAPES

João Pessoa-PB; Junho/2015

G633a Gomes, Adriano de Araújo.
Algoritmo das projeções sucessivas para seleção de variáveis em de calibração de segunda ordem / Adriano de Araújo Gomes.- João Pessoa, 2015.
126f. : il.
Orientadores: Mário César Ugulino de Araújo, Héctor Casimiro Goicoechea
Tese (Doutorado) - UFPB/CCEN
1. Química analítica. 2. Seleção de intervalos. 3. Dados multivias. 4. Filtro interno - efeito. 5. Vantagem de segunda ordem.

UFPB/BC CDU: 543(043)

ALGORITMO DAS PROJEÇÕES SUCESSIVAS PARA SELEÇÃO DE VARIÁVEIS EM CALIBRAÇÃO DE SEGUNDA ORDEM

Tese de Doutorado apresentada pelo aluno Adriano de Araújo Gomes e aprovada pela banca examinadora em 29 de junho de 2015.

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

Prof. Dr. Hector Casimiro Goecoechea
2º. Orientador

Prof. Dr. Kássio Michell Gomes de Lima
Examinador

Prof. Dr. José Germano Veras Neto
Examinador

Prof ª. Dr ª. Liliana de Fátima Bezerra Lira e Pontes
Examinadora

Prof. Dr. Edvan Cirino da Silva
Examinador

Pode parecer óbvio, pode mesmo ser difícil, mas é muito importante. Encerrando ciclos. Não por causa do orgulho, por incapacidade, ou por soberba, mas porque simplesmente aquilo já não se encaixa mais na sua vida. Feche a porta, mude o disco, limpe a casa, sacuda a poeira. Deixe de ser quem era, e se transforme em quem é. Torna-te uma pessoa melhor e assegura-te de que sabes bem quem és tu próprio. E lembra-te: Tudo o que chega, chega sempre por alguma razão.

(Fernando Pessoa)

Com a mais profunda gratidão aos meus pais,

Dedico...

Agradecimentos

- Em primeiro lugar, ao grande DEUS, pelo seu infinito amor e bondade.
- Aos meus pais João Batista e Cícera, por todo amor dedicado ao longo da minha vida.
- Aos meus irmãos Tânia, Vanderli, Francisca e Fernando.
- A toda família, por ter acreditado em mim e ter dado todo apoio e incentivo para prosseguir.
- Ao Prof. Dr. Mário Cesar Ugulino de Araújo (Coord. do LAQA), por ter me recebido no LAQA e pelas várias oportunidades de crescimento profissional concedida e orientação deste trabalho.
- Ao Prof. Dr. Hector Goicoechea, pela incrível oportunidade de crescimento profissional que tive durante oito meses em seu grupo de pesquisa e orientação deste trabalho.
- Aos amigos que fiz no LADAQ/UNL.
- Aos Professores do PPGQ-UFPB pelos ensinamentos durante as disciplinas cursadas.
- Aos orientadores de outras etapas de minha formação, Prof. Dr. Germano Váras e Edvan Cirino.
- Aos amigos que vão além do lado profissional e fazem parte da minha vida.
- Aos amigos que ganhei no LAQA durante todo esse tempo.
- Ao PPGQ-UFPB e a Capes pela bolsa concedida.

Sumário

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SEGLAS E ABREVIATURAS

γ Sensibilidade Analítica.

AAS- (Ácido Acetil Salicílico).

ALS- (Mínimos Quadrados Alternados, do inglês: "*Alternative Least Square*").

BLLS- (Mínimos Quadrados Bilineares, do inglês: "*Bilinear Least Square*").

CANDECOMP- (Decomposição Canônica, do inglês: "*Canonical Decomposition*").

CPF- (Ciprofloxacina).

DAD- (Detector de Arranjo de Diodos, do inglês: "*Diode Array Detection*").

DNF- (Danofloxacina).

EEM- (Matriz Excitação Emissão, do inglês: "*Excitation Emission Matrix*").

EFI- (Efeito de Filtro Interno).

FEN- (Fenilefrina).

FFSA- (do inglês: "*Forward Floating Selection algorithm").*

FNNLS- (Mínimos Quadrados Não Negativos Rápido, do inglês: "*Fast Non Negative Least Square*").

GA- (Algoritmo Genético, do inglês: "*Genetic Algorithm*").

CG- (Cromatografia a Gás).

HPLC- (Cromatografia Liquida de Alto Desempenho, do inglês: "*High Performance Liquid Chromatographyc*").

IBU- (Ibuprofeno).

*i*SPA- (Algoritmo das Projeções Sucessivas por intervalos, do inglês: "*intervals Sucssessive Projection Algorithm*").

LC- (Cromatografia a Liquido, do inglês: "*Liquid Chromatographyc*").

LDA- (Análise Discriminante Linear, do inglês: "*Linear Discriminant Analysis*").

L- DOPA- (Ácido (S)-2-amino-3-(3,4-dihidroxifenil)).

LIBIS- (Espectrometria de emissão em plasma induzido por laser, do inglês: "*laser-induced breakdown spectroscopy*").

MCR-(Resolução de Curvas Multivariadas, do inglês: "*Multivariate Curve Resolution*").

MLR- (Regressão Linear Múltipla, do inglês: "*Multiple Linear Regression*").

MS- (Espectrometria de Massa, do inglês: "*Mass Spectrometry*").

MVC2- (Calibração Multivariada 2, do inglês: "*Multivariate Calibration* 2").

NIR- (Infra Vermelho Próximo, do inglês: "*Near Infra red*").

N-*i*SPA- (Algoritmo das Projeções Sucessivas por intervalos multidimensionais, do inglês: "*Multiway intervals Sucssessive Projection Algorithm*").

NIPLAS- (mínimos quadrados parciais iterativo não linear, do inglês: "*non-linear iterative partial least squares"*).

NNLS- (Mínimos Quadrados Não Negativos, do inglês: "*Non Negative Least Square*").

N-PLS- (Minimos Quadrados Parciais Multidimencional, do inglês: "*Multiway Partial Least Square*").

OFL- (Ofloxacina).

OLS- (Mínimos Quadrados Ordinários, do inglês: "*Ordinary Least Square*").

OPA- (método de projeções ortogonais, do inglês: "*Ortogonal Projection Approche*").

PAR- (Paracetamol).

PARAFAC- (Análise de Fatores Paralelos, do inglês: do inglês: "*Palallel Factor Analysis*").

PARALIND- (Análise de Fatores Paralelos com Dependencia Linear, do inglês: "*Palallel Factor Analysis with Linear Dependence*").

pH- (Potêncial Hidrogênionico) .

PCR- (Regressão por Componentes Princiapis, do inglês: "*Principal Component Regression*").

PLS- (Minimos Quadrados Parciais, do inglês: "*Partial Least Square*").

PRESS- (Soma Quadrada do Erro de Predição, do inglês: "*Prediction Erros Sum Square*").

QSAR - (Relação Quantitativa Atividade Estrutura, do inglês: "*Quantitative Structure-Activity Relationship*").

QSPR- (Relação Quantitativa Propriedade Estrutura, "*Quantitative Structure- Property Relationship*").

RBL- (Bilinearização Residual, do inglês: "*Residual Bilinearization*").

RMSE- (Raiz do Erro Médio Quadrático, do inglês: "*Root Mean Square Erros*").

RMSEC-(Raiz do Erro Médio Quadrático de Calibração, do inglês: "*Root Mean Square Erros of Calibration*").

RMSECV-(Raiz do Erro Médio Quadrático de Validação Cruzada, do inglês: "*Root Mean Square Erros of Cross-Validation*").

RMSEV-(Raiz do Erro Médio Quadrático de Valideação, do inglês: "*Root Mean Square Erros of Validation*").

RMSEP- (Raiz do Erro Médio Quadrático de Predição, do inglês: "*Root Mean Square Erros of Prediction*").

SFM- (Matriz de Fluorescência Sincrônica, do inglês: "*Sycronous Florensence Matrix*").

SIMCA- (Modelagem Independente Flexível por Analogia de Classe, do inglês: "*Soft Independent Modelling Class Analogy*").

(SPA- Algoritmo das Projeções Sucessivas, do inglês: "*Sucssessive Projection Algorithm*").

SPXY- (Partição de Amostras X-Y, do inglês: "*Samples Partition X-Y*").

SEN- Sensibilidade.

SVMR- (Regressão em Máquinas de Suporte de Vetores, do inglês: "*Support Vector Machine Regression*").

U-PLS-(Mínimos Quadrados Parciais desdobrados, do inglês: "*Unfolded Partial Least Square*").

UV-Vis- (Ultravioleta e Visível "*Ultraviolet and Visible*").

VIP- (Importância da Variável na Projeção, do inglês: "*Variable Importance Porjection*").

RESUMO

Neste trabalho foi desenvolvida uma nova estratégia para seleção de intervalos empregando o algoritmo das projeções sucessivas (SPA) acoplado a modelos N-PLS e U-PLS, ambos com etapa pós-calibração de bilinearização residual (RBL). O novo algoritmo acoplado a modelos N-PLS/RBL, foi avaliado em dois estudos de casos. O primeiro envolvendo dados simulados para quantificação de dois analitos (A e B) na presença de um único interferente. No segundo foi conduzida a quantificação de ofloxacina em água na presença de interferentes (ciprofloxacina e danofloxacina) por meio da modelagem de dados cromatografia liquida com detecção por arranjo de diodos (LC-DAD). Os resultados obtidos foram comparados ao modelo N-PLS/RBL e a seleção de variáveis com o algoritmo genético (GA-N-PLS/RBL). No primeiro estudo de caso (dados simulados) foram observados valores de RMSEP (x 10^{-3} em unidades arbitrárias) para os analitos A e B da ordem de 6,7 e 47,6; 10,6 e 11,4; 6,0 e 14,0 para o N-PLS/RBL, GA-N-PLS/RBL e o método proposto, respectivamente. No segundo estudo de caso (dados HPLC-DAD) valores de RMSEP (em mg/L) de 0,72 (N-PLS/RBL); 0,70 (GA-N-PLS/RBL) e 0,64 (*i*SPA-N-PLS/RBL) foram obtidos. Quando combinado com o U-PLS/RBL o novo algoritmo foi avaliado na modelagem de EEM em presença efeito de filtro interno. Dados simulados e a quantificação de fenilefrina na presença de paracetamol em amostras de água e interferentes (Ibuprofeno e ácido acetil salicílico) foram usados como estudos de caso. Os resultados obtidos foram comparados ao modelo U-PLS/RBL e ao bem estabelecido método PARAFAC. Para dados simulados foram observado os seguintes valores de RMSEP (em unidades arbitrarias) 1,584; 0,077 e 0,066 para o PARAFAC; U-PLS/RBL e método proposto, respectivamente. Na quantificação de fenilefrina os RMSEP (em μg/L) encontrados foram de 0,164 (PARAFAC); 0,089 (U-PLS/RBL) e 0,069 (*i*SPA-U-PLS/RBL). Em todos os casos foi demostrado que seleção de variáveis é uma ferramenta útil capaz de melhorar a acurácia quando comparados aos respectivos modelos globais (modelo sem seleção de variáveis) e tornar os modelos mais parcimoniosos. Foi observado ainda para todos os casos, que a perda de sensibilidade promovida pela seleção de variáveis é compensada pelo uso de canais mais seletivos, justificando os menores valores de RMSEP obtidos. E por fim, foi também observado que os modelos baseados em seleção de variáveis como o método proposto foram isentos de bias significativos a 95% de confiança.

Palavras-chave: Seleção de intervalos, dados multivias, efeito de filtro interno, vantagem de segunda ordem.

ABSTRACT

In this work it was developed a new strategy for intervals selection using the successive projections algorithm (SPA) coupled to N-PLS and U-PLS models, both with residual bilinearização (RBL) as a post-calibration step. The new algorithm coupled to N-PLS/RBL models was evaluated in two cases of studies. The first was simulated data for quantitation of two analytes (A and B) in the presence of a single interfering. On the second study was conducted a quantitation of ofloxacin in water in the presence of interferents (ciprofloxacin and danofloxacin) by means of liquid chromatography with diode array detection (LC-DAD) data modeling. The results were compared to the N-PLS/RBL model and the variables selection with the genetic algorithm (GA-N-PLS/RBL). In the first case of study (simulated data) were observed RMSEP values (x 10-3 in arbitrary units) for the analytes A and B in the order of 6.7 to 47.6; 10.6 to 11.4; and 6.0 to 14.0 for the N-PLS/RBL, Ga-N-PLS/RBL and the proposed method, respectively. On the second case of study (HPLC-DAD data) RMSEP value (mg/L) of 0.72 (N-PLS/RBL); 0.70 (GA-N-PLS/RBL) and 0.64 (iSPA N-PLS/RBL) were obtained. When combined with the U-PLS/RBL, the new algorithm was evaluated in the EEM modeling in the presence of inner filter effect. Simulated data and quantitation of phenylephrine in the presence of acetaminophen in water sample and interferences (ibuprofen and acetylsalicylic acid) were used as a case of studies. The results were compared to the U-PLS/RBL and e twell established method PARAFAC. For simulated data was observed the following RMSEP values (in arbitrary units) 1.584; 0.077 and 0.066 for PARAFAC; U-PLS/RBL and the proposed method, respectively. In the quantitation of phenylephrine the found RMSEP (in μg/L) were of 0.164 (PARAFAC); 0.089 (U-PLS/RBL) and 0.069 (ISPA-U-PLS/RBL). In all cases it was shown that variables selection is a useful tool capable of improving accuracy when compared with the respective global models (model without variables selection) leading to more parsimonious models. It was observed in all cases, that the sensitivity loss promoted by variables selection is compensated by using more selective channels, justifying the obtained RMSEP smaller values. Finally, it was also observed that the models based on variables selection such as the proposed method were free from significant bias at 95% confidence.

Keywords: Intervals selection, multiway data, inner filter effect, second order advantage.

Capítulo I

Introdução

1.0 INTRODUÇÃO

1.1 Caracterização geral da problemática e proposta

Os avanços no campo da instrumentação analítica têm permitido a obtenção de uma grande quantidade de informação, em um curto intervalo de tempo, por amostra. Como bons exemplos podem ser citadas as técnicas analíticas hifenadas, tais como, cromatografia a gás ou a líquido com detecção por espectrometria de massa ou massa/massa (CG-MS-MS e LC-MS-MS). Outros exemplos são as técnicas cromatográficas bidimensionais (LC×LC e CG×CG) [1-2]. Ao mesmo tempo em que se alcançam tais avanços no campo da instrumentação, o desenvolvimento de novas metodologias analíticas aponta na direção de menor consumo de reagentes/amostras e também que o método seja preferencialmente não destrutivo e não invasivo, o que caracteriza a ideia de "*química analítica verde*" [3].

Nesta perspectiva, muito se tem avançado com o uso da quimiometria no tratamento de sinal analítico registrado em múltiplos canais (dados multivariados) [4]. Quantificação de espécies de interesse em matrizes complexas como combustíveis [5], amostras biológicas [6] e alimentos [7], por exemplo, empregando na grande maioria dos casos espectrometria no infravermelho próximo (NIR, "*Near Infrared*") de forma não destrutiva e sem necessidade de tratamentos prévios da amostra, são relatados na literatura [6-7]. O tipo de sinal gerado num espectrômetro como um NIR é categorizado como dado de "*primeira ordem*", que consiste de um vetor de informação por amostra, resultando em uma matriz bidimensional **X**, quando é justaposto o sinal de todas as amostras [8].

Os dados de primeira ordem são considerados de fácil aquisição, contudo a modelagem matemática apresenta certas limitações e inconvenientes. A primeira delas é que o conjunto de amostras de calibração deve ser grande o suficiente para garantir bons estimadores dos parâmetros populacionais do modelo construído. Quando se trata de matrizes complexas, os valores da variável dependente (concentração do analito, por exemplo) devem ser determinados por outro método, chamado método de referência, que embora em menor proporção, exige o uso de reagentes com geração de resíduos. Outro inconveniente dos modelos de primeira ordem é que estes podem até indicar a presença de constituintes não modelados, mas não tem exatidão para efetuar predições confiáveis [8-9].

Construir modelos com pequenos conjuntos de calibração empregando soluções padrões para determinação de analitos em matrizes complexas, explorando todo potencial da instrumentação analítica moderna, tem se tornado um desafio. Até o momento as propostas mais bem sucedidas são as que fazem uso dos métodos multimodos ("*multiway methods*") ou também conhecido como calibração de ordem superior [10]. Estes métodos fazem uso de equações matemáticas e estatísticas de maior complexidade para diminuir o esforço de bancada, o consumo de reagentes e etapas de processamento das amostras [11]. O sinal analítico é registrado para uma amostra em que o analito encontra-se na presença de diversos constituintes não modelados e a contribuição do analito é recuperada via algoritmos matemáticos, possibilitando sua quantificação com exatidão. Esta característica é conhecida como "*vantagem de segunda ordem*" [8-10].

Métodos de análise de dados de ordem superior vêm demonstrando seu potencial em diversas aplicações [8-10] e ganhando cada vez mais destaque no cenário da química analítica. Quantificações simultâneas em presença de constituintes não modelados em

matrizes complexas como fluidos biológicos [12], águas [13], alimentos [14] vêm sendo reportadas na literatura. A modelagem de dados multimodos (dependendo do algoritmo) também pode ser capaz de lidar com certas peculiaridades do sistema instrumental e/ou composição da matriz. Problemas como corrimento de pico em cromatografia, deficiência de posto, filtro interno em mediadas de fluorescência molecular são exemplos de inconvenientes que podem ser contornados na análise multimodo [8-10].

Na calibração multivariada de primeira ordem, o uso de técnicas de seleção de variáveis tem ajudado na escolha de sensores mais informativos e no desenvolvimento de modelos que geram melhores resultados. Isso está associado ao fato da seleção de variáveis ser um processo de busca combinatória baseado na minimização de uma dada função de custo [15]. Dado um conjunto de (J) sensores, a tarefa é escolher um subconjunto de variáveis j (como $j < J$) que produza modelos com melhor acurácia, maior simplicidade e robustez [16]. Muitas estratégias para seleção de variáveis são encontradas na literatura, boa parte destas técnicas são baseadas em métodos combinatórios estocásticos.

Seleção de variáveis vai além da busca por valores de erros de predição menores que o modelo global "*full model*". Mas visa também à construção de modelos mais simples e mais interpretáveis, empregando um menor número de parâmetros, em linha com o principio da parcimônia. Como discutido por Seasholtz e Kowalski, se de alguma forma duas estratégias modelam adequadamente um conjunto de dados, o que emprega um menor número de parâmetros terá maior capacidade de generalização ou poder preditivo sobre um conjunto de teste independente [16].

No contexto de calibração multimodos, a seleção de variáveis têm sido pouco explorada, o que se reflete na escassez de trabalhos descritos na literatura. Em quase sua

totalidade os trabalhos que fazem seleção de variáveis em dados multimodos não apresentaram um algoritmo formalmente ou apenas usaram dados multivias no contexto de análise exploratória. Com relação aos modelos de calibração multimodos, os estudos envolvendo seleção de variáveis são ainda iniciais havendo, portanto, um campo vasto a ser explorado.

Assim, neste trabalho é apresentado o desenvolvimento de novas estratégias de seleção de variáveis que combina o Algoritmo das Projeções Sucessivas (SPA) para eliminação de variáveis não informativas, com os métodos de regressão baseados em variáveis latentes, mínimos quadrados parciais multidimensionais (N-PLS) e mínimos quadrados parciais desdobrados (U-PLS). Ambas as estratégias empregam a Bilinearização Residual (RBL) para alcançar vantagem de segunda ordem.

1.2 Objetivos

1.2.1 Objetivos gerais

Desenvolver um método para seleção de variáveis, em forma de intervalos, baseados no Algoritmo das Projeções Sucessivas acoplado ao método de mínimos quadrados parciais em modo desdobrado (U-PLS) e multidimensional (N-PLS), ambos com etapa RBL.

1.2.2 Objetivos específicos

- Desenvolver o algoritmo denominado *i*SPA combinado ao N-PLS/RBL e ao U-PLS/RBL.
- Avaliar o desempenho do *i*SPA-N-PLS/RBL em dados simulados, e dados HPLC-DAD gerados para quantificação do antibiótico ofloxacina (OFL) em amostras de águas.

- Comparar os resultados obtidos com o método proposto aos resultados do método N-PLS/RBL sem seleção de variáveis e ao método N-PLS/RBL combinado com o Algoritmo Genético (GA).
- Avaliar o desempenho do *i*SPA-U-PLS/RBL na seleção de intervalos em matrizes excitação-emissão (EEM) em presença de efeito de filtro interno para dados simulados e na quantificação de fenilefrina (FEN) em amostras de águas.
- Comparar os resultados obtidos com o método proposto aos resultados do método U-PLS/RBL sem seleção de variáveis e ao bem conhecido Análise de Fatores Paralelos (PARAFAC).
- Implementar as rotinas *i*SPA-N-PLS/RBL e *i*SPA-U-PLS/RBL na forma de interface gráfica (N-*i*SPA-toolbox) em ambiente MatLab.

Capítulo II

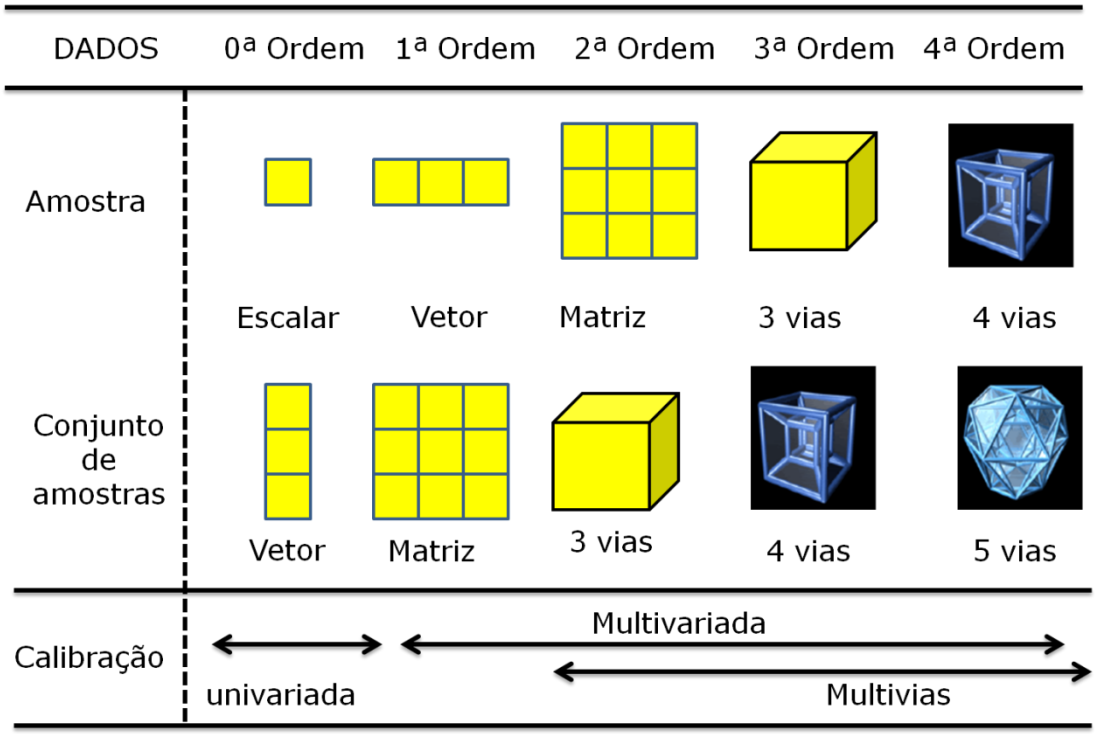
Fundamentação Teórica

2.0 FUNDAMENTAÇÃO TEÓRICA

2.1 Tipos de dados analíticos

Como concentração não é uma grandeza mensurável diretamente, sempre se buscou instrumentos que fossem capazes de medir propriedades que possuam relação com a quantidade do composto químico de interesse em uma dada amostra, e por meio da modelagem matemática desta informação obter a concentração de forma indireta [17]. De acordo com a quantidade de informação gerada por amostra, os dados analíticos podem ser categorizados [18] como indicado na **TABELA 1**.

Tabela 1: Categorização dos dados analíticos**.**



Adaptado da referencia [18]

A área de um pico cromatográfico é um exemplo de dado de ordem zero. Para cada amostra analisada apenas um único valor é obtido como sinal analítico. Para um

conjunto de amostras ou padrões, estes valores obtidos por amostra são dispostos em uma estrutura similar a um vetor.

Supondo uma situação em que não seja possível encontrar condições cromatográficas que possibilitem total resolução do analito dos demais constituintes da amostra, todo cromatograma obtido para cada amostra pode ser usado na construção de modelos de calibração multivariados. Esta estratégia é denominada de calibração de primeira ordem [9]. Para cada amostra é obtido um vetor de dados (neste caso hipotético o cromatograma). A disposição destes vetores um abaixo do outro forma uma matriz com dimensões $I \times J$; em que I seria a quantidade de amostras ou corridas cromatográficas e J os tempos de eluição, registrados em um único sensor.

Se o cromatógrafo disponível possui detecção do tipo de arranjos de diodos (DAD, "*Diode Array Detector*") na região do ultravioleta e visível, poderíamos registrar para cada tempo de eluição um espectro de absorção molecular para uma faixa espectral predefinida. Nste caso, a informação analítica associada a cada amostra consiste de uma matriz de tempos de eluição *versus* comprimentos de onda com dimensões $J \times K$; em que J representa os tempos de eluição e K os comprimentos de onda. A organização destas matrizes da origem a um conjunto de dados de três vias (*three-way*) [8-10,18].

Dados em mais vias podem ser obtidos por variação de pH, tempo de cinética, para sistemas que envolvem reação química com tempo reacional mensurável e também o uso de técnicas analíticas hifenadas, como a cromatografia bidimensional com detecção por espectrometria de massa (GC×GC×MS) ou matrizes excitação-emissão registradas a distintos pH [1-2,8].

Uma característica fundamental dos dados a ser levada em consideração na escolha de um algoritmo para modelagem, é a trilinearidade/bilinearidade do conjunto de dados.

A bilinearidade/trilinearidade dos dados está ligada á instrumentação analítica e as peculiaridades do sistema químico estudado.

A bilinearidade é um conceito que está associado a matrizes, e pode ser entendida como a possibilidade de representar uma matriz de resposta instrumental como o produto de dois vetores. Matematicamente podemos interpretar que a função resposta instrumental $X(r_1,r_2)$ pode ser separada no produto de duas funções independentes $X_1(r_1)$ e $X_2(r_2)$. Quando esta separação não é possível dizemos que a matriz é não-bilinear [19].

De modo ilustrativo, uma matriz $J \times K$ de resposta instrumental obtida em um espectrofluorímetro. Uma matriz excitação emissão (EEM, "*emission-excitation matrix"*) para um sistema constituído de um único fluoróforo, em que cada coluna K corresponde a um espectro de emissão obtido por excitação no comprimento de onda J, conforme ilustrado na **FIGURA 1**, abaixo.

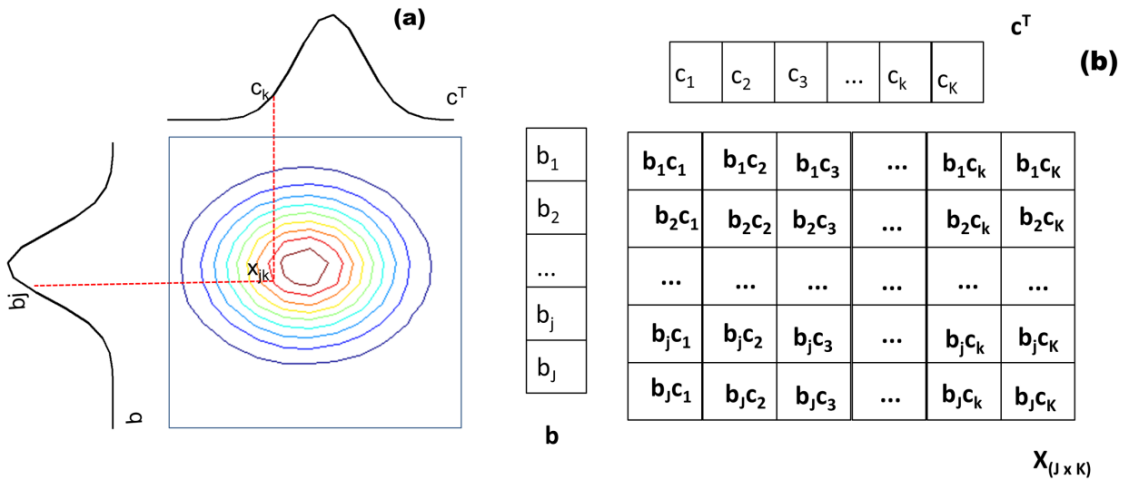


Figura 1: Representação (a) gráfica por superfícies de contorno e (b) matricial de um sinal bilinear para um único componente. Adaptado da referencia [18].

Como ilustrado na **FIGURA 1a** cada ponto x_{jk} da matriz **X** (**FIGURA 1b**) é representado pelo produto b_jc_k. O perfil do modo 1 (b) é independente do modo 2 (c), em outras palavras, significa dizer que o perfil espectral de emissão (c) não depende do

comprimento de onda de excitação, de forma que a matriz **X** pode ser representada pelo produto $\mathbf{bc}^T$, onde **b** é o vetor do perfil instrumental do modo 1 e **c** é o vetor do perfil instrumental no modo 2.

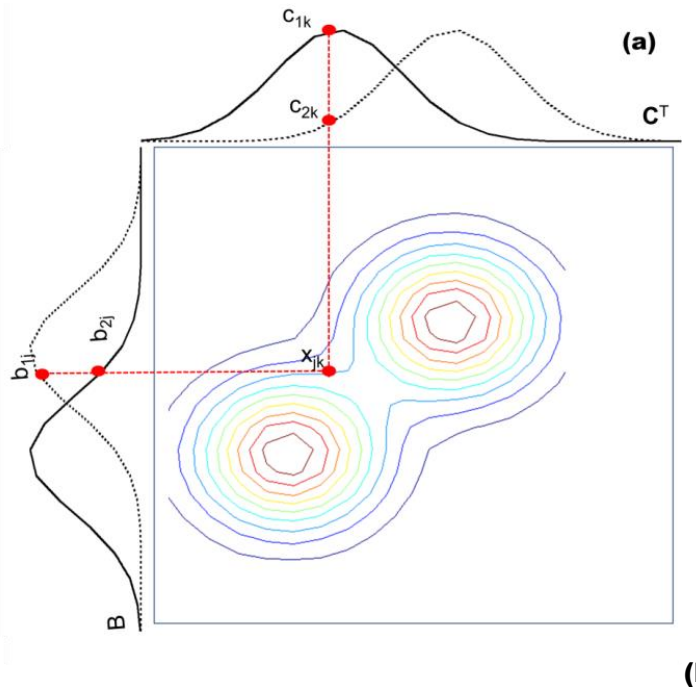
Para o caso de se ter um sistema com dois ou mais componentes, para que a matriz **X** continue bilinear o princípio da aditividade dos sinais deve ser obedecido, como ilustrado na **FIGURA 2**. Cada elemento x_{jk} da matriz **X** é dado pelo somatório do sinal de cada constituinte do sistema químico como indicado na **Eq. 1**, sendo que **X** pode ser representado agora pelo produto matricial $\mathbf{BC}^T$. O subescrito *n* na **Eq. 1** corresponde à quantidade de compostos químicos (fluoróforos), *b* e *c* são os elementos das matrizes **B** e **C** respectivamente [8-10,18].

$$x_{jk} = \sum_{n=1}^{N} b_{Jn} c_{Kn} \quad (1)$$

Para dados bilineares *n* deve ser igual ao número de constituintes químicos do sistema, ou seja, posto químico igual ao posto matemático, caso contrário a matriz **X** é dita não bilinear. Este comportamento pode sofrer desvios na presença de ruído instrumental pronunciado.

De modo similar as EEM são as matrizes do tipo LC-DAD, em que cada linha corresponde a um espectro de absorção molecular para um dado tempo de retenção *J*. E cada elemento x_{jk} da matriz LC-DAD corresponde ao produto $b_j c_k$, em que *b* representa a concentração de uma dada espécie no tempo de eluição *j* e *c* é a absortividade molar da referida espécie no comprimento de onda *K*.

Contudo, as peculiaridades de cada técnica analítica e/ou sistema químico (amostra) pode fazer com que a bilinearidade não seja obedecida, e esta se torna uma questão importante, uma vez que a grande maioria dos algoritmos assume que as matrizes de respostas instrumentais são bilineares.



(b)

c_{11}	c_{12}	c_{13}	...	c_{1k}	c_{1K}
c_{21}	c_{22}	c_{23}	...	c_{2k}	c_{2K}

C^T

b_{11}	b_{12}
b_{21}	b_{22}
...	...
b_{j1}	b_{j2}
b_{J1}	b_{J2}

B

$b_{11}c_{11}+b_{12}c_{21}$	$b_{11}c_{12}+b_{12}c_{22}$	$b_{11}c_{13}+b_{12}c_{23}$	...	$b_{11}c_{1k}+b_{12}c_{2k}$	$b_{11}c_{1K}+b_{12}c_{2K}$
$b_{21}c_{22}+b_{12}c_{21}$	$b_{21}c_{12}+b_{22}c_{22}$	$b_{21}c_{13}+b_{22}c_{23}$	...	$b_{21}c_{1k}+b_{22}c_{2k}$	$b_{21}c_{1K}+b_{22}c_{1K}$
...	...	...	...	...	...
$b_{j1}c_{11}+b_{j2}c_{21}$	$b_{j1}c_{12}+b_{j2}c_{22}$	$b_{j1}c_{13}+b_{j2}c_{23}$	...	$b_{j1}c_{1k}+b_{2j}c_{2k}$	$b_{1j}c_{1K}+b_{j2}c_{2K}$
$b_{J1}c_{11}+b_{J2}c_{21}$	$b_{J1}c_{12}+b_{J2}c_{22}$	$b_{J1}c_{113}+b_{J2}c_{23}$	...	$b_{J1}c_{1k}+bJ_2c_{2k}$	$b_{J1}c_{1K}+b_{J2}c_{2K}$

$X_{(J \times K)}$

Figura 2: Representação (a) gráfica e (b) matricial de um sinal cumulativo e bilinear para dois componentes. Adaptado da referencia [18].

Se a matriz $\mathbf{X}_{(J\times K)}$ fosse obtida por fluorescência sincrônica, em que cada espectro de emissão (linhas de **X**) é gerado pelo deslocamento concomitante dos monocromadores de excitação e emissão a diferentes Δλ [17]. Os modos instrumentais seriam Δλ e comprimentos de onda de emissão. Sendo assim, ao contrario das EEM, nas matrizes de fluorescência sincrônicas (SFM) o perfil de cada espectro de emissão *K*, depende do Δλ, sendo, portanto, funções não separáveis e tornando as matrizes sincrônicas não-bilineares. Na **FIGURA 3** é mostrado um gráfico bidimensional dos vários espectros de emissão de uma EEM (**FIGURA 3a**) e de uma matriz sincrônica (**FIGURA 3b**) para um único fluoróforo.

É possível ver que para EEM, os diferentes espectros de emissão obtidos a distintos comprimentos de onda de excitação possuem igual perfil, diferenciando-se apenas pela magnitude, caracterizando uma matriz bilinear. Já os espectros sincrônicos mostrados na **FIGURA 3b** apresentam perfil que depende do Δλ, ou seja, existe uma interdependência dos modos instrumentais.

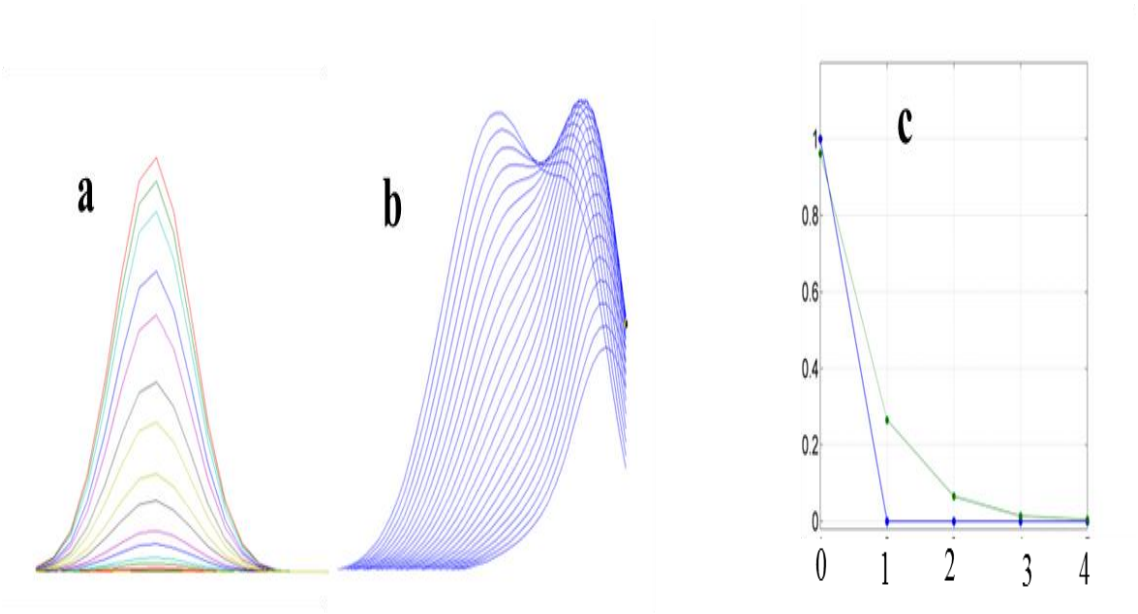


Figura 3: Representação bidimensional de uma matriz (a) EEM, (b) SFM e em (c) autovalores da EEM (**-o**) e da SFM (**--o**).

Uma forma de examinar a bilinearidade de uma matriz é através do seu posto, em que o posto matemático devia ser igual ao "*posto químico*", em casos que a bilinearidade é obedecida. Uma forma de acessar o posto de uma matriz é visualizar os autovalores da mesma [20], para o exemplo acima os autovalores da EEM e da SFM são mostrados na **FIGURA 3c**.

Por sua vez, o conceito de trilinearidade poder ser generalizado a partir do conceito de bilinearidade, em que um arranjo de três vias ($I \times J \times K$) pode ser representado como mostrado na **Eq. 2**. Onde a_{In} está associado a concentração das n espécies.

$$x_{\mathrm{ijk}} = \sum_{n=1}^{N} a_{In} b_{Jn} c_{Kn} \qquad (2)$$

A trilinearidade ou não dos dados de três vias ($I \times J \times K$) e a bilinearidade ou não das matrizes ($J \times K$) restringe os tipos de modelagem que podem ser aplicadas bem como a organização dos dados. As disposições dos dados para modelagem multimodos podem ser do tipo cubo, matriz aumentada e matrizes desdobradas [8,18] conforme ilustração da **FIGURA 4**.

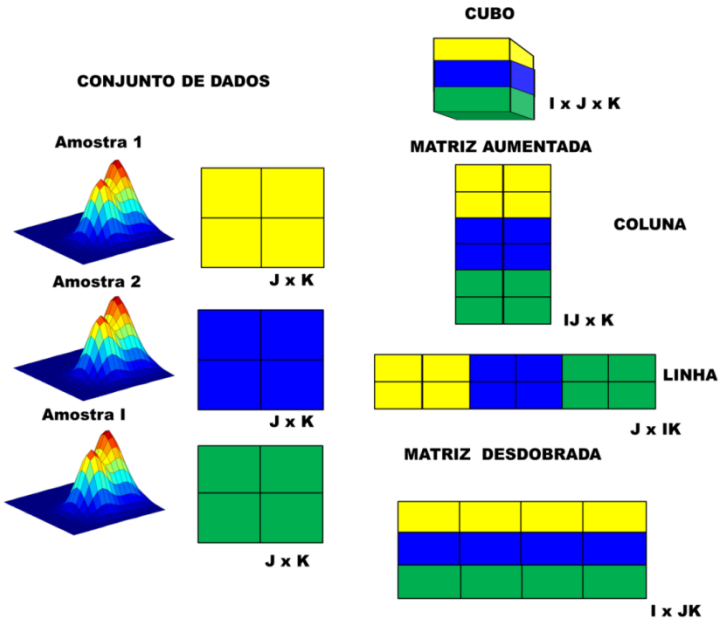


Figura 4: Representação esquemática dos tipos de disposição para dados multivias.

Quando cada matriz de resposta instrumental *i* é bilinear e o conjunto das *I* matrizes são trilineares todas as disposições (cubo, matriz aumentada e matriz desdobrada) podem ser utilizadas para modelagem dos dados. Em casos que as matrizes de resposta instrumental são bilineares, eventuais variações amostra-amostra podem ocorrer e romper a trilinearidade os dados. Neste caso as disposições em forma de matriz aumentada (linha ou coluna) ou matrizes desdobradas podem ser utilizadas. O caso mais comum em que as matrizes individuais são bilineares, porém com quebra de trilinearidade são dados do tipo LC-DAD. O deslocamento de pico que pode ocorre de amostra para amostra torna os dados não trilineares. Por fim, sistemas químicos e/ou instrumentos que geram matrizes não bilineares apresentam como única possibilidade de modelagem a forma de matrizes desdobradas.

2.2 Geração de dados de segunda ordem

Dados de segunda ordem correspondem a sinais analíticos obtidos pelo monitoramento de amostras em dois arranjos de sensores simultâneos, em que a informação registrada corresponde a matrizes de dados por amostras. Como técnicas analíticas com potencial para geração de dados de segunda ordem podemos citar a cromatografia a liquido e a gás com detecção por espectrometria de massa (GC-MS e LC-MS), eletroforese e LC acoplados a sistemas de detecção espectrais como absorção molecular e fluorescência, espectroscopia de fluorescência total 3D e sistemas com variação de pH com detecção espectral [18,21]. As técnicas analíticas mais simples e mais empregadas na geração de dados de segunda ordem são, sem duvida, as espectrais e as cromatográficas, respondendo por cerca de 90% de todas as aplicações encontradas na literatura [18].

2.3 Calibração em química analítica

A calibração é o procedimento matemático e estatístico de construção de modelos empíricos que relaciona a concentração (ou propriedade físico-química) de padrões ou amostras conhecidas com um sinal instrumental que possua relação com esta propriedade ou concentração. É de praxe o uso de modelos lineares pela maior simplicidade dos mesmos [22].

No processo de calibração são estimados coeficientes de regressão que transformam sinal medido em concentração. Como indicado na **TABELA 1**, a calibração pode ser univariada, multivariada ou multivias [18].

A calibração univariada, relaciona o sinal analítico medido em um único canal com a concentração dos padrões. Matematicamente [23] o modelo é representado pela **Eq. 3**.

$$\hat{\mathbf{y}}_\mathbf{i} = \mathbf{b_0} + \mathbf{b_1 x_i} + \mathbf{e_i} \qquad (3)$$

Onde $\hat{\mathbf{y}}_\mathbf{i}$, é o valor predito pelo modelo para i-esima amostra, b_0 e b_1, são os parâmetros do modelo, x_i é o sinal medido e_i é o resíduo deixado pelo modelo. O método dos mínimos quadrados ordinários (OLS, "*Ordinary Least Square*") é empregado para obter estimativas dos coeficientes de regressão do modelo na ausência de heterocedasticidade dos resíduos [23-24].

Do ponto de vista matemático, a calibração univariada é bastante simples, entretanto como consequência da estrutura da modelagem, total seletividade do canal analítico é requerida para que o modelo tenha exatidão. Do contrário, procedimentos prévios de extração e separação devem ser empregados para remover os interferentes. Por tanto, o modelo univariado não é capaz de fazer predições confiáveis na presença de interferentes ou mesmo indicar a presença dos mesmos [8,18].

Por outro lado, se o sinal medido em múltiplos canais forem empregado na construção do modelo que relaciona concentração e sinal analítico, há um ganho em

sensibilidade pelo uso de mais canais analíticos. Além disso, o modelo de calibração multivariado apresenta a capacidade de indicar a presença de interferentes em amostras desconhecidas, embora não consiga possa obter predições com exatidão. Esta característica é conhecida na literatura como vantagem de primeira ordem [8-10, 18, 21]. O modelo multivariado pode ser expresso [25] por uma equação matricial como indicado na **Eq. 4**.

$$\hat{\mathbf{y}} = \mathbf{Xb} + \mathbf{e} \tag{4}$$

Onde $\hat{\mathbf{y}}$ é o vetor de dimensão I×1, com os valores preditos pelo modelo, $\mathbf{X}$ é uma matriz de sinal analítico I×J, $\mathbf{b}$ é o vetor de coeficientes de regressão (estimado por OLS) de dimensão J×1 e $\mathbf{e}$ é o vetor de resíduos com as mesmas dimensões de $\mathbf{y}$.

O modelo construído como indicado na **Eq. 4**, que emprega a concentração como função da resposta analítica, é conhecido como método inverso de calibração. O termo inverso é relativo à Lei de Lambert-Beer. Esta abordagem apresenta como vantagens quando comparada a calibração direta (sinal como função da concentração) a capacidade de modelar constituintes não calibrados [26].

A solução da **Eq. 4** para $\mathbf{b}$ pelo método OLS envolve a resolução de um sistema de equações lineares apresentado na forma matricial na **Eq. 5**. Para que o mesmo seja possível e determinado, o sistema linear deve possuir mais equações que incógnitas, em outras palavras, deve haver mais amostras de calibração que sensores analíticos. O que na grande maioria dos casos em química analítica não é obedecido [27] invalidando o uso de modelos baseados em regressão linera mutipla.

$$\mathbf{b} = (\mathbf{X^T X})^{-1}\mathbf{X^T y} \tag{5}$$

Na **Eq. 5**, podemos observar que para sua resolução é necessário a inversão da matriz de covariância conjunta ($\mathbf{X}^T\mathbf{X}$), é conhecido da álgebra linear que para uma matriz ser invertível é requerido que seu determinante seja diferente de zero [28]. Ainda com

respeiro a resolução da **Eq. 5** para **b**, é válido notar que os vetores colunas em $\mathbf{X}^T\mathbf{X}$ devem ser idealmente lineramente independentes.

Dados espectrais, cromatográficos ou de outro instrumento, geralmente são fortemente correlacionados entre si, o que chamamos de dados multicolineares. A multicolinearidade presente em uma matriz, quando não torna seu determinante igual à zero, o aproxima de tal valor. Isso torna os coeficientes de regressão obtidos pela **Eq. 5**, instáveis numericamente, promovendo a propagação de erros [27-29].

O que se deseja idealmente, é que o sinal analítico medido em cada canal contenha informação distinta dos demais, que sejam linearmente independentes. Como isso não verifica na prática normalmente, os métodos para construção de modelos multivariados de primeira ordem visa contornar estes inconvenientes.

Os métodos mais conhecidos e usados na literatura são mínimos quadrados parciais (PLS, "*Partial Least Squares*") ou regressão por componentes principais (PCR, "*Principal Component Regression*"). Estas técnicas comprimem os dados de modo a concentrar a variância em poucas variáveis que são mutuamente ortogonais o próximo disso, resolvendo o problema de dimensionalidade e multicolinearidade em **X** [30-31]. Outra abordagem consiste em resolver a **Eq. 5**, empregando apenas uma quantidade de sensores que não seja superior ao número de amostras de calibração e que apresentem baixa multicolinearidade. Neste caso pode-se utilizar a regressão linear múltipla (MLR, "*Mutiple Linear Regression*") com um estagio prévio de seleção de variáveis [32].

Com estas abordagens é possível superar limitações enfrentadas pela calibração univariada. Contudo, se um constituinte inesperado (que não está presente nas amostras de calibração) aparecer em uma das amostras de predição o modelo multivariado não será hábil para prever **y** com exatidão se tornando inválido nestas condições [8,18].

Como alternativa pode-se recorrer aos métodos multivariados em multivias, em que para cada canal *J*, outro vetor de canais *K* é usado para obter informação analítica em forma matricial por amostra. Por exemplo, as matrizes excitação-emissão, que para cada comprimento de onda de excitação se registra um espectro de emissão. Este tipo de dado (ver **TABELA 1**) permite a construção de modelos multivias de segunda ordem que possibilita predição com exatidão, mesmo frente a interferentes. Na **TABELA 2**, resumem-se as características dos métodos de calibração com base na quantidade de arranjo de sensores empregados [8-10,18].

Tabela 2: Características dos métodos de calibração em função da ordem dos dados [21].

Ordem da Calibração	**Quantificação dos analitos**	**Presença de interferentes**	
		Detecção	Exatidão
0	Um	Não	Não
1	Múltiplos	Sim	Não
2	Múltiplos	Sim	Sim
3	Múltiplos	Sim	Sim

Como observado na **TABELA 2**, a possibilidade de quantificação simultânea e detecção de interferentes representa a vantagem de primeira ordem. Esta vantagem está associada à capacidade de se efetuar predições confiáveis na presença de constituintes não modelados [18]. Embora não indicado na **TABELA 2**, a vantagem de terceira ordem está relacionada a ganho abrupto de sensibilidade [33]. Vantagens de outras ordens são ainda tema de investigação por parte de quimiometristas [33]. Outra

característica dos modelos de calibração é que as vantagens são cumulativas com o aumento da ordem dos dados.

2.4 Métodos de Calibração multivias

Os métodos de calibração de segunda ordem podem ser categorizados de acordo com é alcançado a vantagem de segunda ordem. O primeiro grupo sõa os métodos que portam esta vantagem de forma intrínseca, como é o caso das abordagens que obtém perfis puros dos constituintes do sistema, a exemplo do Análise de Fatores paralelos (PARAFAC) [34], resolução de curvas multivariadas – Mínimos quadrados alternados (MCR-ALS) [35]. O outro grupo de métodos multivias, são os que fazem uso de uma etapa pós-calibração para alcançar vantagem de segunda ordem, a exemplo dos mínimos quadrados parciais desdobrados (U-PLS) e mínimos quadrados parciais multidimensionais (N-PLS) [36-37].

2.4.1 Análise de Fatores Paralelos-PARAFAC

O método conhecido como PARAFAC tem suas origens ligada as ciências humanas, mais precisamente na linguística e foi desenvolvido por R. Harshman, Carroll e Chang de forma simultânea e independente sendo denominado de Decomposição Canônica (CANDECOMP, "*Canonical Decomposition*") [38]. Porém, as aplicações do PARAFAC em química analíticas são mais tardias [38-39].

O trabalho de doutoramento do Rasmus Bro, sob orientação do professor Age Smilde [40], representa um marco na popularização do PARAFAC na comunidade de química, mais especificamente na química analítica. Sua tese intitulada "*Multi-way Analysis in the Food Industry: Models, Algorithms, and Applications*" [40], trata do PARAFAC entre outros métodos quimiométricos do ponto de vista de seus aspectos teóricos, implementação computacional e aplicações na indústria de alimentos [40].

As rotinas em ambiente MatLab do PARAFAC e demais algoritmos implementados foram disponibilizadas, o que potencializou seu uso e aplicações por outros grupos de pesquisa. Isso mostra a importância da pesquisa na área de desenvolvimento de novos modelos e implementação dos existentes de forma amigável para uma maior difusão na comunidade científica, haja vista que a programação e a álgebra linear envolvida nos métodos multímodos não são triviais para profissionais com formação em química.

2.4.1.1 PARAFAC: Fundamentos matemáticos

Do ponto de vista matemático o PARAFAC pode ser entendido como uma generalização de PCA [41], que é um método para dados bilineares ou também pode ser visto como um caso restrito do método de tucker3 [42]. De modo similar ao modelo PCA os fatores PARAFAC são compostos por duas matrizes de pesos (**B** e **C**) e uma de escores (**A**), como mostrado na **Eq. 6**.

$$\underline{\mathbf{X}} = \mathbf{A}(\mathbf{C}|\otimes|\mathbf{B})^{\mathbf{T}} + \underline{\mathbf{E}} \qquad (6)$$

Onde **A**, **B** e **C** tem dimensões I × F, J × F e K × F, respectivamente, |⊗| é o operador Khatri-Rao e **E** é o tensor de resíduos com as mesmas dimensões de **X**. Esta equação é bastante similar à representação matemática do método Tucker3 [40]. Entretanto vai aparecer o tensor **G**, que é o CORE do modelo Tucker3 ($\underline{\mathbf{G}}_{(d \times l \times h)}$). No modelo PARAFAC **G** é uma hiperidentidade [20]. A tensor CORE **G**, no modelo PARAFAC apresenta valor 1 quando d=l=h e zero para todas as demais posições. Isso indica que um dado fator em um dos modos só está relacionado com os outros fatores dos outros modos para mesmo valor de *f*.

Na **FIGURA 5** é mostrada uma representação gráfica da decomposição de um arranjo de dados de três vias com PARAFAC [20, 40].

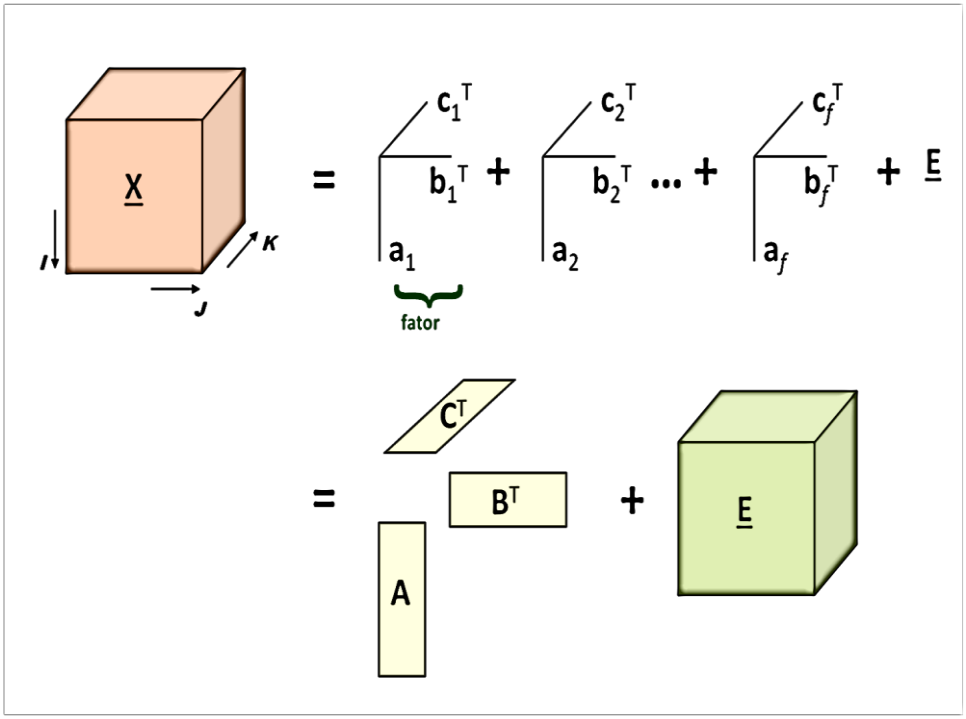


Figura 5: Representação gráfica do modelo PARAFAC. Adapatado da referência [20]

A estrutura em tríades do PARAFAC, ao contrário dos métodos bilineares que possuem liberdade rotacional, possui uma única solução capaz de minimizar a soma de quadrados dos resíduos. Esta propriedade é conhecida como unicidade e permite ao modelo PARAFAC, quando bem ajustado, obter os perfis puros dos constituintes do sistema nos modos instrumentais J e K [20].

O número de fatores (*f*) em um modelo PARAFAC corresponde ao número de constituintes da amostra, pelo menos na ausência de deficiência de posto. A escolha do valor apropriado da quantidade de fatores pode ser feita baseada no conhecimento químico do sistema, por processo de validação cruzada e/ou reamostragem [40].

Bro e colaboradores apresentaram um método para determinação automática do valor de *f* denominado de CORCONDIA ("*Core Consistency Diagnostic*") [43]. A métrica CORCONDIA está baseada na interpretação do modelo PARAFAC como um caso restrito do método Tucker3. Para um modelo Tucker3 construído com base nas

matrizes de pesos de um modelo PARAFAC perfeitamente ajustado, o tensor core **G**, deve ser uma hiperidentidade **I**. O valor de CORCONDIA é um valor percentual que indica o quanto o tensor **G** obtido pelo ajuste de uma modelo Tucker3 aos pesos do modelo PARAFAC se aproxima de **I**, calculado pela **Eq. 7**.

$$\text{CORCONDIA} = 100 * [1 - \frac{\sum_d^f \sum_l^f \sum_h^f (g_{dlh} - q_{dlh})^2}{\sum_d^f \sum_l^f \sum_h^f q_{dlh}^2}] \quad (7)$$

Os valores g_{dlh} são obtidos pelo ajuste Tucker3 aos pesos do modelo PARAFAC e q_{dlh} são elementos de uma hiperidentidade perfeita. Idealmente o valor de CORCONDIA deve ser de 100% para dados perfeitamente trilineares. Entretanto, a literatura recomenda valores acima de 60% como indicativo aceitável da consistência trilinear dos dados. O valor de CORCONDIA pode servir para indicar o ajuste no modelo PARAFAC, mas não deve ser usado como único guia [8,16,18]. O conhecimento químico do sistema sob investigação é uma ferramenta valiosa.

2.4.1.2 Implementação do PARAFAC: Mínimos Quadrados Alternados

A solução do PARAFAC consiste em estimar as matrizes **A**, **B** e **C**. O método mais empregado para este propósito é o algoritmo dos mínimos quadrados alternados (ALS, "*Alternanting Least Squares*"). O ALS se baseia no princípio de que duas das matrizes de pesos devem ser conhecidas para que se possa estimar a matriz desconhecida [20,38-40].

Se estimativas inicias de **B** e **C** estão disponíveis, **A** é facilmente obtido pelo método dos mínimos quadrados. Chamando (**B**⊗**C**) de **Z**, temos que o modelo PARAFAC pode ser representado por **X**=**AZ** e **A** é estimado por $\mathbf{A}=\mathbf{X}\mathbf{Z}^{\mathbf{T}}(\mathbf{Z}\mathbf{Z}^{\mathbf{T}})^{-1}$. O ALS pode ser descrito nas seguintes etapas:

- Decide-se o número de fatores *f*.
- Inicializam **B** e **C**.

- 3-Estima **A** empregando **X**, **B** e **C**.
- 4-Estima **B**.
- 5- Estima **C**.
- 6- Repetem-se as etapas de 3 a 5 até a convergência.

A é uma matriz com dimensões I×F, em que cada coluna consiste de um vetor de pesos proporcional as concentrações, **B** e **C** têm dimensões J×F e K×F, com os perfis nos modos intrumentais. Na etapa 3, **X** é desdobrada em uma matriz I×JK.

As etapas apresentadas acima compreendem o funcionamento do algoritmo ALS. Embora sejam apenas 6 etapas, para casos que envolvem dados de alta dimensionalidade o processo de minimização pode levar dezenas de minutos para ser alcançada. Em alguns casos, os modos instrumentais podem apresentar alta colinearidade entre as variáveis, levando a problema de instabilidade numérica na etapa 3 (método dos mínimos quadrados) do ALS [20, 39, 40].

Um aspecto de grande importância na otimização ALS está relacionado aos valores de inicialização das matrizes **B** e **C**. Valores que apresentem similaridades com os perfis reais podem diminuir o esforço computacional e prevenir que o modelo fique preso a mínimos locais [16, 20, 40]. Várias propostas de inicialização do ALS são encontradas na literatura, como o uso de valores randômicos ou valores baseados em autovalores generalizados. Alguns autores têm relatado que se os dados são trilineares, mínimo local é uma problema muito incomum na otimização ALS, embora existam divergências na literatura em especial para casos mais complexos com modelos com mais de cinco fatores [8,18].

Depois de inicializar o ALS, iterações serão executadas até um número grande predefinido ou que seja atingido um critério de convergência estabelecido. Um critério

comum utilizado é interromper as iterações quando a mudança relativa no ajuste entre duas iterações é inferior a um determinado valor [40].

Outra característica importante do PARAFAC é que na fase de otimização ALS restrições podem ser impostas à solução. Estas restrições levam a modelos com menor fração de variância explicada, por tornar a modelagem menos flexível ("*hard*"). Entretanto, os pesos obtidos são mais interpretáveis e possuem maior sentido físico e/ou químico. As principais restrições são: não negatividade, unimodalidade e ortogonabilidade [39-40].

A não negatividade é a restrição que impõe que a solução obtida pelo OLS deve conter apenas valores iguais ou maiores que zero. Em outras palavras, valores negativos não são permitidos. Os métodos de implementação mais comuns para a restrição de não negatividade são: zero forçado, em que valores negativos são substituídos por zeros; mínimos quadrados não negativos (nnls, "*non negative least square*") e mínimos quadrados não negativos rápido (fnnls, "*fast non negative least square*"). Nestes dois últimos, a solução do OLS só se admite valores iguais ou maiores que zero. A restrição de não negatividade é bastante usada na obtenção dos perfis instrumentais puros, para evitar soluções negativas, que não possuem sentido físico, como absorbância negativa, por exemplo [38-40].

Matematicamente, a unimodalidade pode ser entendida com base na definição de que uma função $f(x)$ é dita unimodal se no intervalo $a \leq x \leq b$ se, e somente se, ela for monotônica em ambos os lados do ponto de ótimo x no intervalo. Essa definição implica em um único máximo ou único mínimo no intervalo $a \leq x \leq b$. Do ponto de vista da resolução de problemas químicos envolvendo PARAFAC, o caso em que se aplica esta restrição são em dados de cromatografia, em que se sabe previamente que para cada analito corresponde um único pico com apenas um máximo [20,40].

Ao contrário do PCA que possui solução cumulativa, o PARAFAC não apresenta esta propriedade. Em um modelo PARAFAC com *f* fatores, estes memos *f* fatores não são iguais aos *f* fatores de um modelo com *f* +1 fatores. Esta característica dos modelos PARAFAC é consequência da não existência de ortogonabilidade entre fatores. Mas, a ortogonabilidade pode ser imposta aos fatores PARAFAC como uma restrição na etapa de otimização do ALS. Normalmente não se usa esta restrição, pois fatores ortogonais tendem a ser abstratos e sem sentido físico [20,40].

O PARAFAC apresenta como limitação o uso restrito a dados que cumprem com a trilinearidade. Muitos sinais analíticos rompem este critério, como por exemplo, dados de cromatografia liquida, em que pode não ocorrer reprodutibilidade dos tempos de retenção [8,18,20]. Uma adaptação do PARAFAC foi proposta de modo a lidar com dados com perda de trilinearidade em um dos modos instrumentais, e foi denominado PARAFAC2 [44].

Outro problema analítico que torna inválido o modelo PARAFAC, são casos em que os perfis a serem estimados apresentam problema de dependência linear. Para este tipo de dados, um método que pode ser entendido como uma generalização do PARAFAC foi proposta por Bahram e Bro, e é conhecido como Perfis Paralelos com Dependência Linear (PARALIND, "*PARAllel profiles with LINear Dependencies*") [45].

2.4.2 Partial Least Square (PLS)

A regressão por mínimos quadrados parciais foi desenvolvida por Herman Wold e colaboradores em 1975 [30] e representa a ferramenta da quimiometria mais explorada e versátil [46]. Existem muitas formas distintivas de computar um modelo de regressão PLS [47]. Segundo Andersson [47], todas as abordagens são equivalentes, o principal diferencial está associado à instabilidade numérica do modelo e ao esforço

computacional envolvido. Uma das formas mais difundidas, é sem duvida o algoritmo dos mínimos quadrados parciais iterativo não linear (NIPALS, *non-linear iterative partial least squares*) [48]. Contudo outras estrátegias também são reportadas na literatura [47].

Todas as descrições dadas a seguir são referentes ao método PLS-1, onde se assume que a variável dependente **y** é um vetor I×1. Contudo, os conceitos discutidos aqui são facilmente generalizados para o PLS-2 (**Y** é uma matriz I×D) [49]. Inicialmente, a matriz **Z**(I×J) e o vetor **c**(I×1), referente ao conjunto de calibração são previamente processados gerando $\mathbf{X}_{(I\times J)}$ e $\mathbf{y}_{(I\times 1)}$ respectivamente. O algoritmo de NIPALS pode ser representado pelas seguintes etapas [50]:

1) Calcula-se a matriz de pesos ponderados (**W,** *loadings weights*) para variável latente *A* **Eq. 8**.

$$\mathbf{w}_A = \mathbf{X}^T\mathbf{y} \tag{8}$$

2) Os pesos ponderados são normalizados para comprimento 1 empregando a **Eq. 9.**

$$\mathbf{w} = \frac{\mathbf{w}_A}{\sqrt{\mathbf{w}_A^T\mathbf{w}_A}} \tag{9}$$

Os pesos ponderados representam a direção do espaço muldimencional com máxima correlação entre $\mathbf{X}_{(I\times J)}$ e $\mathbf{y}_{(I\times 1)}$.

3) A matriz de escores de **X** é calculada pela projeção de **X** em **W** como mostrado em **Eq. 10.**

$$\mathbf{t}_A = \mathbf{X}\mathbf{W}_{Anor} \tag{10}$$

4) Matriz de pesos é então calculada **Eq 11**.

$$\mathbf{L}_A = \mathbf{X}^T\mathbf{T}/\mathbf{T}^T \tag{11}$$

5) De forma similar, calcula-se os pesos em **y** (**Eq. 12**)

$$\mathbf{Q}_A = \mathbf{y}^T\mathbf{T}/\mathbf{T}^T \quad (12)$$

A contribuição da variável latente A é removida pela diferença $\mathbf{X}_{A-1}=\mathbf{X}-\mathbf{T}_A\mathbf{L}_A^T$ e $\mathbf{y}_{A-1}=\mathbf{y}-\mathbf{TQ}$. A nova variável latente é calculada substituindo **X** e **y** por $\mathbf{X}_{A-1}$ e $\mathbf{y}_{A-1}$ respectivamente. Após um número determinado de *A* variáveis latentes, os coeficientes de regressão para um modelo linear é dado por **Eq. 13**.

$$\mathbf{v} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\,\mathbf{y} \quad (13)$$

2.4.2.1 U-PLS

A descrição do método PLS-1 acima é bem conhecida para métodos de calibração primeira ordem. Para modelagem de dados de segunda ordem o PLS apresenta a variante conhecida como U-PLS, em que o termo "U" do inglês significa desdobrar ou vetorizar (*unfold*). Proposto em 1990 por Ohman e Wold [36], o U-PLS, atua sobre um tensor do tipo $\underline{\mathbf{X}}$ (I×J×K), desdobrando cada matriz de respostas instrumentais ($\mathbf{X}_{(J\times K)}$) em um vetor **x** (1×JK), gerando a matriz desdobrada **uX** (I×JK). Na sequência, o caráter multivias dos dados é desconsiderado e um modelo PLS-1 como descrito acima é empregado para estabelecer um modelo de regressão entre sinal e concentração [36,51]. O modelo U-PLS em nada difere do método PLS-1, que corresponde a uma decomposição bilinear dos dados instrumentais como indicado na **FIGURA 6**.

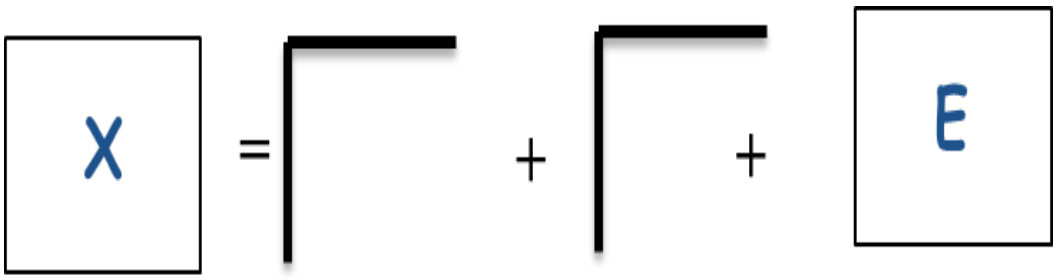


Figura 6: Representação da decomposição bilinear de uma matriz **X** pelo U-PLS.

2.4.2.2 N-PLS

O PLS trilinear ou como é mais popularmente conhecido N-PLS, foi formalmente proposto por Bro [37] em 1996, embora outros estudos anteriores tenham relatado o uso de PLS para decomposições trilineares [20,37,40]. O N-PLS foi apresentado como uma alternativa ao U-PLS, onde a estrutura trilinear dos dados é considerada e um modelo mais estável e menos complexo é obtido [37]. Quando comparado ao PARAFAC, o N-PLS apresenta a vantagem de um menor esforço computacional [40], uma vez que está baseado na resolução de um problema de autovetores [20].

Em essência, o N-PLS, proposto por Bro, não é diferente do PLS-1, consistindo apenas de uma generalização para dados multivias. Em ambos os métodos a decomposição dos dados instrumentais visa maximizar a covariância entre **y** (variável dependente) e os escores de **X**. Ao contrario do PLS-1, no N-PLS cada fator (variável latente) obtido da decomposição **X** (I×J×K) corresponde a uma tríade [37]. Cada tríade é caracterizada por um vetor de escores (**t**) e dois vetores de pesos ponderados ($\mathbf{w}^j$ e $\mathbf{w}^k$, loadings weights) que são os pesos ponderados nos modos instrumentais J e K respectivamente. Estes possuem igual significado quando comparado ao PLS-1, ou seja, correspondem as direção de máxima covariância entre **X** (I×J×K) e **y**. A tríade (representada graficamente na **FIGURA 7**) pode ser expressa como **Eq. 14**:

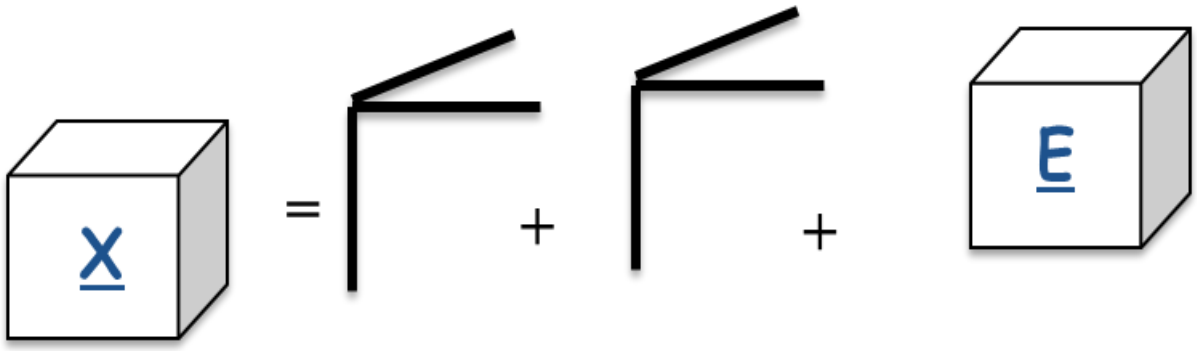


Figura 7: Representação da decomposição em três vias do tensor **X** (I×J×K) pelo método N-PLS.

$$x_{ijk} = t_i\, w_j^j w_k^k \tag{14}$$

De modo similar ao PLS-1, no modelo trilinear $\mathbf{w}^j$ e $\mathbf{w}^k$ busca-se minimizar a soma dos quadrados dos resíduos [20] de acordo com a **Eq. 15**:

$$e^2 = (x_{ijk} - t_i\, w_j^j w_k^k)^2 \tag{15}$$

A solução pelo método dos mínimos quadrados é dado por **Eq. 16**:

$$t_i = \sum_{j=1}^{j} \sum_{k=1}^{k} (\, z_{jk} w_j^j w_k^k) \tag{16}$$

Onde z_{jk} são os elementos da matriz **Z** com dimensões (J ×K*)* correspondente a soma das I matriz que compões o tensor $\underline{\mathbf{X}}$ (I×J×K), ponderadas pela concentração do analito, como mostrado na **Eq. 17**.

$$\mathbf{Z} = \mathbf{X}_1 y_1 + \mathbf{X}_2 y_2 + \mathbf{X}_3 y_3 + \cdots + \mathbf{X}_I y_I \tag{17}$$

O próximo passo é a determinação de $\mathbf{w}^j$ e $\mathbf{w}^k$, que pode ser facilmente obtido por decomposição em valores singulares da matriz **Z**. Na sequência **t** é estimado empregando a **Eq. 16**. Na etapa seguinte é obtido o vetor de coeficientes de regressão como indicado na **Eq. 18**.

$$\mathbf{v} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\,\mathbf{y} \tag{18}$$

Assim como no PLS-1, a contribuição do *A*-ésimo fator é removida, e o próximo fator é computado sobre os resíduos remanescentes, onde cada amostra $\mathbf{X_I}$ é substituída por $[\mathbf{X}_i - t_i\, \mathbf{w}^j\, (\mathbf{w}^k)^T]$ e **y** por (**y** – **Tv**). É possível também, estimar os resíduos instrumentais em $\underline{\mathbf{X}}$ (I×J×K), estimando as matrizes de pesos $\mathbf{P^J}$ e $\mathbf{P^K}$.

O número ótimo de fatores (*A)*, nos modelos U-PLS e N-PLS pode ser acessado por procedimentos como validação-cruzada, uso de um conjunto externo de validação, monte-carlo [52-53], em que se observa a variação do erro de predição em função do número de fatores.

2.4.2.3 Bilinearização Residual

Ao contrário do método PARAFAC, ambas as abordagens de uso do PLS para dados multivias discutidos não portam vantagem de segunda ordem de forma intrínseca. Embora o processo de calibração envolva dados de segunda ordem os métodos U-PLS e N-PLS não são capazes de fazer predições confiáveis na presença de constituintes não modelados [51,54]. Este inconveniente foi contornado pelo uso de uma etapa de pós-calibração conhecida como bilinearização residual (RBL) [51]. Então, os modelos U-PLS e N-PLS passam a ser chamados de U-PLS/RBL e N-PLS/RBL, respectivamente [54-55].

A etapa de predição de uma dada amostra desconhecida $\mathbf{X_u}$, envolve em primeiro lugar a obtenção dos seus escores, que podem ser obtidos pela projeção do sinal instrumental da amostra desconhecido sobre os pesos do conjunto de calibração [56]. Então, a concentração pode ser estimada como indicado na **Eq. 19**.

$$\hat{y}_u = \mathbf{t}_u^T \mathbf{v} \tag{19}$$

Onde $\hat{y}_u$ é a concentração predita, $\mathbf{t_u}$ são os escores da amostra desconhecida e $\mathbf{v}$ é o vetor das estimativas dos coeficientes de regressão obtido como indicado nas **Eq. 13** e **18**. No caso eventual em que um constituinte não modelado possa estar presente na amostra $\mathbf{X_u}$, os escores $\mathbf{t_u}$ não servem para uma predição apropriada de $\hat{y}_u$. A presença de um constituinte não modelado em uma amostra desconhecida pode ser detectada pela

inspeção do gráfico de resíduos da respectiva amostra como mostrado na **FIGURA 8** [57].

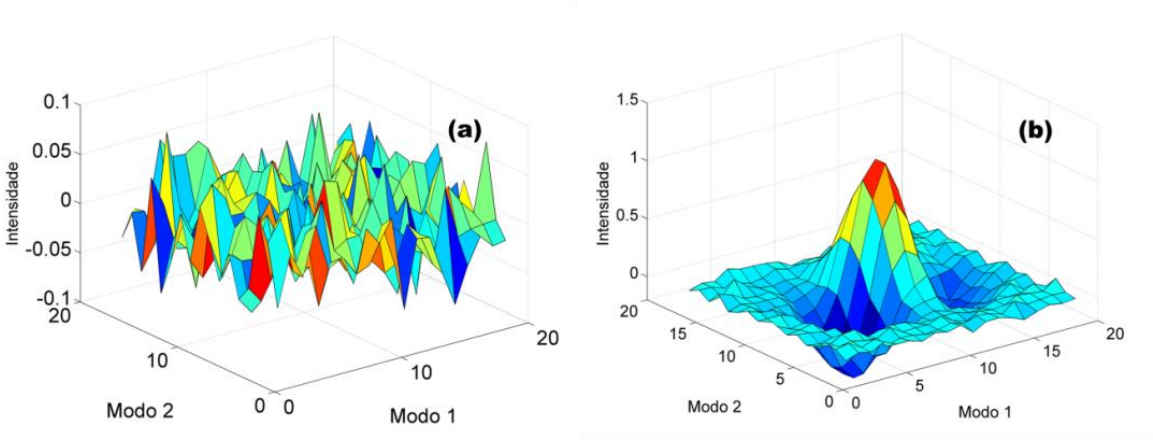


Figura 8: Matriz de resíduo instrumental de uma amostra de testes (a) na ausência (b) e na de presença de constituintes não modelados.

Na **FIGURA 8a** é observado gráfico de resíduos típico para uma amostra de predição modelada adequadamente, com presença de ruídos de baixa intensidade e aleatórios, evidenciando a inexistência de constituintes não modelados em que a vantagem de segunda ordem não é requerida. Ao contrário, na **FIGURA 8b** é possível observar claramente a presença de um perfil característico, contribuição dos constituintes não modelados. Cenário no qual os modelos U-PLS e N-PLS não serão capazes de efetuar predições confiáveis. Entretanto, distintamente dos métodos de primeira ordem, os dados de segunda ordem permitem o uso da etapa RBL, assegurando predições exatas, ou seja, alcançando a vantagem de segunda ordem.

Os gráficos mostrados nas **FIGURAS 8a** e **8b** são as matrizes de resíduos e podem ser representadas por um escalar sem perda de representatividade. Este escalar é o desvio padrão residual (s_p) para amostra de predição **Xu**. O que se faz na pratica é comparar s_p com o ruído instrumental típico, que corresponde ao desvio padrão residual

estimado para todo o conjunto de calibração (s_{cal}). O valores de s_{cal} e s_p para os modelos U-PLS e N-PLS são calculados conforme indicado abaixo, respectivamente [57-58].

$$s_{\mathrm{cal}} = \left\|\mathrm{vec}(\mathbf{uX}\mathrm{cal} - \mathbf{TP}^{\mathrm{T}}\right\| / \sqrt{(JK - A)I} \tag{20}$$

$$\mathrm{s_p} = \left\|\mathrm{vec}(\mathbf{X_u}) - \mathbf{Pt_u}\right\| / \sqrt{(\mathrm{JK} - \mathrm{A})} \tag{21}$$

$$\mathrm{s_{cal}} = \sum_{i=1}^{I} \frac{||\mathbf{X_{cal\,i}} - \mathrm{reshape}\{\mathbf{T_u}[(\mathbf{w^j})|\otimes|(\mathbf{w^k})]\}||}{(\mathrm{JKI} - \mathrm{A})\mathrm{I}^{1/2}} \tag{22}$$

$$\mathrm{s_p} = \frac{||\mathbf{X_u} - \mathrm{reshape}\{\mathbf{t_u}[(\mathbf{w^j})|\otimes|(\mathbf{w^k})]\}||}{(\mathrm{JKI} - \mathrm{A})^{1/2}} \tag{23}$$

Onde "*reshape*" corresponde à operação de converter um vetor JK×1 em uma matriz J×K e $\|\,.\,\|$ corresponde a norma Euclidiana. Valores de s_p superior a s_{cal} é indicativo de interferentes na amostra $\mathbf{X_u}$, e que o procedimento RBL deve ser usado. O RBL consiste em decompor em valores singulares (**Eq. 24**) a matriz de resíduos ($\mathbf{E_p}$) da amostra $\mathbf{X_u}$.

$$\mathbf{B}_{\mathrm{unex}}\mathbf{G}_{\mathrm{unex}}(\mathbf{C}_{\mathrm{unex}})^{\mathrm{T}} = \mathrm{SVD}\ (\mathbf{E_p}) \tag{24}$$

$\mathbf{B_{unex}}$ e $\mathbf{C_{unex}}$ são as matrizes de autovetores de $\mathbf{E_p}$ no espaço linha e no espaço coluna respectivamente, enquanto $\mathbf{G_{unex}}$ corresponde a matriz de autovalores [59]. As matrizes $\mathbf{B_{unex}}$, $\mathbf{C_{unex}}$ **e** $\mathbf{G_{unex}}$ são truncadas para N_i fatores. Em outras palavras o número de fatores RBL (N_i) é o posto de $\mathbf{E_p}$. O produto $\mathbf{B_{unex}C_{unex}}\ \mathbf{(G_{unex})^T}$, denominado de $\mathbf{S_{int}}$, contêm informação referente ao perfil instrumental dos constituintes não modelados e são usados para modificar os escores $\mathbf{t_u}$ da amostra de teste $\mathbf{X_u}$ para minimizar $\mathbf{E_p}$ empregando um procedimento de otimização não linear do tipo Gauss-Newton [60]. A

minimização de $\mathbf{E_p}$ via Gauss-Newton ocorre quando s_p e s_{cal} são similares. São empregadas as **Eq. 25** e **26** para o U-PLS e N-PLS respectivamente.

$$\mathbf{vec}(\mathbf{X_u}) = \mathbf{Pt_u} + \mathrm{vec}\,(\mathbf{S_{int}}) + \ \mathrm{vec}(\mathbf{E_u}) \tag{25}$$

$$\mathbf{X_u} = \mathrm{reshape}\{\mathbf{t_u}[(\mathbf{w}^{\mathrm{j}})|\otimes|(\mathbf{w}^{\mathrm{k}})]\}\,|| + \mathbf{S_{int}} + \mathbf{E_u} \tag{26}$$

Onde $\mathbf{E_u}$ é a matrizes de resíduos obtido após o uso de N_i fatores RBL. Com base nas **Eq. 25** e **26** é possível observar que no procedimento Gauss-Newton, os pesos de calibração são mantidos e modifica-se $\mathbf{t_u}$ para minimizar $\mathbf{E_p}$ ate $\mathbf{E_u}$. Os perfis dos interferentes contidos em $\mathbf{S_{int}}$ são continuamente atualizados empregando a **Eq. 24**.

A escolha do valor ótimo de N_i é feito observando a variação de s_u (desvio padrão residual pós RBL) em função N_i. Quando para um dado valor de N_i, s_u é comparável a s_{cal}, N_i corresponde ao melhor ajuste RBL. O valor de s_u é dado como mostrado em **Eq. 27**.

$$S_{\mathrm{u}} = \ \|\mathbf{E_u}\| / \sqrt{[J - N_{\mathrm{i}})(K - N_{\mathrm{i}}) - A} \tag{27}$$

Na **FIGURA 9a** é mostrado um gráfico típico da variação de s_u em função de N_i para um caso hipotético. Observar-se que para N_i igual zero s_u é próprio s_p, a medida que o número de RBL aumenta, ocorre um abaixamento de s_u até atingir um valor ótimo (na mesma magnitude de s_{cal} indicado pela linha solida azul) para N_i igual a 5, sugerindo que a amostra hipotética $\mathbf{X_u}$, em questão possui 5 constituintes não modelados.

É importante notar que para cinco fatores (N_i igual a 5) s_u se torna muito semelhante a s_{cal} e que para valores maiores de N_i (6 e 7) não ocorre variação significativa de s_{cal}, mostrando uma estabilização de $s_{u.}$. Contudo, para um uso excessivo de fatores (N_i igual

a 8) ocorre um decréscimo de s_u para níveis inferiores ao resíduo instrumental (que tem como estimador s_{cal}), indicando um caso de sobreajuste [61].

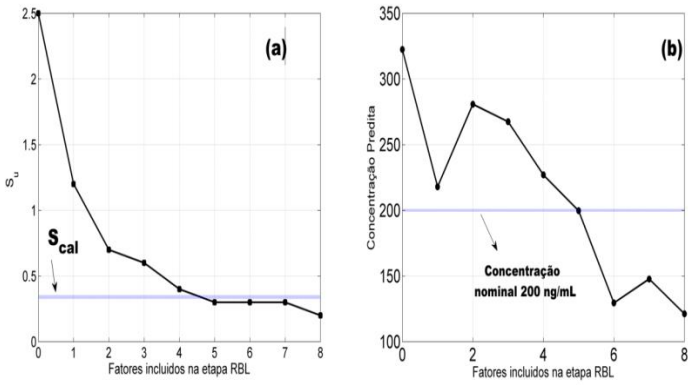


Figura 9: Em (a) variação de S_u e em (b) variação da concentração predita ambas com a inclusão de fatores RBL. A linha sólida azul representa o nível do ruído instrumental das amostras de calibração em (a) e em (b) a concentração nominal [61].

De forma concomitante, na **FIGURA 9b** observas-e a variação da concentração predita em função de N_i. Note que quando o valor ideal de fatores é empregado, uma predição concordante com o valor esperado é obtido. No entanto, o gráfico da **FIGURA 9b** não está disponível em situações reais para amostras desconhecidas, mas aqui tem como finalidade indicar que o uso de valores de s_u similares a s_{cal} conduz a escolhas corretas de N_i.

2.5 Seleção de Variáveis

Seleção de variáveis pode ser compreendida como um problema de otimização combinatorial com restrições, em que o objetivo é encontrar um subconjunto de preditores, capaz de produzir modelos de calibração com melhor exatidão e robustez, quando comparado ao modelo com todas as variáveis "*full model*" [62]. Os métodos de seleção de variáveis buscam ainda produzir modelos mais simples ou parcimoniosos [16,63].

No contexto de dados de primeira ordem, seleção de variáveis é amplamente empregada e difundida em diversos campos da quimiometria, como por exemplo,

classificação [64], transferência de calibração [65] e calibração multivariada [66-67] para os mais diversos métodos de regressão como MLR [68], PLS [69], dentre outros. As técnicas de seleção de variáveis, de modo geral, podem ser categorizadas como indicado no diagrama da **FIGURA 10**.

Figura 10: Diagrama esquemático de classificação dos métodos de seleção de variáveis Adaptado de [78].

Os métodos randômicos possuem alguma variável de entrada aleatória e, consequentemente, o subconjunto de variáveis selecionado é estocástico, ou seja, está associado a certo grau de probabilidade [70]. Os métodos randômicos mais conhecidos são os que simulam processos naturais como Algoritmo Genético [71] (GA, "*Genetic Algorithm*"), Colônia de Formigas (AC, "*Ant Colony*") [72], Busca de Tabu (TS, "*Tabu Search*") [73], por exemplo. Por outro lado, os métodos determinísticos não estão associados a nenhuma probabilidade a priori, apresentando como solução um único

subconjunto de variáveis, sendo o algoritmo das projeções sucessivas [74] (SPA) um exemplo de método determinístico.

Com respeito à forma do subconjunto de variáveis selecionadas, esta pode ser sob a forma de variáveis individuais discretas (conjuntos descontínuos) ou de intervalo ou combinação de intervalos contínuos. O primeiro método é muito empregado como alternativa aos métodos de compressão para resolver problemas de multicolinearidade em regressão MLR. Já os métodos de seleção de intervalos são indicados para métodos que utilizam rotações ortogonais previamente, bem como, só é adequado para atuar em variáveis continuas, como espectros, voltamogramas e cromatogramas, por exemplo. Não há motivação para se aplicar métodos de seleção de intervalos em dados discretos, como dados de QSAR, ou de parâmetros físico-químicos, uma vez que as posições das variáveis em **X** são arbitradas pelo analista.

A forma de atuação dos métodos de seleção também os diferencia, e pode ocorrer sob a forma de filtros, externos e internos [70]. Os métodos do tipo filtro atuam sobre um dado modelo previamente ajustado, em que é estabelecido um limiar de corte. Consequentemente, variáveis acima ou baixo (como na definição do limiar de corte) deste limiar são selecionadas. Jack-knife [75] e estratégias que usam ponderação dos pesos [76] são alguns exemplos de métodos do tipo filtro. Os métodos externos geram subconjuntos de variáveis e os analisam um a um, atribuindo um valor que está associado a uma dada função de custo [77]. Abordagens como GA, SPA, AC e seleção de intervalos, são exemplos de métodos de seleção externa. Por fim, os métodos internos, menos comuns, o processo de seleção ocorre concomitantemente com o processo de modelagem. Como exemplos pode ser citado o método de limiar de corte flexível proposto por Sæbø *et al* no âmbito de regressão PLS [78].

2.5.2 Seleção de variáveis em dados multivias

Essa seção não tem por objetivo fazer uma revisão exaustiva da literatura acerca de seleção de variáveis em dados e/ou calibração multivias, mas apenas mostrar as principais contribuições nesta área e como estes trabalhos foram conduzidos, indicando a instrumentação analítica abordada, método de modelagem e seleção de variáveis.

O primeiro registro encontrado acerca de seleção de variáveis e considerado relevante foi proposto por Andersson *et al* [79]. Neste trabalho, os autores propõem o método PARAFAC-ponderado (*Weighted* PARAFAC) para contornar problemas de não linearidade em dados do tipo LC-DAD. Segundo os autores, os dados LC-DAD podem ser afetados por interações do tipo solvente-soluto e saturações do detector. O método PARAFAC-ponderado elimina as variáveis (tempos de eluição e comprimentos de onda) onde ocorre grande desvio da lei de Lambert-Beer [79], uma vez que estes desvios causam quebra de trilinearidade, uma característica requerida pelo PARAFAC. Desta forma, o método proposto garante um bom ajuste do modelo PARAFAC e, consequentemente, bons resultados qualitativos (recuperação dos perfis puros em ambos os modos instrumentais) e quantitativos (erro de predição baixos).

Do ponto de vista da implementação, os autores fazeram uso de um tensor de pesos **W** associado a cada elemento x_{ijk} do tensor de dados instrumentais, de modo a penalizar sinais registrados nos sensores com grandes desvios de trilinearidade. Cada elemento de w_{ijk} corresponde um valor limite (*cut-off value*) [79].

Wu *et al* [80], também empregando PARAFAC, desenvolveram um método de seleção de variáveis baseado em algoritmo genético em dados de três e quatro vias [80]. Neste estudo não é conduzida uma determinação quantitativa, apenas análises exploratórias são efetuadas. O objetivo do método GA-PARAFAC de Wu *et al*, é, empregando um subconjunto de variáveis, reter toda informação relevante dos dados. O

método é proposto como alternativa aos métodos ponderados, uma vez que o modelo PARAFAC, ao contrario de PCA, é não cumulativo e a ponderação de um dado sensor no modo J pode alterar os resultados no modo K. O algoritmo genético foi implementado com codificação binária para gerar as cadeias de variáveis nos modos J e K e o método generalizado de análise de custo foi empregado para atribuir o grau de semelhança entre os pesos do modelo global e os pesos de cada subconjunto de variáveis. Para ambos os estudos de caso os autores relataram ter alcançado bons resultados e que a fração de variância relevante foi devidamente preservada.

Stordrange *et al*, investigaram o uso de métodos multivias como PARAFAC e Tucker3 [81] para modelar dados de espectrometria NIR, registrado no processo de produção de um composto orgânico em várias bateladas. Modelos foram construídos em diferentes faixas de comprimentos de onda para predizer a concentração do composto sintetizado. Estas faixas foram escolhidas de modo arbitrário, não constituindo um método de seleção de variáveis para dados multivias, sendo apenas uma forma de avaliar se uma dada região espectral é mais informativa em detrimento de outra. O erro de predição foi o indicador da qualidade das variáveis selecionadas. Segundo os autores a etapa prévia de seleção de variáveis se mostrou essencial para interpretabilidade dos resultados obtidos.

Gourvénec *et al* [82], em seu estudo abordaram o uso de MCR combinado com o método de projeções ortogonais (OPA) para modelar dados de espectroscopia NIR obtidos de processos em batelada. A seleção de variáveis via GA foi apresenta como alternativa para acompanhar um processo de produção, registrando consecutivos espectros NIR.

Em um contexto de controle de processo similar ao discutido por Gourvénec *et al* [82], Chu *et al*, [83] apresentaram o algoritmo de seleção de variáveis baseado no método "*Forward Floating Selection algorihtm (FFSA)*" [84]. Os autores empregaram espectrometria NIR para controlar o processo de polimerização do policloreto de vinila, e o FFSA é combinado com N-PLS. O erro médio quadrático de predição é empregado como função de custo para guiar a escolha das variáveis mais informativas.

Levando-se em consideração o desenvolvimento de métodos seleção de variáveis em calibração multivias, a proposta de Carreiro *et al*, pode ser considerada a única contribuição previa na literatura [85]. Os autores desenvolveram um método que combina seleção de variáveis via GA com mínimos quadrados bilineares (BLLS). O método denominado GA-BLLS é avaliado na determinação de resíduos de cinco pesticidas (carbaril, metil tiofanato, simazina, dimetoato e seu metabolito ftalimida) em amostras de vinho tinto por meio da modelagem de dados gerados em um sistema LC-DAD, operando com eluição em modo isocrático. Os autores relataram que eventuais corrimentos de pico foram corrigidos previamente.

O algoritmo GA-BLLS foi implementado em cinco etapas. A primeira é a codificação das variáveis empregando sistema binário (0 ou 1). Dado um tensor $\underline{\mathbf{X}}$ com dimensões I × J × K, onde as fatias ("*slices*") J × K do tensor $\underline{\mathbf{X}}$ são matrizes LC-DAD registradas em J tempos de eluição e para cada tempo J é registrado um espectro no DAD com K comprimentos de onda. Vetores com dimensões 1× (J + K) são gerados pela vetorização das matrizes LC-DAD. Estes vetores são chamados de cromossomos e seus elementos são os genes. Cada cromossomo possui J + K genes, sendo os J primeiros genes relacionados com os tempos de eluição e os K últimos relacionados com os comprimentos de onda. Os genes recebem de forma randômica o valor 0 ou 1.

Sendo que valor 0 significa que este gene não faz parte do modelo (variável não incluída) e valor 1 significa que esta variável está incluída.

Na sequência, ocorre a geração da população inicial, que consiste de uma matriz **R** com dimensões Q × (J+K). Q é o número de indivíduos (ou cromossomos) da população inicial. No trabalho de Carneiro *et al*, foi empregado uma população com 100 indivíduos e cada individuo com 10% dos genes com valor igual 1, ou seja, cada individuo corresponde a uma cadeia de dez variáveis escolhidas de forma randômica dentro das J + K variável disponíveis [85].

As etapas de cruzamento e mutação aconteceram com probabilidades de 70 e 1 %, respectivamente. A população a cada geração foi mantida fixa em 100 cromossomos. Para cada cromossomo é atribuído um valor de aptidão, que neste casso corresponde a erro de predição baseado em uma modelo BLLS com etapa pós-calibração de bilinearização residual.

O trabalho proposto por Carneiro *et al*, mostrou o potencial de aplicar seleção de variáveis em dados de segunda ordem . Contudo a principal fragilidade desta proposta se encontra no fato do GA ser susceptível a muitos parâmetros de otimização que dependem da experiência do analista com a técnica, o que pode não ser uma tarefa trivial. O GA apresenta ainda variáveis de entrada que estão associadas a uma dada probabilidade, o que lhe confere caráter estocástico. Dependendo do ponto de vista, esta característica pode ser um atributo negativo.

Favilla *et al*, [86] empregaram o conceito de importância da variável na projeção (VIP, "*variable importance in the projection*") como ferramenta de seleção de variáveis acoplado a modelos N-PLS no contexto de calibração e classificação. O método VIP mede a importância de cada variável de um conjunto de J variáveis. No primeiro estudo

de caso foi investigada a composição de misturas para fabricação de pão, acompanhando o processo de produção através do registro de espectros NIR das misturas. O segundo estudo de caso envolve a classificação (em termos de origem geográfica) de amostras de azeite de oliva empregando GC-MS. Um terceiro estudo de caso empregou imagens de ressonância magnética. Para todos os casos, o método VIP mostrou ser uma ferramenta útil.

Hantao *et al,* [87] utilizaram o conceito de Taxa de Fisher ("*Fisher Rate"*) [88] para seleção de variáveis em modelos PARAFAC para análise exploratório de amostras de clones de *Eucalyptus*, empregando cromatografia bidimensional a gás acoplada a espectrometria de massas (GC× GC-MS).

Com base nos trabalhos mostrados, é possível ver que o único trabalho que propões um algoritmo de seleção de variáveis em calibração de ordem superior e destaca aspectos da vantagem de segunda ordem foi à proposta de Carneiro *et al* [85]. Mesmo considerando todos os trabalhos mencionados acima ainda são pouco expressivas as contribuições neste campo, o que justifica estudos dedicados a desenvolver novas estratégias de seleção de variáveis para calibração que emprega dados multivias.

2.6 Algoritmo das Projeções Sucessivas (SPA)

O Algoritmo das Projeções Sucessivas (SPA, "*Successive Projection Algorithm*") é um método combinatório que varre o espaço multidimensional das variáveis construindo subconjunto de variáveis que sejam minimamente correlacionadas entre si, etapa denominada fase I. Posteriormente, estas cadeias são avaliadas com base em função de custo (*Jcost*) para um dado modelo matemático específico, sendo escolhido o subconjunto de variáveis que minimiza *Jcost*.

Proposto em 2001 por Araújo e coautores [89], o SPA tinha por objetivo resolver problema de multicolinearidade em análise multicomponente simultânea por espectrometria de absorção molecular empregando regressão linear múltipla [89]. Nesta proposta, o então nomeado SPA-MLR, atuava com duas fases: geração (Fase I) e avaliação das cadeias (Fase II) [74,89].

Na fase I são conduzidas operações de projeções, de modo a obter subconjuntos de variáveis com baixa correlação entre si. Os dados empregados nesta etapa são apenas as informações registradas para as amostras de calibração ($\mathbf{X_{cal}}$ com dimensões I × J). A operação de projeções, que emprega as colunas de $\mathbf{X_{cal}}$ previamente centralizada na média, consiste de um ciclo do tipo "*forward*", para cada iteração uma nova variável é adiciona ao subconjunto inicializado com $\mathbf{x_j}$ para j variando de 1 ate J. Portanto, partindo de $\mathbf{x_j}$, para j igual a 1, fazendo $\mathbf{x}_j=\mathbf{z}^1$, calcula-se a matriz de projeção **P,** ortogonal a $\mathbf{z}^1$ como mostrado pela **Eq. 28**.

$$\mathbf{P^1} = \mathbf{I} - \frac{(\mathbf{Z^1 Z^{1T}})}{(\mathbf{Z^{1T} Z^1})} \tag{28}$$

Em que $\mathbf{P}_1$ é a matriz de projeção ortogonal a $\mathbf{z}_1$, com dimensões I × I, e **I** é uma identidade com dimensões apropriadas. Multiplicando cada coluna de $\mathbf{X_{cal}}$ por $\mathbf{P^1}$ obtem-se as projeções das demais variáveis ($\mathbf{P}^1\mathbf{x_2}$, $\mathbf{P}^1\mathbf{x}_{3...}$ $\mathbf{P}^1\mathbf{x}_j$) no plano ortogonal a $\mathbf{z}^1$. É importante lembrar que quanto maior a projeção de uma variável em **P** menor é sua correlação com $\mathbf{z}^1$. O subconjunto de variáveis iniciado com $\{\mathbf{x}_1\}$ é acrescido da variável que mostrou maior projeção em $\mathbf{P}^1$, $\{\mathbf{x}_1, \mathbf{x}_j^{SEL1}\}$. Na etapa subsequente, os vetores de projeção ($\mathbf{P}^1\mathbf{x_2}$, $\mathbf{P}^1\mathbf{x}_{3...}$ $\mathbf{P}^1\mathbf{x}_j$) das demais variáveis remanescentes são projetadas na linha ortogonal a $\mathbf{P}^1\mathbf{x}_j^{SEL1}$.

O processo é iterativo e continua até atingir um número máximo de variáveis no subconjunto iniciado com $\mathbf{x}_1$. Este número máximo está associado aos graus de

liberdade necessários para que o sistema linear, correspondente ao modelo MLR, seja possível e determinado, assumindo como valor mínimo (*Nmin*) possível 1 e máximo (*Nmax*) I-1 para dados centrados na média.

O procedimento de projeção descrito acima é reiniciado para $\mathbf{z}^{j+1}$ e repetido até $\mathbf{z}^{J}$ de modo que J subconjuntos de variáveis com comprimento *Nmax* são gerados. Os índices das variáveis que compõem cada subconjunto são armazenados em uma matriz **SEL** com dimensões *Nmax* × J.

É válido notar que o resultado da fase I do SPA consiste apenas de uma matriz que contêm (índices) subconjuntos de varáveis com baixa multicolinearidade, quando comparado ao conjunto das J variáveis. A priori, são todas candidatas a resolver o problema de inversão da matriz de covariância $\mathbf{X}^{T}\mathbf{X}$ para determinação dos coeficientes de regressão por mínimos quadrados ordinários.

Na fase II, os J subconjuntos são avaliados com respeito à correlação com a variável dependente **y**. Esta avaliação é baseada na construção de uma modelo matemático, neste caso MLR, para os subconjunto gerado na fase I. Iniciando com *Nmin* variáveis do subconjunto J até *Nmax* para o mesmo subconjunto, de modo que o número de modelos computados é dado por [(*Nmax*-*Nmin*) + 1]×J. Para cada modelo (subconjunto de variáveis) está associado um valor *Jcost*, que mede a qualidade do subconjunto J em prever adequadamente **y**. Jcost normalmente corresponde à raiz do erro médio quadrático (RMSE, *root mean square error*) que pode ser computado por validação cruzada (RMSECV, *root mean square error for cross-validation*) ou empregando um conjunto independente de amostras (RMSEV, *root mean square error of validation*).

Devido sua característica "*forward*" o SPA se torna susceptível a seleção de variáveis com baixa correlação com **y.** Este inconveniente foi corrigido adequadamente por Galvão *et al*, pela adição de uma fase III ao SPA-MLR para remoção de variáveis

não informativas [90]. Nesta etapa o subconjunto de variáveis selecionado L, é disposto em ordem decrescente de relevância. O índice de relevância de cada variável para o modelo é dado pela **Eq. 29**.

$$r_j = s_j|b_j| \tag{29}$$

Onde *r* é o índice de relevância da variável j, definido como o produto entre o desvio padrão da variável (s_j) pelo modulo do seu respectivo coeficiente de regressão (b_j). Um processo de avaliação da relevância das variáveis é feito comparado $Jcost_1^2$ para L variáveis com $Jcost_2^2$ obtido para L-1 variáveis, por meio de um teste *F* com α igual a 0.25. Se a hipótese nula (H_0) for aceita, a variável de menor valor de *r* é removida. Isso significa que o modelo para as L-1 variáveis de maiores valores de *r* produz um modelo com igual capacidade preditiva quando comparado ao modelo com as L variáveis. A avaliação da relevância prossegue até que a hipótese não nula (H_1) seja aceita no teste *F*. A fase III do SPA permite uma "limpeza" do subconjunto de variáveis previamente selecionadas na fase II, levando a modelos MLR mais simples e parcimoniosos.

O SPA conta com diversas modificações, tanto no contexto MLR [91-93] como para atuar como ferramenta de seleção de variáveis combinado a outras ferramentas quimiometricas, como classificação [94], transferência de calibração [95], seleção de amostras [96], seleção de variáveis em modelos MLR quando amostras de predição apresenta constituintes não modelados [97] e SPA com correlação ponderada [98].

Recentemente, Gomes *et al*, propuseram um algoritmo que combina seleção de variáveis em forma de intervalos com modelos de regressão PLS[99]. Esta abordagem mostrou melhores resultados quando comparado ao SPA-MLR convencional, certamente devido a melhor capacidade de modelagem de ruído e ligeiros desvios de linearidade da estrutura de regressão em variáveis latentes [99].

Um tipo de aplicação do SPA comum na literatura é o uso das vaiáveis previamente selecionadas pelo SPA-MLR ou SPA-LDA (Análise discriminante Linear) como informação de entrada para outras modelagens como PLS, Modelagem Flexível Independente por Analogia de Classe (SIMCA, "*Soft independent modelling of class analogies*"), e regressão por maquinas de suporte de vetores (SVMR, "support vector machine regression"), por exemplo [100-101]. Contudo não há evidência que subconjuntos de variáveis otimizadas para modelos LDA e MLR sejam a melhor solução para outros métodos de regressão e/ou classificação. Este tipo de aplicação deve ser usada com cautela.

Algumas modificações no SPA foram propostas simplesmente com foco computacional, ou seja, redução do tempo de cálculo. A fase II do SPA, que consiste na obtenção de diversos modelos MLR independentes entre si, foi paralelizada para diminuir o tempo de execução [92].

No contexto de aplicações, o SPA como ferramenta de seleção de variáveis, apresenta diversas contribuições na literatura [74], envolvendo diversas matrizes como combustíveis [102], fármacos [103], amostras ambientais [104], alimentos [105], bebidas [106] e dados de Relação Quantitativa Atividade Estrutura (QSAR, "*Quantitative Structure- Activity Relationship*") e Relação Quantitativa Propriedade Estrutura (QSPR, "*Quantitative Structure- Property Relationship*") [107-108].

As aplicações relatadas acima fazem usos das mais diversas técnicas analíticas instrumentais como espectrometria de absorção molecular UV-Vis [109], infravermelho [110], fluorescência molecular [111], Espectrometria de emissão em plasma induzido por laser (LIBIS, "*laser-induced breakdown spectroscopy*") [112], imagens digitais [113], dentre outras [74]. Uma versão na forma de "*freeware*" do SPA-MLR pode ser encontrada para download no sítio indicado em [114].

Capítulo III

Experimental

3. EXPERIMENTAL

3.1 Estudos de caso para avaliação do *i*SPA-N-PLS/RBL

3.1.1 Dados simulados –I

O conjunto de dados simulados –I foi empregado para avaliar método de seleção de variáveis proposto (N-*i*SPA), acoplado ao N-PLS/RBL. Este estudo simulado envolveu a determinação de dois analitos (aqui identificados como analitos A e B).

Com intuído de simular dados de cromatografia com detecção por arranjo de diodos, matrizes LC-DAD foram geradas partindo do perfil gaussiano puro de cada constituinte. Para todos os casos, as matrizes possuíam dimensões 20×20 (vinte tempos de eluição × vinte comprimentos de onda).

O conjunto de calibração foi construído (misturas simuladas de A e B) com base em um planejamento fatorial completo (dois fatores e cinco níveis, totalizando 25 amostras) com faixa de concentração de 1 a 5 unidades. O conjunto de teste foi construído com 100 amostras com concentrações randômicas de A e B entre 2 e 4 unidades. Além dos analitos A e B, foi adicionado um interferente com contribuição constante de 3,5 unidades. Neste cenário a vantagem de segunda ordem é requerida para alcançar boa acurácia.

Os perfis dos analitos (A e B) e do constituinte não modelado são mostrados na **FIGURA 11**. Os perfis representados pelas linhas sólidas azuis e verdes correspondem aos analitos A e B respectivamente, enquanto a linha sólida vermelha é o interferente presente apenas no conjunto de teste. É possível ver a forte interferência simultânea do interferente sobre o sinal dos analitos (A e B).

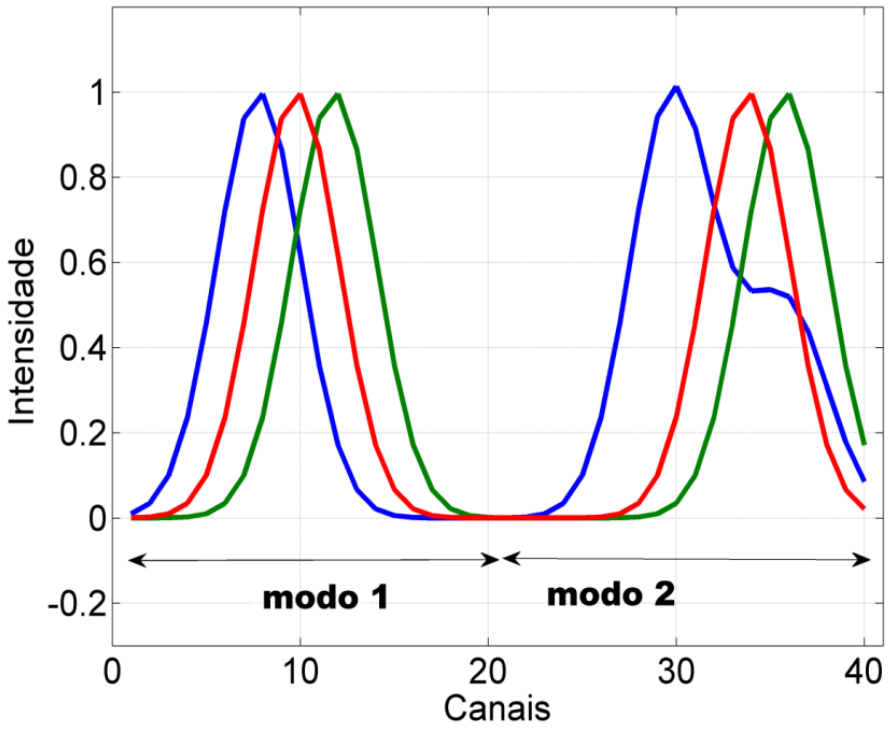


Figura 11: Perfil puro dos analito A (linha azul) e B (linha verde). A linha vermelha representa o constituinte não modelado.

O modo 1 corresponde ao modo cromatográfico, constituído de picos unimodais. O modo 2, por sua vez, representa o modo espectral. Ruído gaussiano com desvio padrão de 1%, com relação ao sinal máximo de calibração, foi adicionado em as amostras de calibração e teste. Na concentração o ruído adicionado foi de 0,01%.

3.1.2 Determinação de Ofloxacina

Este sistema experimental consistiu na determinação da quinolona ofloxacina (OFL) em amostras de água, na presença de outras duas quinolonas (ciporfloxacina-CPF e danofloxacina-DNF) não modeladas, empregando cromatografia líquida de alto desempenho com detecção por arranjo de diodos (HPLC-DAD).

As medidas cromatográficas foram conduzidas em um cromatógrafo modelo Agilent Model 1100 LC instrument (Agilent Technologies, Waldbronn, Germany, equipado com um desgaseificador, bomba quartenária, amostrador automático, forno para o

compartimento da coluna, detector com arranjo de diodos na região ultravioleta-visível e um pacote computacional (CHEMSTATION) para controle do instrumento e aquisição dos dados. A coluna cromatográfica empregada foi do tipo Zorbax Eclipse XDB-C18 com dimensões 4,6 × 7,5mm 3,5-micron (Agilent Technologies, Waldbronn, Germany). O cromatógrafo descrito acima e usado neste trabalho é ilustrado na **FIGURA 12**.

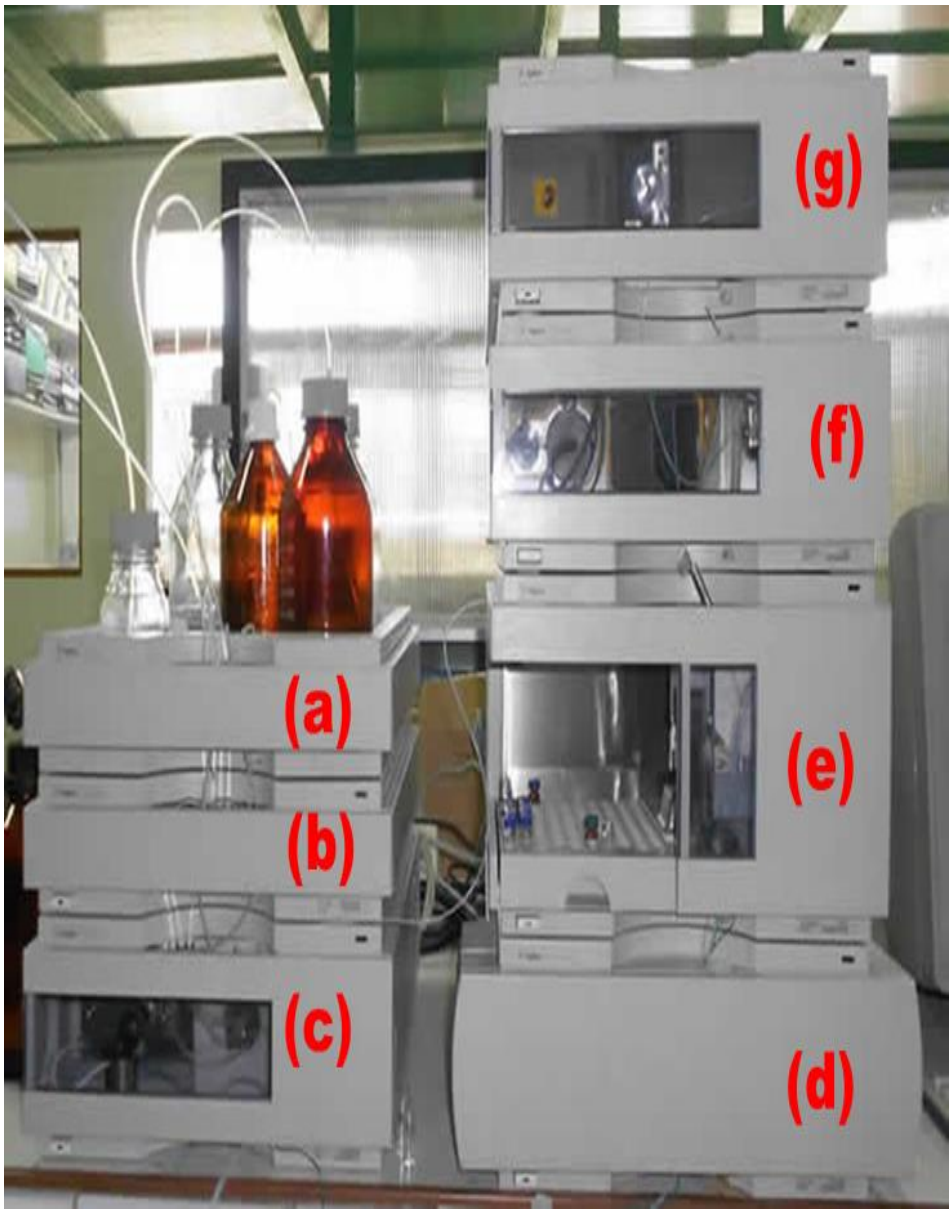


Figura 12: Ilustração do HPLC–DAD usado na aquisição de dados deste trabalho. Compartimento dos solventes (a), desgaseificador (b), bomba quaternária (c), compartimento da coluna (d), injetor automático (e), DAD (f) e detector de fluorescência (g).

A temperatura do forno da coluna foi controlada em 35 ◦C. A fase móvel empregada foi uma mistura tampão acetato de sódio/ácido acético (10 mmol L^{-1}) - metanol-acetonitrila (71:20:9, v/v). As soluções padrões de OFL foram eluídas em modo isocrático a uma vazão de 1,80 mL min^{-1}. O volume injetado foi de 100 µL.

Todos os padrões e solventes usados foram de grau analítico. OFL foi adquirido da Sigma (Germany), CPF e DNF foram comprados da Fluka (Switzerland). Metanol (MeOH) e acetonitrila ambas grau HPLC foram adquiridos junto a J.T. Baker (Deventer, The Netherlands). Água ultrapura foi obtida empregando um sistema purificado microporo Milli-Q (Bedford, MA, USA). Ácido acético grau analítico foi comprador da Cicarelli (Santa Fe, Argentina) e acetato de sódio trihidratado (NaAc . $3H_2O$) também grau analítico foi comprado da ANEDRA, (Research AG S.A., Argentina). As soluções estoques das quinolonas foram preparadas em metanol com nível de concentração de 200,0 mg L^{-1} e foram mantidas sob refrigeração a 4 ºC na ausência de luz. A partir das soluções estoque foram preparados os padrões de calibração, diluindo alíquotas adequadas em água.

O conjunto de calibração consiste de cinco padrões nas concentrações 2,0; 4,0; 6,0; 8,0 e 10,0 mg L^{-1} de OFL. O conjunto de teste foi construído fortificando doze amostras de água com diferentes concentrações de ofloxacina, ciprofloxacina e danofloxacina (veja **TABELA 3**). O tempo da corrida cromatográfica para todos os padrões e amostras foi de 2 minutos. A faixa espectral monitorada foi de 200 a 400 nm, com resolução de 1 nm, gerando matrizes tempo × absorbância com dimensões 294×201 por amostra.

Tabela 3: Composição das amostras do conjunto de teste. Concentrações estão expressas em mg $L^{-1.}$

Amostra*	OFL	CPF	DNF
1	8,83	13,24	0,79
2	6,00	9,00	1,50
3	2,00	9,00	1,50
4	3,17	4,76	1,04
5	10,00	9,00	1,50
6	3,17	13,24	2,21
7	6,00	13,24	0,50
8	6,00	15,00	2,50
8	6,00	9,00	1,50
10	8,83	4,76	2,21
11	6,00	15,00	1,50
12	6,00	9,00	2,50

*Os valores de concentração da tabela acima foram escolhidos com base nas respectivas faixas lineares.

3.2 Estudos de caso para avaliação do *i*SPA-U-PLS/RBL

3.2.1 Dados simulados- II

O conjunto de dados simulados-II visou imitar EEMs afetadas por efeito de filtro interno (EFI) de uma espécie (F) sobre o sinal do analito. Um conjunto de calibração consistindo de seis EEMs (todas com dimensões 31×31) foi gerado em triplicata (originando um tensor de calibração 18×31×31). Os perfis puros usados para obter as EEMs são mostrados na **FIGURA 13a**.

A faixa de concentração do analito está entre 1 e 6 unidades, com incremento de 1 unidade. A faixa de concentração da espécie F (linha verde) apresenta valores

randômicos entre 2 e 4. O efeito de filtro interno causado por F sobre o sinal do analito foi estimado [115] de acordo com e **Eq. 30**, e que está ilustrada na **FIGURA 13b** abaixo.

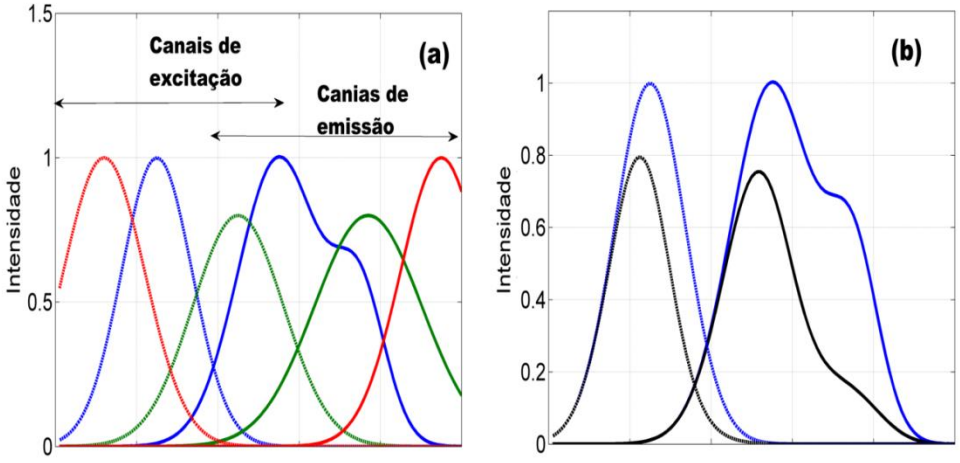


Figura 13: Perfis puros usados na construção do conjunto de dados simulado II (a) perfil puro para o (—) analito, (—) espécie F (—) constituinte não modelado. As linhas sólidas e pontilhadas são os perfis de emissão e excitação respectivamente em (b) Perfil do analito na ausência do EFI (—) e na presença da espécie F (—).

$$\mathbf{X}_{\mathrm{cal}i} = \{y_{\mathrm{cal}i}\mathbf{S}_1 \times \exp[-\mathcal{E}_{2j} + \mathcal{E}_{2k}]y_{ifi}\} + y_{ifi}\mathbf{S}_2 \qquad (30)$$

Onde $\mathbf{X}_{\mathrm{cal}i}$ é a iésima amostra de calibração com dimensões $J{\times}K$, e $y_{\mathrm{cal}i}$ é um escalar que representa a concentração do analito em $\mathbf{X}_{\mathrm{cal}i}$. $\mathbf{S_1}$ é a EEM que contem o sinal do analito em concentração unitária. O termo $\exp[-\mathcal{E}_{2j} + \mathcal{E}_{2k}]y_{ifi}$ representa a contribuição do EFI causado por F sobre o sinal do analito. $[\mathcal{E}_{2j}]y_{ifi}$, e $[\mathcal{E}_{2j}]y_{ifi}$ representam a absorção de F em cada canal dos modos J e K respectivamente, e $y_{ifi}\mathbf{S}_2$ é o sinal da espécie F na concentração y_{if}.

O conjunto de teste consistiu de cinquenta amostras em concentrações randômicas na faixa de 2 a 5 unidades para o analito e espécies F. As EEM do conjunto de teste foram geradas com base na **Eq. 31**.

$$\mathbf{X}_{\text{test}i} = \{y_{\text{test}i}\mathbf{S}_1 \times \exp[-\varepsilon_{2j} + \varepsilon_{2k}]y_{\text{test}-ifi}\} + y_{\text{test}-ifi}\mathbf{S}_2 + y_{\text{int}i}\mathbf{S}_3 \quad (31)$$

Onde o termo $y_{\text{int}i}\mathbf{S}_3$ corresponde a um constituinte não modelado (linha vermelha), adicionado apenas às amostras de teste em concentração (2 a 5 unidades) também randômica. Para todos os casos, os valores de concentração estão afetados por um ruído de 1% e no sinal de 5%.

3.2.2 Determinação de Fenilefrina

Todos os reagentes empregado foram adquiridos junto ao Laboratório de Controle de Qualidade de Medicamentos (LCQM) da Faculdade de Bioquímica e Ciências Biológicas da Universidad Nacional del Litoral/Santa Fé, Argentina. Soluções estoques de fenilefrina (FEN), paracetamol (PAR) e ácido acetil salicílico (AAS) foram preparadas dissolvendo-as em água ultra pura para obter padrões nas concentrações de 100, 200 e 200 mg L^{-1} respectivamente. A solução estoque de ibuprofeno (IBU) foi preparada em metanol. Na sequência 1 mL da solução metílica de IBU foi transferida para uma balão volumétrico de 10 mL, o metanol foi evaporado sob fluxo de nitrogênio, o balão foi aferido com água ultra pura, de modo que a concentração de IBU fosse de 200 mg L^{-1}, esta foi considerada a solução de trabalho do IBU.

O conjunto de calibração foi preparado pela diluição de volumes apropriados da solução estoque de FEN de modo a obter as seguintes concentrações: 0,248; 0,379; 0,496 0,627; e 0,744 μg mL^{-1} em duplicata. Em todos os padrões de calibração foram adicionados alíquotas de PAR, de modo que sua concentração final em todos os casos foi de 10,00 μg mL^{-1}.

O conjunto de amostras de teste foi preparado segundo um planejamento composto central fracionário (2^{4-1}) com quatro fatores e cinco níveis. As concentrações de todas as amostras do conjunto de teste são mostradas na **TABELA 4**.

Tabela 4: Composição das amostras do conjunto de teste. Todas as concentrações estão expressas em μg mL^{-1}.

Amostra	FEN	Fator 2	Fator 3	Fator 4
1	0,277	0,280	0,545	8,04
2	0,365	0,280	0,545	17,0
3	0,277	0,369	0,545	17,0
4	0,365	0,369	0,545	8,04
5	0,277	0,280	0,743	17,0
6	0,365	0,280	0,743	8,04
7	0,277	0,369	0,743	8,04
8	0,365	0,369	0,743	17,0
9	0,248	0,324	0,644	12,5
10	0,394	0,324	0,644	12,5
11	0,321	0,250	0,644	12,5
12	0,321	0,400	0,644	12,5
13	0,321	0,324	0,495	12,5
14	0,321	0,324	0,792	12,5
15	0,321	0,324	0,644	5,0
16	0,321	0,324	0,644	20,0
17	0,321	0,324	0,644	12,5

Todas as medidas espectroscópicas foram feitas usando um Perkin-Elmer LS-55 luminescence spectrometer (veja **FIGURA 14**) equipado com uma lâmpada de descarga de xenônio, monocromadores do tipo Monk-Gillieson e uma fotomultiplicadora conectada ao micro computar via cabo serial RS232C. Em todas as medidas foi empregada celula de quartzo com caminho óptico de 1,0 cm.

As matrizes excitação emissão forram registradas varrendo a excitação de 215 a 240 nm com resolução de 2 nm e registrando os espectros de emissão na faixa de 270 a 360 nm com resolução de 0,5 nm, de modo que cada EEM registrada possui dimensões 181 × 13. A largura de banda da fenda dos monocromadores de excitação e emissão foi mantida fixa em 10 nm e o detector operando com 650 V.

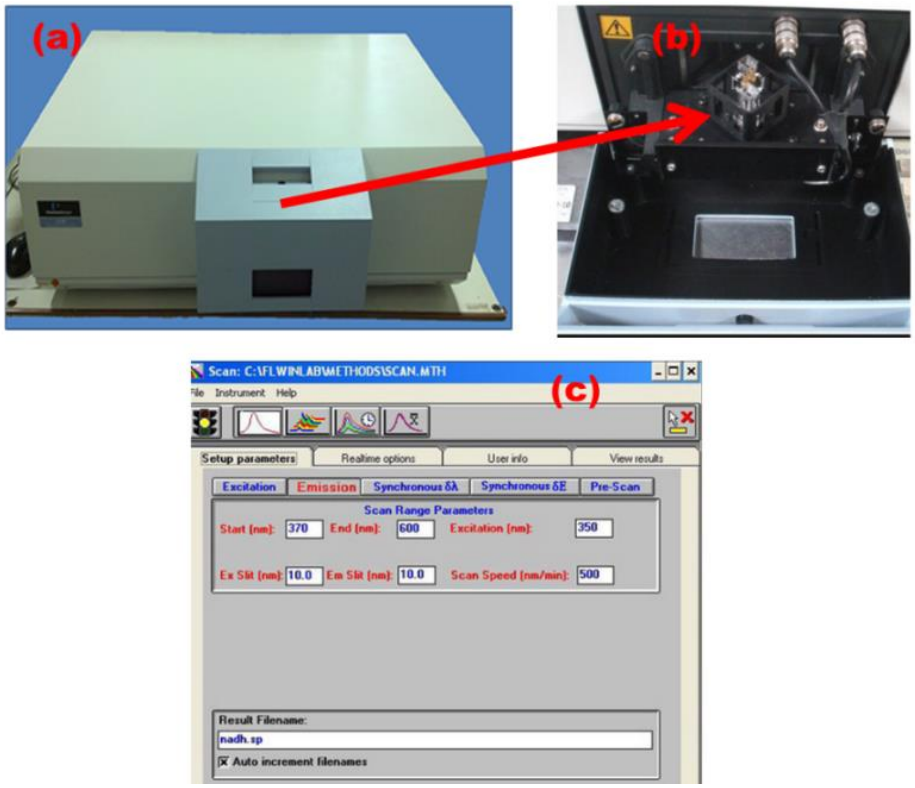


Figura 14: Ilustração (a) do Perkin-Elmer LS-55 luminescence spectrometer usado para geração das EEM. Em (b) é mostrado o compartimento da amostra e em (c) a janela principal do software de controle e aquisição de dados.

3.3 Softwares utilizados

O algoritmo N-*i*SPA desenvolvido neste trabalho emprega programação em ambiente MatLab®. Demais cálculos envolvendo modelos PARAFAC, U-PLS/RBL e N-PLS/RBL foram realizados empregando a interface gráfica MVC2 [116], implementada por Olivieri e colaboradores e disponível em www.iquir.conicet.gov.ar /descragas/ mvc2.rar.

A janela principal do pacote MVC2 é mostrada na **FIGURA 15a**. Nesta janela é possível carregar os dados em formato TXT, plotar os gráficos referentes a sinal instrumental das amostras e efetuar a modelagem dos dados. Diferente de outros pacotes dedicados ao tratamento de dados multivias, MVC2 foi planejada com foco no desenvolvimento de métodos quantitativos.

Após a escolha do método de modelagem na janela da **FIGURA 15a**, a aba específica do método selecionado é mostra como na **FIGURA 15b**. Ao descarregar o pacote MVC2, que é um "*freeware*" o usuário conta com um manual completo.

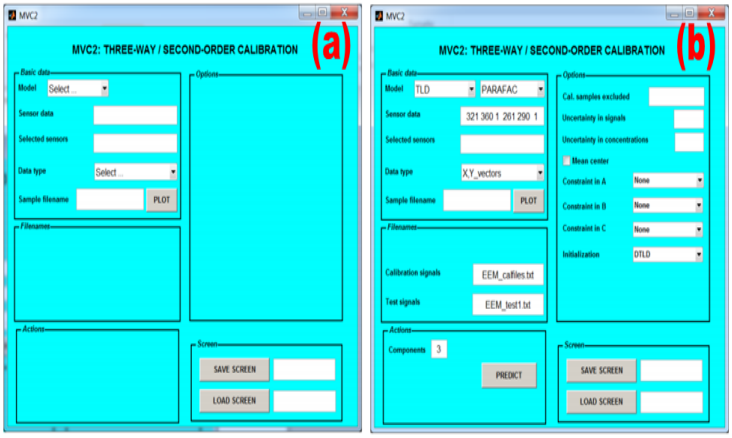


Figura 15: Em (a) é mostrado à janela principal do pacote MVC2 e em (b) a janela com botões específicos para modelagem PARAFAC.

Os cálculos envolvendo o GA-N-PLS/RBL foram conduzidos empregando a rotina desenvolvida e disponibilizada por Carneiro *et al* [85], adaptada para atuar em conjunto com o N-PLS/RBL. Codificação binária, com população inicial de 100 cromossomos gerados aleatoriamente foi empregada. A quantidade de variáveis iniciais em cada modo foi de 10%. Para todos os casos, a probabilidade de reprodução e mutação em ambos os modos foi de 50% e 1% respectivamente. Após 100 gerações os indivíduos mais aptos (variáveis selecionadas) são escolhidos com base na minimização do RMSECV.

Capítulo IV

Algoritmo Proposto

4. ALGORITMO PROPOSTO

4. 1 Descrições do funcionamento

O algoritmo proposto neste trabalho é uma extensão do *i*SPA para seleção de intervalos em regressão PLS, proposto por Gomes *et al* em 2012 [99]. Adaptações foram feiras de modo que o SPA possa atuar como ferramenta de seleção intervalos em calibração multivias, associado a modelos U-PLS e N-PLS. O novo algoritmo foi denominado de N-*i*SPA. Em ambos os casos é usado RBL para alcançar vantagem de segunda ordem.

Inicialmente, a estrutura de dados tridimensional I×J×K é desdobrada em função dos modos instrumentais J e K. Este procedimento gera as matrizes **uXcal-1** e **u-Xcal-2** com dimensões IK×J e IJ×K respectivamente. Tomando com exemplo dados do tipo EEM, em que o modo J é a excitação e o modo K a emissão, a matriz **uXcal-1** tem na dimensão J os espectros de excitação. Enquanto na dimensão IK estão arranjados (um abaixo do outro) os J espectros de emissão. A matriz **u-Xcal-2** tem na dimensão K os espectros de emissão e na dimensão IJ os K espectros de excitação.

As matrizes **uXcal-1** e **u-Xcal-2** consistem da informação de entrada para fase I do N-*i*SPA. Estas matrizes são particionadas em uma quantidade de intervalos arbitraria, que é definida pelo analista e otimizada automaticamente. Assumindo que J variáveis (modo 1), j_1, j_2, j_3,...,j_J e K variáveis (modo 2) k_1, k_2, k_3, ..., k_K são particionados em s^1 e s^2 intervalos não sobrepostos de comprimento s^1_1, s^1_2, s^1_3, ...,s^1_{w1} e s^2_1, s^2_2, s^2_3, ...,s^2_{w2} respectivamente. Geralmente os intervalos tem o mesmo comprimento, em caso de variáveis remanescentes, estas são distribuídas nos primeiros intervalos de modo que $s^1_1+s^1_2+s^1_3+ \ldots+s^1_{s1}=J$ e $s^2_1+s^2_2+s^2_3+\ldots+s^2_{s2}=K$.

Para cada um dos s^1 e s^2 intervalos são selecionadas as variáveis de maior norma para cada intervalo no modo J e no modo K que dão origem as matrizes **iuXcal-1** e **iuXcal-2** com dimensões IK×s^1 e IJ×s^2. Sob as colunas de **iuXcal-1** e **iuXcal-2** são conduzidas as operações de projeção do SPA como descrito em **2.6 Algoritmo das Projeções Sucessivas**. Este procedimento de projeções gera as matrizes **SEL-1** e **SEL-2** como resultado da fase I do N-*i*SPA. Em **SEL-1** e **SEL-2** são armazenados os índices dos subconjuntos de variáveis de maior norma minimamente correlacionadas, que são as "representantes" do seu respectivo intervalo. As matrizes **SEL-1** e **SEL-2** possuem dimensões s^1-1×s^1 e s^2-1×s^2 respectivamente. Na **FIGURA 16** é apresentada uma representação gráfica da fase I do N-*i*SPA.

Na fase II do N-*i*SPA, o usuário deve indicar o modelo de regressão para avaliação dos subconjuntos de intervalos. São disponíveis PLS para dados desdobrados (U-PLS) e PLS multidimensional (N-PLS). O tipo de regressão indicado pelo usuário é usado para construir modelos para cada combinação de intervalos armazenados em **SEL-1** e **SEL-2**. Para cada combinação de intervalos em **SEL-1** todas as combinações em **SEL-2** são avaliadas em ciclos aninhados.

Quando U-PLS é selecionado pelo usuário como método de regressão, a cada iteração da fase II do N-*i*SPA, o tensor I×J×K é atualizado para conter apenas os intervalos definidos pelos controladores de fluxo nos modos J e K. Na sequência, este subtensor é disposto como matriz desdobrada I×s^1s^2 sobre a qual é computado um modelo U-PLS com validação cruzada completa ("*full cross-validation*"). De modo similar é realizado para o método N-PLS, contudo a estrutura de cubo dos dados é mantida. Nesta etapa os erros de validação cruzada (ou por conjunto externo de validação) e o número ótimo de fatores PLS são computados e armazenados.

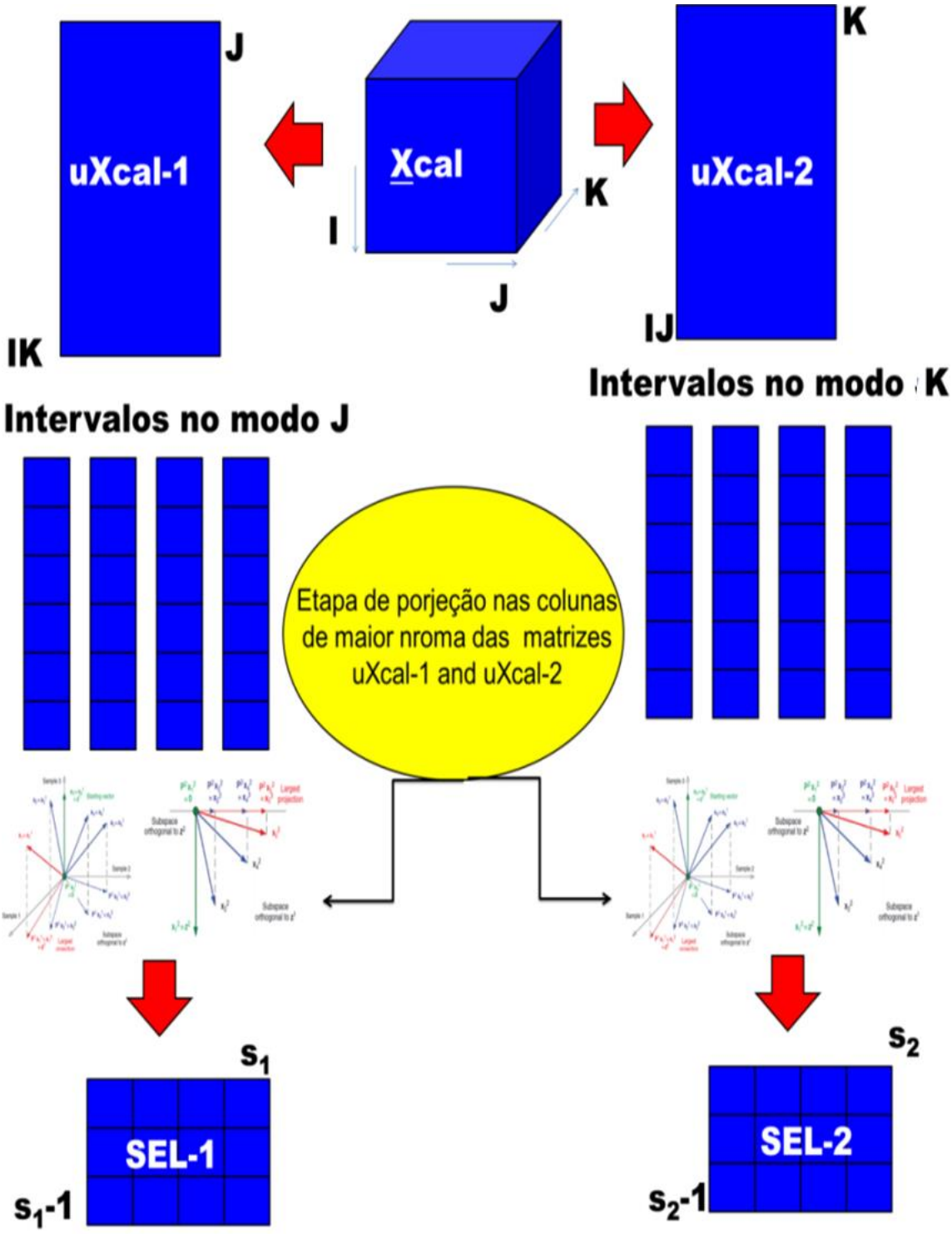


Figura 16: Representação gráfica da fase um do N-*i*SPA.

Com mencionado em ***2.4.2.3 Bilinearização Residual*****,** os métodos N-PLS e U-PLS não possuem a vantagem de segunda ordem, contudo esta pode ser alcançada empregando etapa RBL. A utilização ou não do procedimento de bilinearização dos

resíduos das amostras de predição com intuito de alcançar vantagem de segunda ordem deve ser indicado pelo analista.

Se RBL (l=0) não é empregado, as amostras de calibração e validação possuem a mesma composição. Então para cada subconjunto de intervalos é calculado o valor de *Jcost* (que pode ser o RMSECV ou RMSEV). Um único subconjunto de intervalos nos modos J e K que minimizem o valor de *Jcost* é selecionado.

Para l>0, casos em que a etapa RBL é empregada, para cada ciclo da fase II a amostra de predição **Xu** (J×K) é atualizada para os intervalos em consideração no modos J e K. A predição de **Xu** é conduzida empregando de 1 a l fatores RBL, o valorde s_u associado a cada predição é combinado com o valor de RMSECV ou RMSEV gerando um novo valor de *Jcost*, como indicado na **Eq. 32**.

$$j_{cost} = \sqrt{\frac{(y_i - \hat{y}_i)^2}{I}} + R \tag{32}$$

$$R = |1 - \frac{S_u}{S_{cal}}| \tag{33}$$

O termo R na **Eq. 33** é a comparação de s_u (resido instrumental de $\mathbf{X_u}$ para l fatores RBL) com s_{cal} (como resíduo típico das amostras de calibração). Um bom intervalo para predizer y_u em $\mathbf{X}_u$ deve ter variáveis de alta correlação com y_u e regiões com baixa contribuição dos interferentes, ou que estes, quando presentes, estejam devidamente modelados pela etapa RBL. Portanto, a correlação dos intervalos com **y** é avaliada pelo erro médio de predição (RMSECV), enquanto a modelagem adequada dos constituintes inesperados presentes em $\mathbf{X_u}$ é avaliado por R.

Note que um bom valor de R ocorre quando s_u e s_{cal} são muito próximos, sugerindo que o resíduo da amostra de predição e das amostras de calibração são equivalentes. Quando s_u é maior que s_{cal} (interferentes não modelados adequadamente) R aumenta e os intervalos em consideração não são selecionados para *l* RBL. Quando s_u é muito pequeno quando comparado a s_{cal} é indicativo que uma quantidade excessiva de fatores RBL está sendo empregada, isso torna R praticamente igual a 1, e os intervalos em avaliação para o respectivo número de fatores RBL são evitados. À medida que s_u e $_{scal}$ se tonam parecidos (modelagem adequada dos interferentes) R tende a zero. E a minimização do erro médio de predição para o conjunto de calibração passa a governar a seleção dos intervalos. O erro de predição em *Jcost* e o termo R podem ser normalizados para se tornar comparáveis.

Obviamente que em análise de amostras reais complexas, a composição de amostra para amostra pode variar. Assim estrutura flexível de *Jcost* permite a seleção de diferentes subconjuntos de intervalos para cada amostra do conjunto de teste. Isso permite que o grau de sobreposição do sinal dos interferentes sobre o sinal do analito seja avaliado caso a caso. Na **FIGURA 17** é ilustrado o funcionamento da fase II.

4.2 N-*i*SPA Tool Box: Linhas de comando

O algoritmo de seleção de variáveis desenvolvido neste trabalho foi inicialmente implementado em linhas de comando em arquivos do tipo "*file.m*" em ambiente MatLab. O pacote denominado de N-*i*SPA ToolBox, permite ao usuário modelar dados de segunda ordem com U-PLS e o N-PLS, ambos com bilinearização residual e seleção de intervalos. O algoritmo é composto de duas rotinas principais que executam a seleção de intervalos via *i*SPA, denominadas *i*SPA-U-PLS/RBL e *i*SPA-N-PLS/RBL. Na **TABELA 5** são mostradas as sub-rotinas que compões o N-*i*SPA e suas funções.

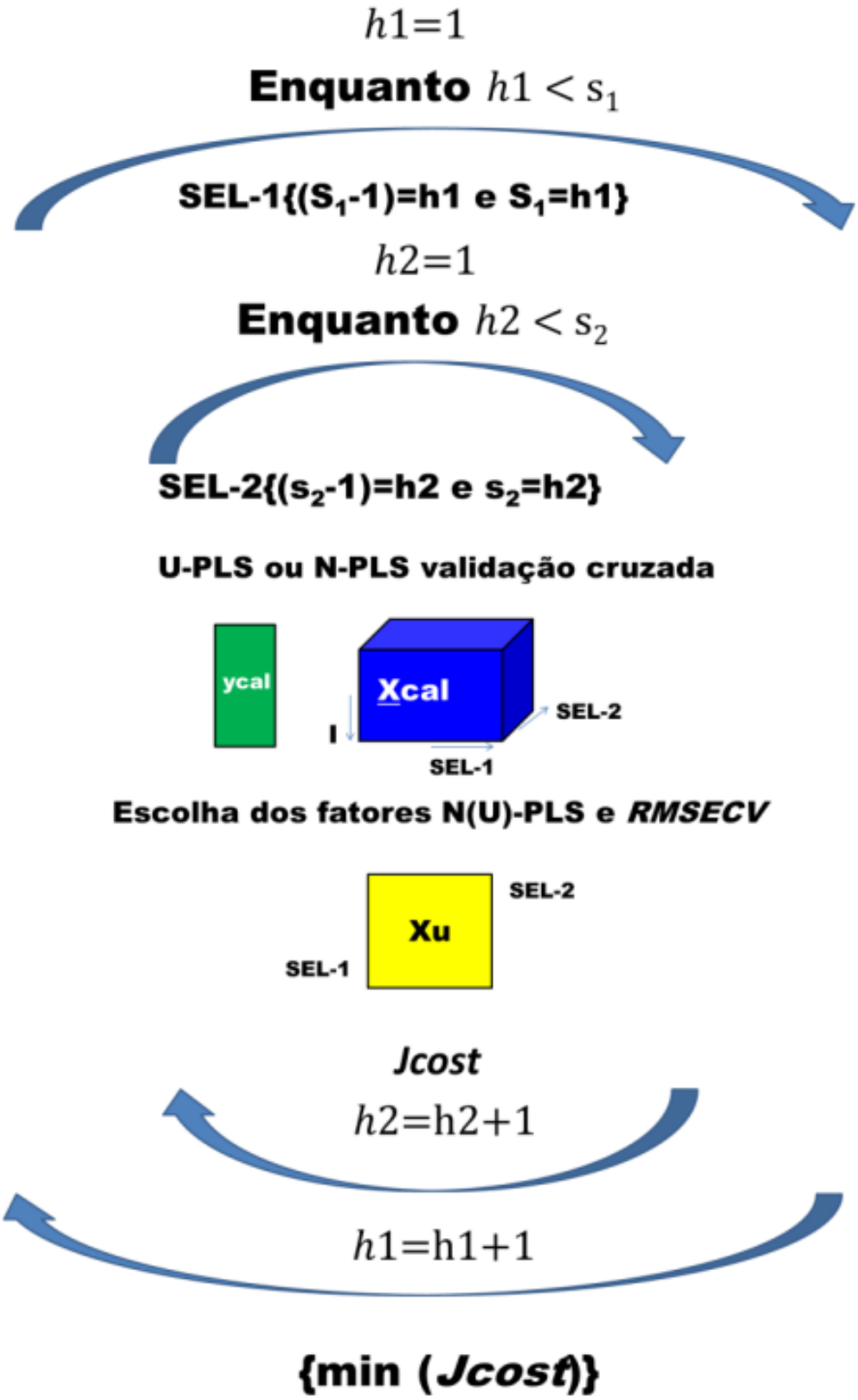


Figura 17: Esquema de funcionamento da fase II do N-*i*SPA.

Tabela 5: Relação dos arquivos.m que compões o pacote N-iSPA.

Aquivo.m	Função	Disponível em:
iSPA-N-PLS/RBL	Rotina principal	Implementada pelo autor
iSPA-U-PLS/RBL	Rotina principal	implementada pelo autor
N-PLS*	Calcula parâmetros do modelo N-PLS	http://www.models.kvl.dk/algorithms
N-PLS-CV*	Calcula modelo N-PLS com cv	www.iquir.conicet.gov.ar/descragas/mvc2.rar.
U-PLS-CV*	Calcula modelo U-PLS com cv	www.iquir.conicet.gov.ar/descragas/mvc2.rar.
RBL para U-PLS*	Etapa RBL para U-PLS	www.iquir.conicet.gov.ar/descragas/mvc2.rar.
RBL para N-PLS*	Etapa RBL para N-PLS	www.iquir.conicet.gov.ar/descragas/mvc2.rar.
solver-U-PLS	Minimização Gauss-Newton para U-PLS	www.iquir.conicet.gov.ar/descragas/mvc2.rar.
solaver-N-PLS	Minimização Gauss-Newton para N-PLS	www.iquir.conicet.gov.ar/descragas/mvc2.rar.

*Rotinas com adaptações feitas pelo autor.

4.3 N-iSPA ToolBox: Interface Gráfica

O N-*i*SPA foi implementado também em de forma interface gráfica para tornar sua difusão e utilização mais amigável. A interface denomina de "**nispa_gui**" é executada em ambiente MatLab, e quando inicializada, a seguinte janela é apresentada como mostrado na **FIGURA 17**.

Nesta interface o usuário pode carregar os dados de calibração e predição, visualizar graficamente os dados por meio do comando "PLOT". Empregando um "*menu popup*" o usuário pode selecionar o método de calibração (U-PLS ou N-PLS), além de informar dados de entrada como número de variáveis latentes, fatores RBL e número de intervalos. Ao final do calculo um relatório de métricas de desempenho é salvo no "*workspace*" do MatLab e exibido na "*command window*". Além das saídas numéricas gráficas são geradas, como intervalos selecionados e região elíptica de confiança.

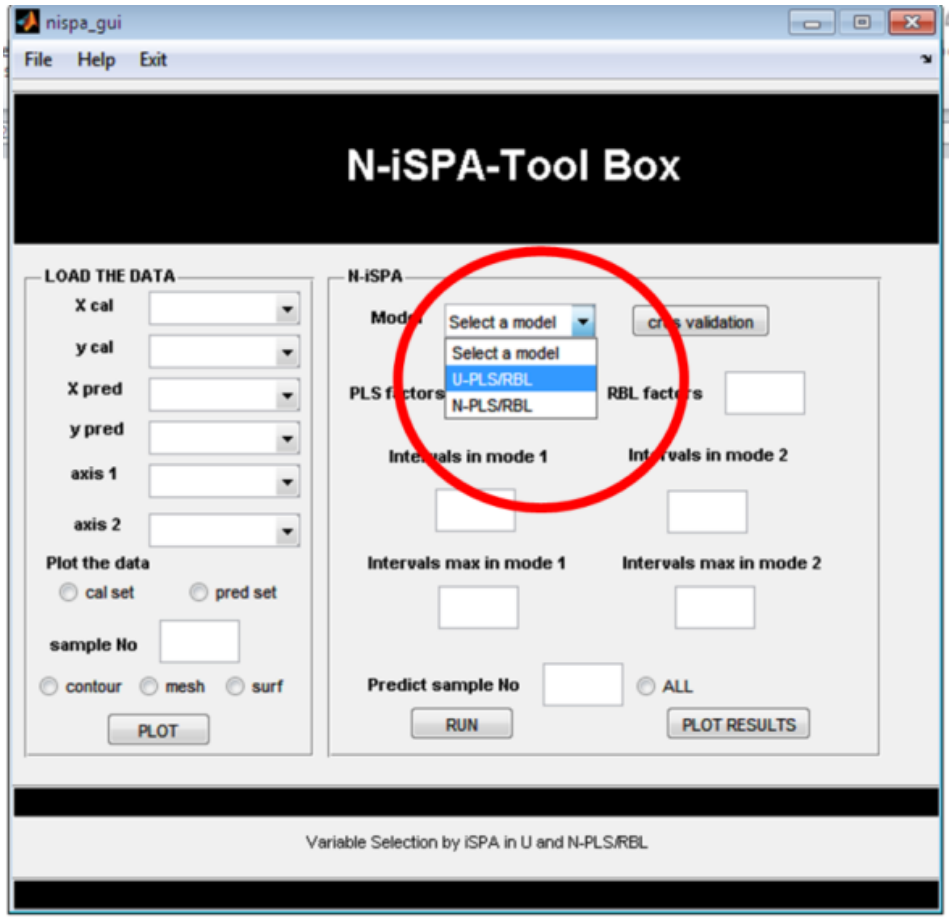


Figura 18: Ilustração da interface gráfica do N-*i*SPA**.**

Capítulo V

Resultados e Discussão

5. RESULTADOS E DISCUSSÃO

5.1 *i*SPA-N-PLS/RBL

O algoritmo *i*SPA-N-PLS/RBL foi avaliado em dois estudos de casos. O primeiro envolve o uso de dados simulados para determinação de dois analitos em presença de um único interferente. Enquanto segundo estudo de caso, envolve a determinação de ofloxacina em amostras de água na presença de dois interfentes.

5.1.1 Dados Simulados-I

Os dados simulados são compostos pelo conjunto de calibração (25×20×20) contendo dois analitos (A e B) e o conjuntos de testes com interferente. Na **FIGURA 19a** é ilustrada uma superfície simulada de um sinal registrado em sistema LC-DAD típico. Inicialmente os dados de calibração foram modelados via validação cruzada completa para cada método, a fim de identificar o número ótimo de variáveis latentes. Os resultados obtidos são mostrados na **FIGURA 19b** e **19c**.

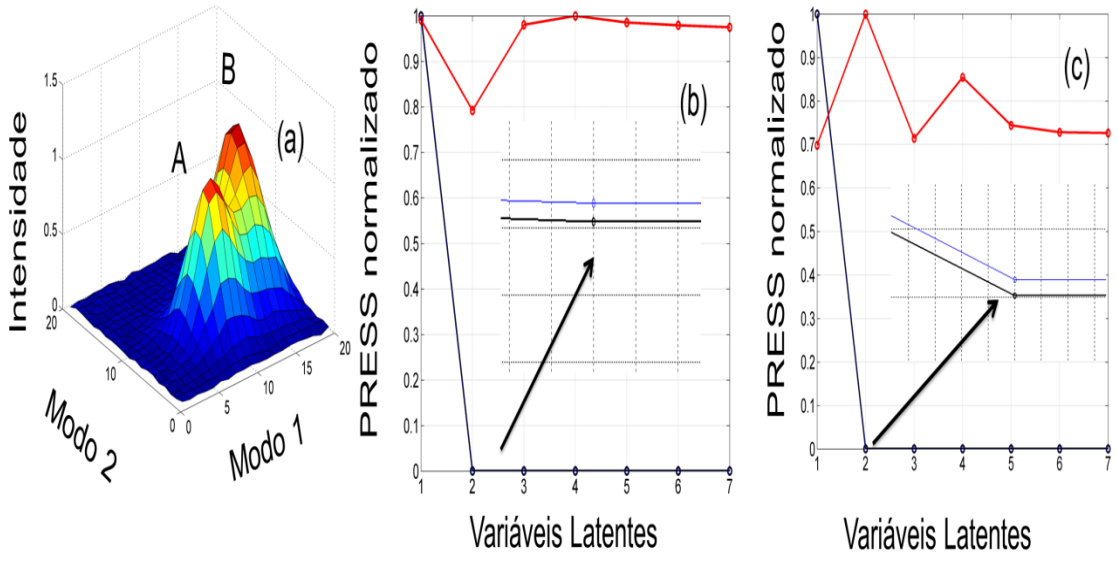


Figura 19: Em (a) sinal simulado típico das amostras de calibração e variação do PRESS normalizado em função do número de variáveis latentes incluídas no modelo para analito (b) A e analito (c) B. A linha preta está relacionado ao modelo N-PLS, a linha azul ao modelo *i*SPA-N-PLS e a linha vermelha ao modelo GA-N-PLS.

Observando as **FIGURAS 19b** e **19c** é possível ver que o número ideal de fatores para os modelos N-PLS (linha preta) é igual a dois. O mesmo ocorre para os modelos baseados em seleção de intervalos com o método proposto (linha azul). Os dois ajustes são muito similares e nas **FIGURAS 19b** e **19c** são mostradas por meio de uma ampliação a variação da soma quadrática do erro de predição (PRESS, "*predicted residual sum of squares*") em função do número de variáveis latentes, em torno de dois.

O Ajuste com o GA-N-PLS (linha vermelha nas **FIGURAS 19b** e **19c**) mostrou um curva atípica. Para o analito A (**FIGURAS 19b**) o GA-N-PLS mostrou um mínimo bem definido para dois fatores. Por outro lado para o analito B, a curva de PRESS versus número de fatores sugere um número maior de fatores. Este comportamento pode estar associado ao fato do GA selecionar váriaveis em regiões de baixa magnitude do sinal e de forma descontinua.

Para todos os casos, dois fatores foram empregados, embora para o GA-N-PLS não foi tão evidente como no caso dos modelos N-PLS e *i*SPA-N-PLS. Éválido lembrar, que o valor ótimo de variáveis latentes é indicado não somente pela visualização gráfica da **FIGURA 19**, mas também baseado em um teste *F*, como sugerido por Halland e Thomas [117] bem como co conhecimento químico do analista a cerca do sistema sob investigação.

Na sequência, foi conduzida a predição das amostras do conjunto de teste. Em todos os casos foi empregado um único fator RBL, correspondente ao único interferente adicionado as amostras do conjunto de teste (veja **FIGURA 11**). A escolha deste valor, além de ser igual ao número de interferentes, foi justificada pela inspeção do gráfico (**FIGURA 20**) que mostra a variação típica da norma do resíduo (s_u) das amostras de teste em função do número de fatores (N_i) RBL.

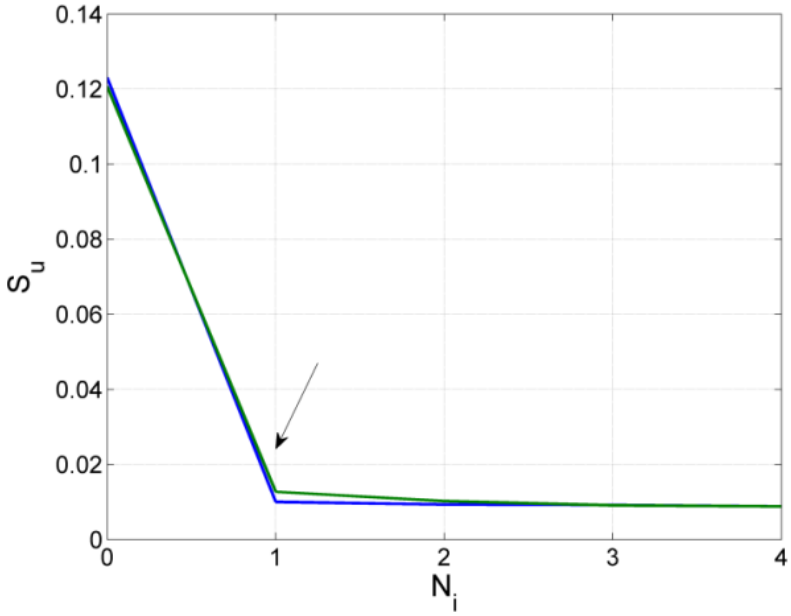


Figura 20: Variação do resíduo da amostra de teste em função da adição de fatores (N_i) para o modelo global. A linha azul corresponde ao analito A e a linha verde ao analito B.

Com base no gráfico da **FIGURA 20**, é possível observar que o ajuste do resíduo das amostras de teste, para um fator RBL, é bastante semelhante ao resíduo típico do conjunto de calibração obtidos com base na **Eq. 21** (por volta de 0,010 e 0,013 para os analitos A e B, respectivamente). Para fatores RBL adicionais não ocorre variação significativa de s_u. Empregando duas variáveis latentes para modelar o conjunto de calibração e um fator RBL para alcançar a vantagem de segunda ordem as amostras de teste foram preditas e os resultados são apresentados na **TABELA 6**.

Com base na **TABELA 6**, vemos que todos os modelos foram hábeis para predizer as concentrações das amostras do conjunto de teste. Para todos os modelos valores de RMSEP comrrspondendestes a erro de predição relativo (REP, "*relative error prediction*") inferiores a 2% foram obtidos em todos os casos. Contudo, resultados melhores foram obtidos quando se empregou seleção de variáveis.

Para o analito A, o melhor resultado foi obtido empregando o método proposto. Por outro lado o GA-N-PLS/RBL mostrou o maior valor de RMSEP, inclusive ligeiramente supeiror ao modelo N-PLS/RBL. Para o analito B, ambos os modelos com seleção de variáveis mostraram-se melhores quando comparado ao modelo N-PLS/RBL, sendo no entanto o menor RMSEP obtido pelo GA-N-PLS/RBL.

Tabela 6: Resultados* da predição para os dados simulados: conjunto de teste II (unidades arbitrárias).

Modelo	Analito	RMSEP $\times 10^{-3}$	SEN	γ	LOD$\times 10^{-2}$
N-PLS/RBL	**A**	6,7	4,49	36	1,5
	B	47,6	3,11	26	2,8
***i*SPA-N-PLS/RBL**	**A**	6,0	1,68	14	2,0
	B	14,0	1,74	15	3,6
GA-N-PLS/RBL	**A**	10,6	1,27	23	1,9
	B	11,4	0,84	4,5	4,2

*Media do conjunto de teste

Ainda com respeito à acurácia dos modelos mostrados na **TABELA 6**, na **FIGURA 21a** e **21b** são mostradas as regiões elípticas de confiança conjunta (EJCR). Estas regiões elípticas correspondem aos intervalos de confiança conjunta do coeficiente angular e linear obtido pelo ajuste de uma reta por OLS entre valores nominais e preditos de cada modelo. Para um modelo que não apresenta *bias* significativo, a elipse deve conter o ponto ideal (1 e 0). Observando as EJCR (**FIGURA 21**) obtidas para o conjunto de teste, vemos que estas estão em concordância com os valores de RMSEP da **TABELA 6**.

É possível ver na **FIGURA 21a** que a elipse relacionada ao método proposto para o analito A apresenta a menor área em comparação com as demais. No caso do Analito B

a menor elipse é obtida para o GA-N-PLS/RBL e o modelo global apresenta um bias significativo, a elipse correspondente não contem o ponto ideal.

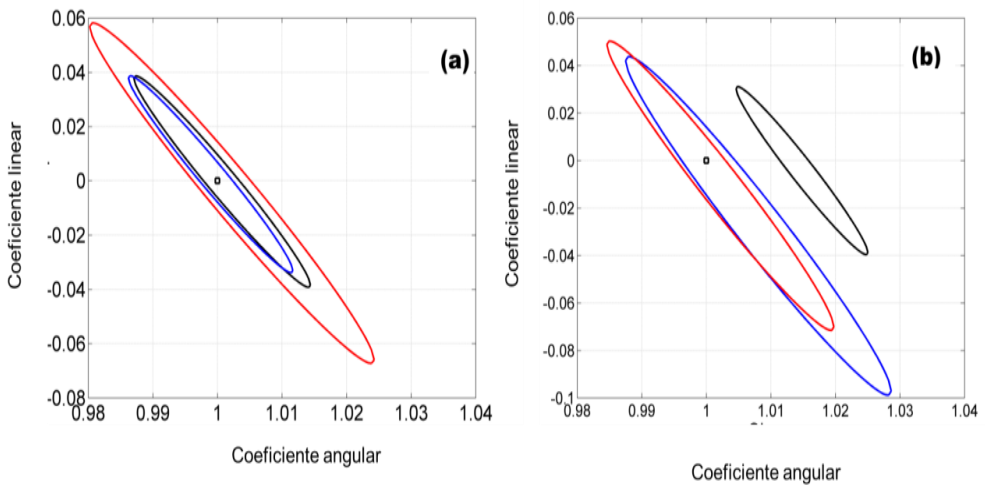


Figura 21: EJCR para o analito A (a) e analito B (b) obtidas para os modelos (linha preta) N-PLS/RBL, (linha azul) *i*SPA-N-PLS/RBL e (linha vermelha) modelo GA-N-PLS/RBL.

Com respeito à sensibilidade (SEN) vemos que a escolha de canais mais seletivos gera modelos mais exatos (pelo menos no caso do *i*SPA-N-PLS/RBL), contudo essa redução pode leva a perda de sensibilidade. Note (veja **TABELA 6**) que o modelo N-PLS/RBl que emprega todos os canais disponíveis apresenta maiores valores de sensibilidade enquanto o modelo baseado em GA apresentou os menores valores de sensibilidade.

A definição de sensibilidade analítica (γ) proposta por Olivieri e adotada neste trabalho [118] leva em consideração o resíduo deixada pelo modelo, ou seja a sensibilidade analítica é definida como a razão entre a SEN e norma do resíduo $\|\mathbf{E}\|$. Logo um melhor ajuste com menos canais pode levar a melhores valores de γ. Ainda na

TABEL 6 vemos que os valores de limite de detecção (LOD) estão condizentes com os valores de SEN obtidos.

Por fim, na **FIGURA 22** são apresentados os canais selecionados em cada caso. A inspeção das variáveis selecionadas ajuda a entender os resultados obtidos neste estudo de caso. A primeira consideração que pode ser feita é com respeito à presença de *bias* significativo no modelo N-PLS/RBL para o analito B. Observando as **FIGURA 22a** e **22b**, é possível ver que o perfil do constituinte não modelado (linha vermelha) se sobrepõe de forma mais acentuada ao analito B (linha verde), isso pode explicar o *bias* observado para o modelo N-PLS/RBL. Se olharmos a posição da elipse do modelo N-PLS/RBL para o analito B vemos que a mesma está posicionada sobre o ponto ideal, sugerindo uma contribuição aditiva ao sinal do analito, ou seja, um *bias* positivo.

Observe (veja **FIGURA 21**) que nos modelos com seleção de variáveis o *bias* está ausente, devido à seleção de sensores mais seletivos, sugerindo que a seleção de variáveis é capaz de contornar problema de *bias* pouco acentuado pela seleção de canais mais seletivos.

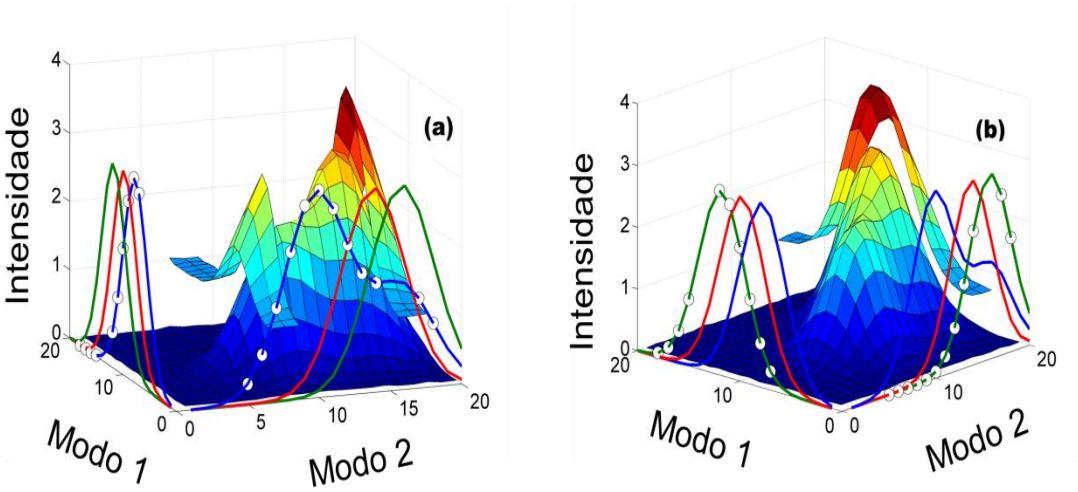


Figura 22: Superfície correspondente ao sinal típico das amostras do conjunto de teste. Deslocado por *offset* é mostrado o intervalo selecionado pelo *i*SPA-N-PLS. As linhas solidas azuis e verdes correspondem aos sinais puros empregados para gerar os dados simulados e as esferas brancas são as variáveis selecionadas pelo GA-N-PLS. Resultado para o (a) Analito A e (b) Analito B.

A segunda consideração está relacionada com o melhor resultado obtido pelo o modelo GA-N-PLS/RBL para os analitoa A e B quando comparado aos demais modelos. Inicialmente observando a **FIGURA 22a** é possivel observar que sobre o modo 1, o GA seleciona apenas variveis na região de sobreposição do sinal do analito A com os demais constituintes do sistema. E para o analito B observa-se na **FIGURA 22b** que enquanto o *i*SPA-N-PLS/RBL seleciona um uma faixa estreita em torno do máximo do sinal, o GA-N-PLS-RBL além de variáveis em torno do máximo do sinal do analito B, seleciona variáveis em região de baixa relação sinal ruído (em especial no modo 2). Contudo estas variáveis correspondem à região de menor contribuição do sinal da espécie não modelada sobre o sinal do analito B, conseuqnetemente este sinal, ainda que com menor maguinitude é mais seletivo. Este comportamento do GA pode explicar os valores de RMSEP obtidos em cada caso.

5.1.2 Dados LC-DAD: Determinação de Ofloxacina

Neste estudo de caso foi conduzida a quantificação de ofloxacina (OFL) em amostras de água por meio da modelagem de matrizes LC-DAD. A OFL (veja estrutura molecular na **FIGURA 23a**) é um composto orgânico pertencente ao grupo das quinolonas. As quinolonas (ou fluorquinolonas, como também são conhecidas por suas propriedades fluorescentes) são antibióticos usados no combate de infecção bacteriana.

Na **FIGURA 23a** é mostrado o cromatograma obtido, segundo as condições de eluição descritas na metodologia deste trabalho, para OFL com tempo de retenção de 0,85 minutos. Na **FIGUAR 23b** é mostrado o perfil espectral (normalizado) da OFL em que é possível ver dois picos, o primeiro em torno de 230 nm e o segundo, e mais intenso, por volta de 290 nm. Já na **FIGUAR 23c**, é mostrado a superfície de contorno LC-DAD obtidos para a amostra de calibração em maior concentração.

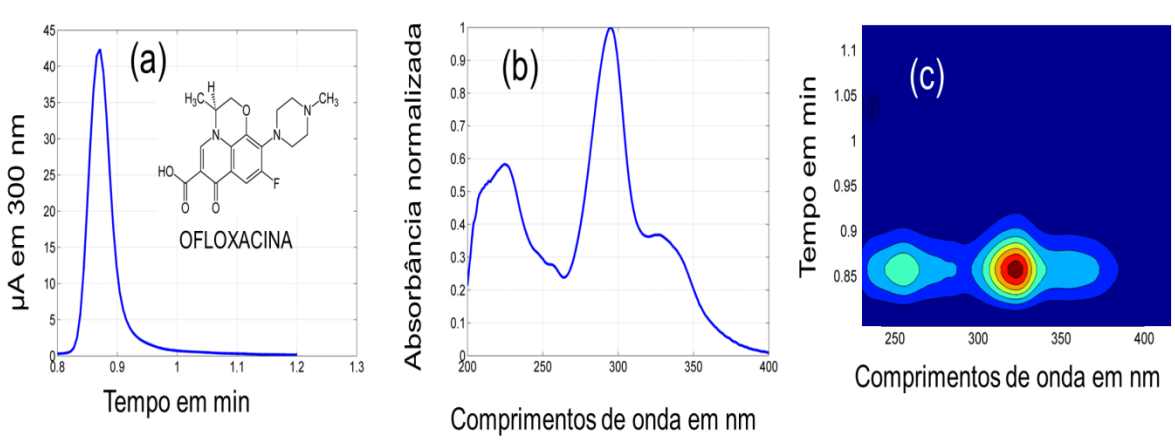


Figura 23 Informações referentes à quinolona ofloxacina quantificada neste estudo de caso. Em (a) é mostrado o cromatograma em 300 nm, (b) o espectro normalizado e (c) um superfície de contorno na concentração de 10 mgL-1

As amostras do conjunto de calibração são padrões puros de OFL em água que foram modelados empregando o N-PLS com validação cruzada. Os resultados obtidos (mostrado na forma dos gráficos da **FIGURA 24**) foram empregados para escolher o número de fatores N-PLS.

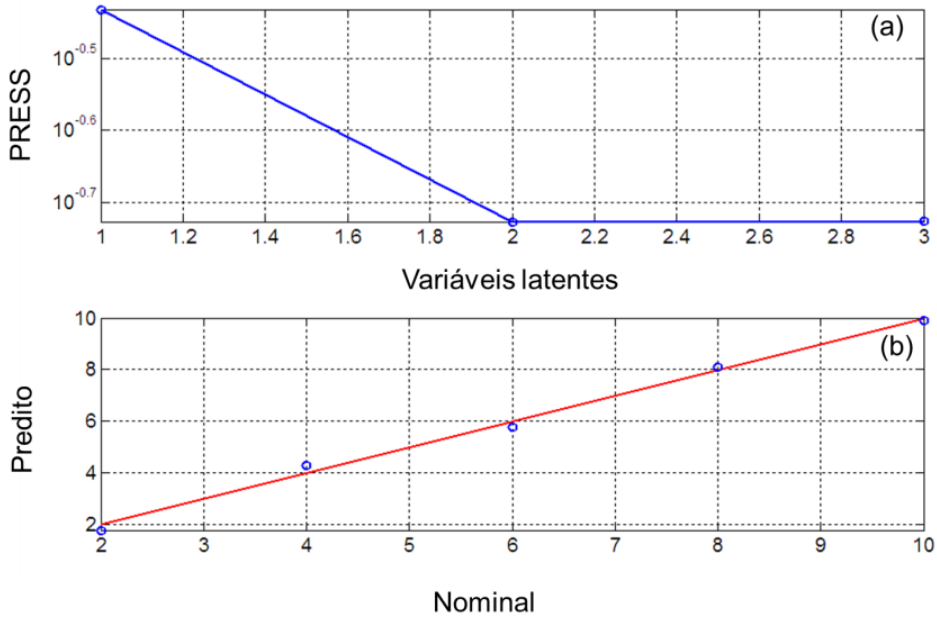


Figura 24: Resultados obtidos por validação cruzadas para as amostras de calibração em (a) a curva de PRESS versus o número de variáveis latentes e em (b) a curva ajusta entre valores nominais e preditos pelo modelo por validação cruzada.

Observando a **FIGURA 24a**, a variação de PRESS versus o número de variáveis latentes incluídas no modelo N-PLS parece sugerir que dois fatores são necessários para modelar os dados LC-DAD. Contudo, o valor de PRESS para um único fator já parece bastante razoável quando comparado aos níveis de concentração do conjunto de calibração. O comportamento do gráfico da **FIGURA 24a** certamente sugere um segundo fator para modelar a linha de base das amtraizes LC-DAD. Levando em consideração o fato de que as amostras de calibração contém um único constituinte e que o ajuste obtido para o modelo com uma única variável latente foi adequado (veja **FIGURA 24b**), não há razões para o uso de duas variáveis latentes. Portanto, apenas uma variável latente foi empregada na etapa de predição das amostra do conjunto de teste. Esta escolha é suportada pela compoisção quimica das amostras de calibração que são conhecidas.

O conjunto de teste, além da OFL, contém outras duas quinolonas não modeladas, a ciprofloxacina (CPF) e a danofloxacina (DNF). Os perfis cromatográficos e as estruturas químicas das quinolonas não modeladas na calibração são mostradas na **FIGURA 25a**. É possível observar que os tempos de retenção da CPF e da DNF são superiores ao tempo de retenção da OFL, contudo ocorre sobreposição (resolução inferior a um) entre as quinilonas não modelados e a ofloxacina. Os tempos de eluição da CPF e DNF são de 0,96 e 1,03 minutos respectivamente.

Observando a **FIGURA 25b**, vemos que os perfis espectrais são muitos parecidos, reflexo a similaridades entre as estruturas químicas das quinolonas envolvidas neste estudo. Na **FIGURA 25c** é possível ver o perfil LC-DAD das misturas das três quinolonas. Embora os máximos do perfil cromatografico de cada quinolona estejam

resolvidos ocorre sobreposição suficiente para afetar as predições empregando as abordagens univariada tradicionais e/ou calibração multivariada de primeira ordem, portanto se faz necessário alcançar vantagem de segunda ordem.

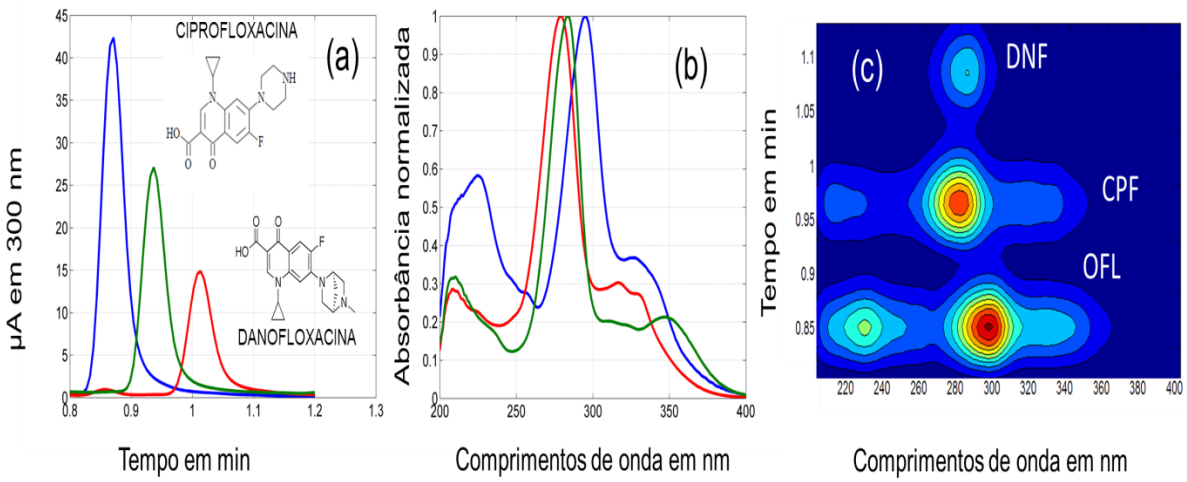


Figura 25: Conjunto de teste, em (a) os perfis cromatográficos e (b) espectrais puros das misturas de teste e em (c) uma típica superfície LC-DAD. Linha azul (OFL); linha verde (CPF) e a linha vermelha (DNF)

Os resultados na etapa de predição para o modelo N-PLS foram obtidos empregando dois fatores RBL, enquanto os modelos baseados em seleção de variáveis foi necessário apenas um único fator RBL para alcançar adequadamente a vantagem de segunda ordem. A variação de s_u típica para as amostras de teste em função do aumento de número de fatores RBL (*Ni*) é mostrada na **FIGURA 26**.

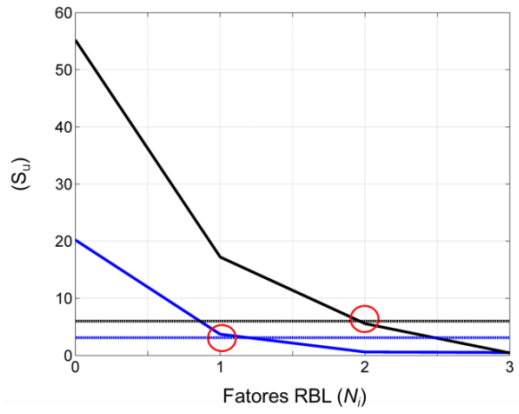


Figura 26: Variação de S*u* em função do aumento de N_i típico para as amostras de teste para o modelo N-PLS/RBL (linha preta) e *i*SPA-N-PLS/RBL (linha azul).

Observando a **FIGURA 26**, a primeira consideração a ser feita é que é o resíduo de calibração obtido pelo modelo *i*SPA-N-PLS foi inferior ao obtido pelo modelo N-PLS. Estes valores estão indicados pelas linhas horizontais pontilhadas. O típico resíduo obtido quando não se aplica a etapa RBL (N_i=0) na predição das amostras de teste para o modelo N-PLS/RBL é superior quando comparado ao modelo com seleção de variáveis. Este fato já sugere uma melhor ajuste do modelo de calibração para o modelo baseado em seleção de variáveis quando comparado ao modelo global.

Para que o resíduo da amostra de teste (s_u) alcance valores compatíveis com o resíduo típico da calibração (s_{cal}) foram necessários dois fatores RBL (N_i=2), quando se emprega o modelo N-PLS/RBL. Este resultado está em concordância com a composição química do sistema, uma vez que duas quinolonas não modeladas foram adicionadas as amostras de teste. Para o modelo *i*SPA-N-PLS/RBL apenas um fator foi necessário para que o s_u alcance valores concordantes com s_{cal}. Resultados similares ao *i*SPA-N-PLS/RBL foram obtidos para o GA-N-PLS/RBL.

Empregando uma única variável latente em todos os casos, dois fatores RBL para o modelo N-PLS/RBL e apenas um para os modelos baseados em seleção de variáveis os resultados obtidos estão resumidos na **TABELA 7**.

De acordo com os valores de RMSEP obtidos vemos que todos os modelos foram capazes de predizer a concentração de OFL nas amostras de teste. Como observado em todos os estudos de caso, o erro de predição apresentou melhorias quando se aplica seleção de variáveis, sendo esta vantagem mais pronunciada para o método proposto quando comparado ao GA-N-PLS.

Tabela 7: Resumo dos resultados da predição expressos em (mg L^{-1}).

Modelos	Métricas de desempenho				
	RMSEP	SEN	γ (L mg^{-1})	LOD	LOQ
N-PLS/RBL	0,72	556,74	7,58	0,05	0,13
GA-N-PLS/RBL	0,70	191,22	3,28	0,07	0,22
***i*SPA-N-PLS/RBL**	0,64	119,04	2,20	0,08	0,23

Com respeito à sensibilidade, vemos uma redução considerável quando se aplica seleção de variáveis. Este decrescimo em sensibilidade ocorre pelo uso de menos canais analiticos na construção do modelo N-PLS. Por outro lado, este decréscimo é compensado por um melhor ajuste dos dados pelo uso de canais mais seletivos, refletindo em valores de γ, LOD e LOQ comparáveis ao modelo N-PLS/RBL.

Os resultados obtidos sugerem que o processo de seleção de variáveis seleção de variáveis encontra o melhor compromisso entre sensibilidade e seletividade, ou seja, a sensibilidade é reduzida, mas não a níveis que afete o ajuste. Em contrapartida, a remoção de regiões não informativas e o uso dos apenas dos canais mais seletivos promove uma redução no erro médio de predição.

A acurácia dos modelos pode ser ainda corroborada pelas analise das respectivas EJCR mostradas na obtidas para os modelos N-PLS/RBL; GA-N-PLS/RBL e o método porposto **FIGURA 27**.

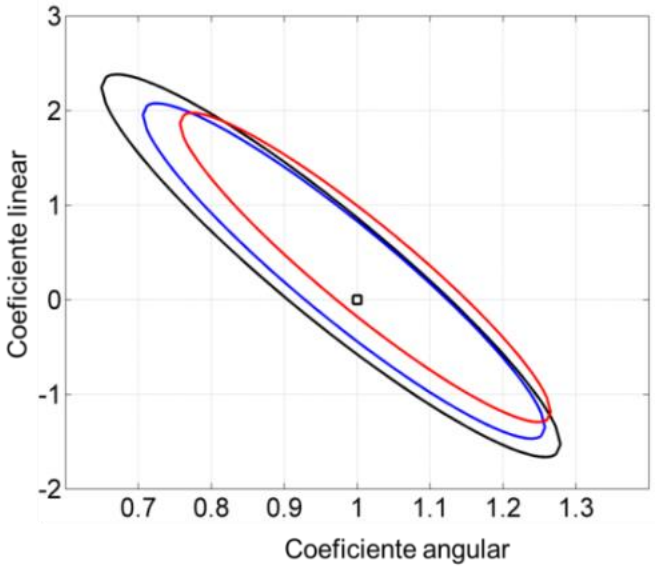


Figura 27: EJCR obtidas para os modelos (linha preta) N-PLS/RBL, (linha azul) *i*SPA-N-PLS/RBL e (linha vermelha) modelo GA-N-PLS/RBL.

Como mostrado acima, todas as EJCR contêm o ponto ideal, sugerindo que com 95% de confiança nenhum dos modelos apresentaram *bias* significativo. E por último, são mostradas as variáveis selecionados pelo *i*SAP-N-PLS/RBL e GA-N-PLS/RBL na **FIGURA 28**.

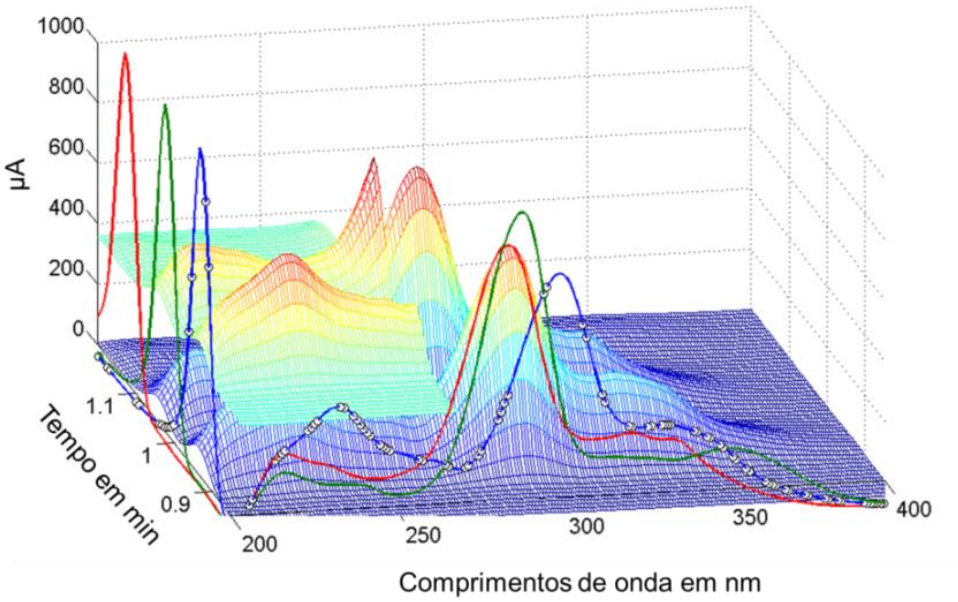


Figura 28: Perfil típico das misturas de teste e deslocado por offset o intervalo selecionado pelo *i*SPA São (o) sensores selecionados pelo GA. As linhas sólidas são os perfis puros para (-) OFL, (-) CPF e (-) DNF.

Enquanto o GA seleciona variáveis em ambos os modos instrumentais, o *i*SPA-N-PLS/RBL seleciona apenas um intervalo sobre modo espectral e todos as variáveis do modo cromatografico são incluidas no modelo. O intervalo selecionado pelo metodo proposto corresponde a região mais seletiva do sobre o perfil espectral da OFL. É possível perceber ainda que nesta faixa os espectros os constituintes não modelados CPF e DNF são praticamente iguais, isso explica por que o *i*SPA-N-PLS/RBL alcança a vantagem de segunda ordem com um único fator RBL, certamente como os perfis da CPF e DNF, são praticamente igual (extremamente correlacionado), quando se aplica seleção de variáveis são recuperados como um único perfil por combinação linear na etapa RBL.

5.2 *i*SPA-U-PLS/RBL

O algoritmo *i*SPA-U-PLS/RBL desenvolvido neste trabalho, foi aplicado em um contexto especifico, a modelagem de EEM na presença de filtro interno. O método proposto foi comparado ao PARAFAC e ao modelo U-PLS/RBL sem seleção de variáveis. O método proposto foi avaliado em dois estudos de caso. O primeiro envolveu dados simulados e no segundo foi conduzida a determinação de fenilefrina na presença de paracetamol em amostras de água.

5.2.1 Dados Simulados-II

Os dados simulados empregados neste estudo de caso envolvem um conjunto de calibração (18×31×31) contendo dois constituintes, o analito (A) e a espécie química que promove o efeito de filtro interno (F) sobre o sinal do analito. A faixa de calibração está compreendida entre 1 e 6 para o analito e concentrações randômicas entre 2 e 5 para a espécie F. O conjunto de teste (50×31×31) contém ainda um constituinte não modelado. Todas as espécies do conjunto de teste possuem concentrações randômicas entre 2 e 5 unidades. Na **FIGURA 29** é mostrado o típico sinal do analito na ausência do filtro interno (**FIGURA 29a**) e em presença do efeito de filtro interno (**FIGURA 29b**).

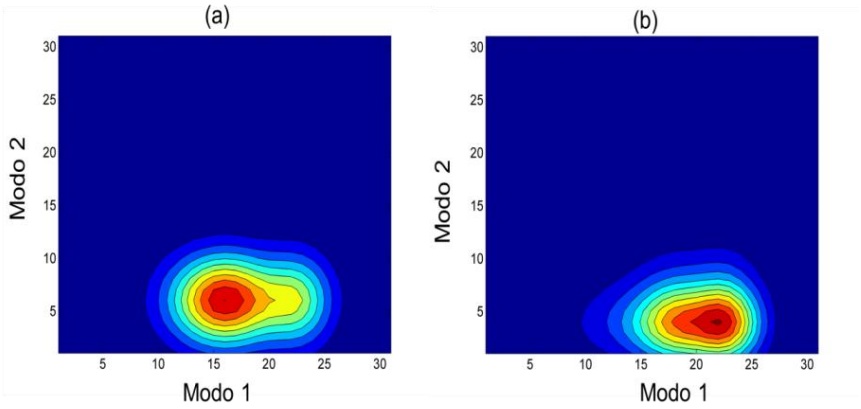


Figura 29: Sinal do analito puro em (a) e em (b) o sinal do analito puro sob efeito de filtro interno.

Na **FIGURA 29a** é mostrado o sinal do analito puro e em **FIGURA 29b** o sinal do analito puro quando sob efeito do filtro interno, a deformação do sinal em ambos os modos instrumentais é responsável pela quebra da trilinearidade dos dados em dois modos. Esta deformação causada pelo efeito de filtro interno inválida a maioria dos metodos de calibração em multivias.

Neste trabalho a abordagem utilizada para modelar o efeito de filtro interno foi a adição da espécie F nas amostras de calibração. Inicialmente o conjunto de calibração foi avaliado por validação cruzada para os modelos U-PLS e *i*SPA-U-PLS para escolha do número adequado de variáveis latentes. Para o modelo PARAFAC foi empregado, além o conhecimento do sistema, a variação de CORE em função do aumento do número de fatores. Os resultados obtidos nesta etapa para todos os modelos são mostrados na **FIGURA 30**.

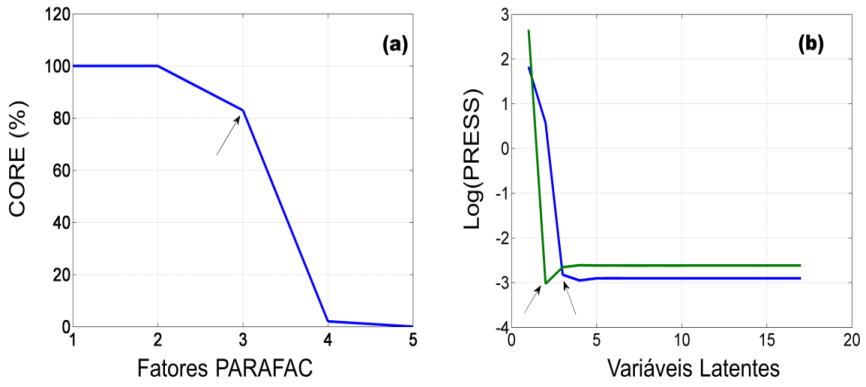


Figura 29: Resultados da escolha do número de fatores em (a) para o modelo PARAFAC e em (b) para modelos baseados em variáveis latentes.

O gráfico da **FIGURA 30a** para o modelo PARAFAC foi obtido empregando a decomposição dos dados de calibração concomitantemente com os dados do conjunto de teste. De acordo com a variação de CORE, em função do número de fatores PARAFAC empregado na decomposição dos dados. É possivel observar que para 4 ou mais fatores ocorre a completa perda da trilinearidade. Portanto, 3 representa, com cerca

de 82% de consistência de trilinearidade, o melhor número de fatores para o modelo PARAFAC. Este número de fatores está em concordância com a composição do sistema investigado. Como forma complementar de avaliar o desempenho da decomposição PARAFAC na **FIGURA 31** são mostrados os perfis recuperados relativos aos três fatores.

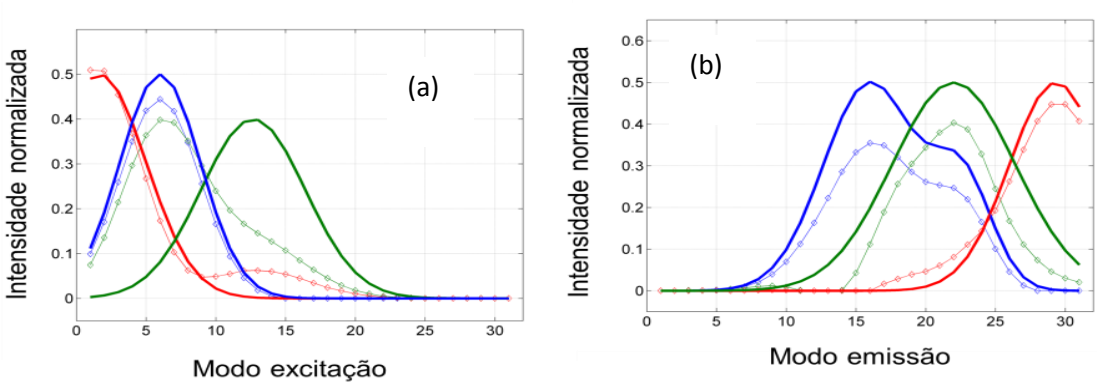


Figura 30: Perfil simulado puro e normalizado (linha solida) e perfil recuperado pelo PARAFAC (linhas com losangos) para o (—) analito, (—) espécie F (—) constituinte não modelado.

Com base na **FIGURA 31**, é possivel observar que o PARAFAC recupera perfis bastante deformados e com intensidades bastante diferentes dos perfis reais. A inconsistencia observada entre perfil real e recuperado pelo PARAFAC está asscíado a perda de trilinearidade dos dados EEM pormovido por variações que ocorrem de amostra para amostra devido o efeito de filtro intermo. O modo de emissão é o mais afetado pelo efeito de filtro interno, consequentemente sua recuperação, pelo PARAFAC, foi mais afetada quando comparado ao modo de excitação.

Esta deficiência do PARAFAC em resolver apropriadamente os perfis espectrais certamenta afetara a capacidade preditiva, e modelos com baixo desempenho serão obtidos.

Os perfis de log(PRESS) em função do número de variáveis latentes mostrados na **FIGURA 30b**, são obtidos empregando unicamente o conjunto de calibração. É possível observar que para o modelo U-PLS/RBL (linha azul) o mínimo da função de custo é obtido para três variáveis latentes. Este número de fatores está em concordância com a composição do sistema e suas peculiaridades. Dois fatores estão relacionados com as espécies existentes (Analito e espécie F) e a terceira variável latente está associada à modelagem do efeito de filtro interno. Ainda na **FIGURA 30b**, para o modelo *i*SPA-U-PLS, ao contrário do modelo U-PLS/RBL, apenas duas variáveis latentes foram sugeridas pela variação de PRESS em função do número de fatores (linha verde).

Para alcançar vantagem de segunda ordem quando se aplica os modelos U-PLS e *i*SPA-U-PLS é empregado a etapa pós calibração RBL. De forma similar aos estudos de casos já discutidos neste trabalho, foi inspecionada a variação de s_u com o aumento do número de fatores RBL, como mostrado na **FIGURA 32**.

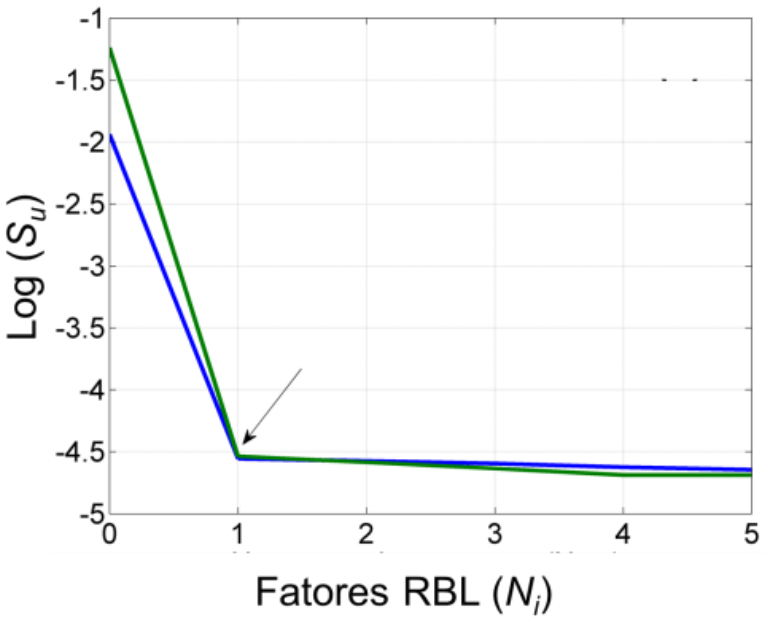


Figura 31: Variação do resíduo da amostra de teste em função da adição de fatores (N_i) para o modelo global (linha azul) e para o *i*SPA-U-PLS/RBL(linha verde).

Com base no gráfico da **FIGURA 32**, é observado que para alcançar a vantagem de segunda ordem, para ambos os modelos, um único fator RBL é necessário. Após o primeiro fator o resíduo típico das amostras de teste são estabilizados e não ocorrem variações significativas para fatores RBL adicionais. Empregando quatro fatores PARAFAC, três e duas variáveis latentes para os modelos U-PLS/RBL e *i*SPA-U-PLS/RBL respectivamente e ambos com um único fator RBL todas as amostras de teste foram preditas e os resultados são mostrados na **TABELA 8**.

Tabela 8: Resumo* da predição dados simulados (unidades arbitrárias).

Modelos	**Métricas de desempenho**				
	RMSEP	SEN	γ^{-1}	LOD	LOQ
PARAFAC	1,58	2,5	0,02	2,1	6,4
U-PLS/RBL	0,08	2,1	0,01	0,1	0,2
***i*SPA-U-PLS/RBL**	0,07	0,5	0,02	0,1	0,3

*Media do conjunto de teste

Observando a **TABELA 8,** com relação aos valores de RMSEP, como esperado o PARAFAC apresentou um ajuste inadequado, devido à quebra de trilinearidade causada pelo efeito de filtro interno. Por outro lado os modelos U-PLS/RBL e *i*SPA-U-PLS/RBL foram capazes de modelar adequadamente os dados simulados de EEM com filtro interno e alcançar com sucesso vantagem de segunda ordem. Em adição o modelo proposto (baseado em seleção de intervalos) foi capaz de promover melhorias em termos de acurácia (menor RMSEP) quando comparado ao modelo U-PLS/RBL. Do

ponto de vista da parcimônia, também ocorrerão olhorias, uma vez que menos variáveis latentes foram usadas para relacionar as EEM com a concentração do analito.

A acurácia dos modelos baseados no método PLS (U-PLS/RBL e *i*SPA-U-PLS/RBL) são corroboradas pelas respectivas EJCR (veja **FIGURA 33**), em que para ambos os casos o ponto ideal está contido na elipse. O mesmo não ocorre para o modelo PARAFAC.

Observando a posição de todas as EJCR com relação ao ponto ideal (1 e 0) vemos que as elipses apontam para um *bias* negativo, que é significante para o PARAFAC e insignificante para os demais modelos. Como o efeito de filtro interno atenua o sinal do analito uma tendncia *bias* negativo está em concordância com o esperado.

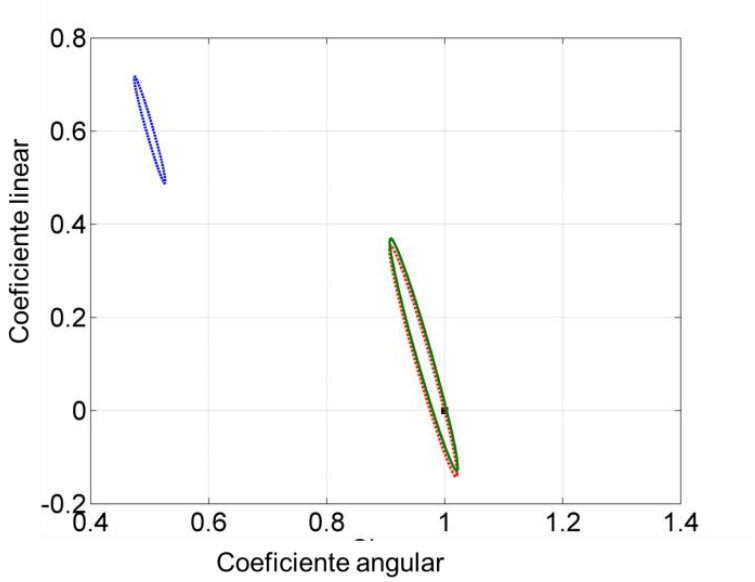


Figura 32: EJCR obtidas para os modelos (linha azul) PARAFAC, (linha vermelha) *U*-PLS/RBL e (linha verde) modelo *i*SPA-U-PLS/RBL.

O PARAFAC apresenta o maior valor de SEN, contudo a γ, o LOD e o LOQ são afetados negativamente pelo ajuste inadequado do modelo PARAFAC as EEM. O modelo *i*SPA-U-PLS/RBL apresenta queda na sensibilidade devido o uso de uma

quantidade de sensores reduzida contudo os valores de LOD e LOQ são comparáveis ao modelo U-PLS/RBL. O intervalo selecionado pelo *i*SPA-U-PLS/RBL é mostrado na **FIGURA 34**.

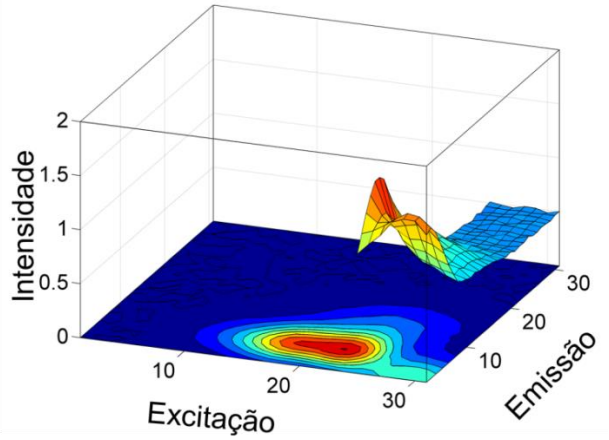


Figura 33: Superfície de contorno típica para as amostras do conjunto de teste e descolado por *offset* o intervalo selecionado pelo *i*SPA-U-PLS/RBL.

Observando a **FIGURA 34**, é possível visualizar que o *i*SPA-U-PLS/RBL seleciona um intervalo estreito ao longo do modo excitação (modo 1). Esta região da EEM é a menos afetada pelo efeito do filtro interno. O fato do *i*SPA-U-PLS/RBL empregar apenas a região menos deformada pelo efeito de filtro interno explica porque duas variáveis latentes foram suficiente para ajustar os dados de calibração, ao paso que o modelo U-PLS/RBL empregou 3 fatores.

5.2.2 Determinação de Fenilefrina em presença de paracetamol

O método proposto (*i*SPA-U-PLS/RBL) foi também avaliado em um estudo de caso envolvendo dados EEM para quantificação da fenilefrina (FEN) na presença de paracetamol (PAR). A FEN é um fármaco com muitas aplicações, dentre elas é usado como descongestionante nasal e agente cardiotônico. O PAR é conhecido como interferente na determinação de FEN, por ser capaz de absorver radiação nos

comprimentos de onda de excitação e emissão da FEN. A estrutura molecular da FEN e do PAR são mostrados na **FIGURA 35a**.

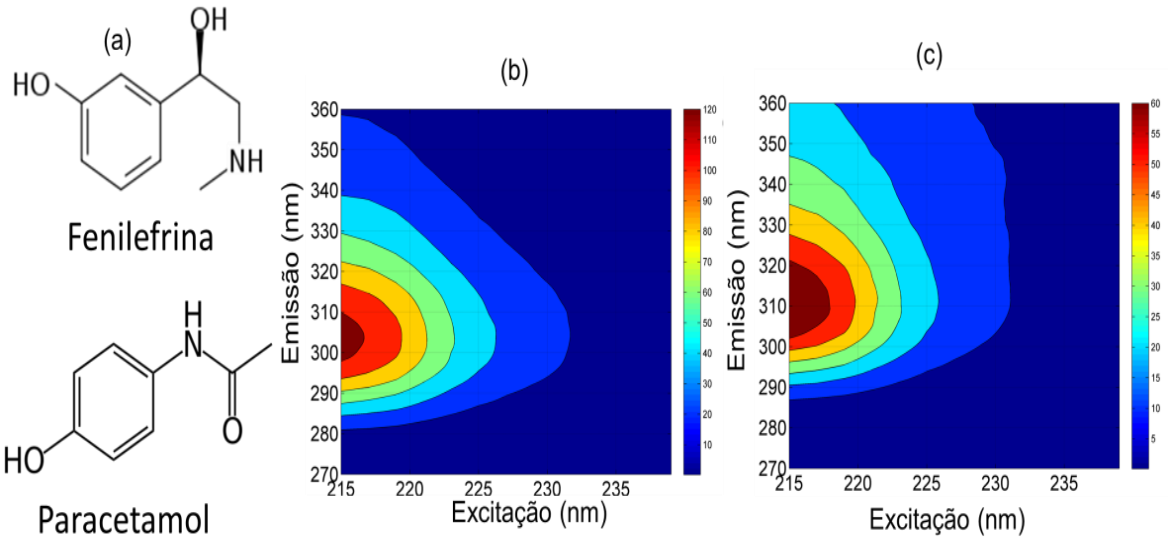


Figura 34: Em (a) estrutura molecular da FEN e do PAR, (b) suprfície de contorno da FEN pura e (c) na presença do PAR.

A semelhança entre as duas espécies (PAR e FEN) é bastante grande, pois praticamente os mesmos grupos fluoróforos estão presentes em ambos. Isso explica o efeito de filtro interno causado pelo PAR sobre o sinal de fluorescência da FEN. Na **FIGURA 35b** é mostrado o sinal de fluorescência da FEN pura na concentração de 0,50 µg/mL em água e na **FIGURA 35c** foi adicionado PAR de modo que sua concentração na amostra usada para registar o sinal da fluorescência fosse de 10,0 µg/mL.

Examinando a escala da barra de cores das **FIGURAS 35b** e **35c**, é possível ver a atenuação do sinal da fluorescência causada pela presença do PAR, bem como a distorção do sinal da FEN, similar ao observado para os dados simulados. Estas distorções no perfil espectral e na intensidade podem variar de amostra para amostra e promover a quebra da trilinearidade dos dados em dois modos.

Assim como nos dados simulados, neste estudo de caso o método proposto foi comparado ao modelo U-PLS/RBL e ao PARAFAC. Inicialmente foi investigado o número de fatores (fatores PARAFAC e variáveis latentes para os modelos PLS) necessários para ajustar os modelos aos dados de calibração.

Para o PARAFAC foi empregado à variação de CORE (**FIGURA 36**) em função do número de fatores. Para os modelos baseados em PLS foi usada a variação de PRESS em função do número de variáveis latentes.

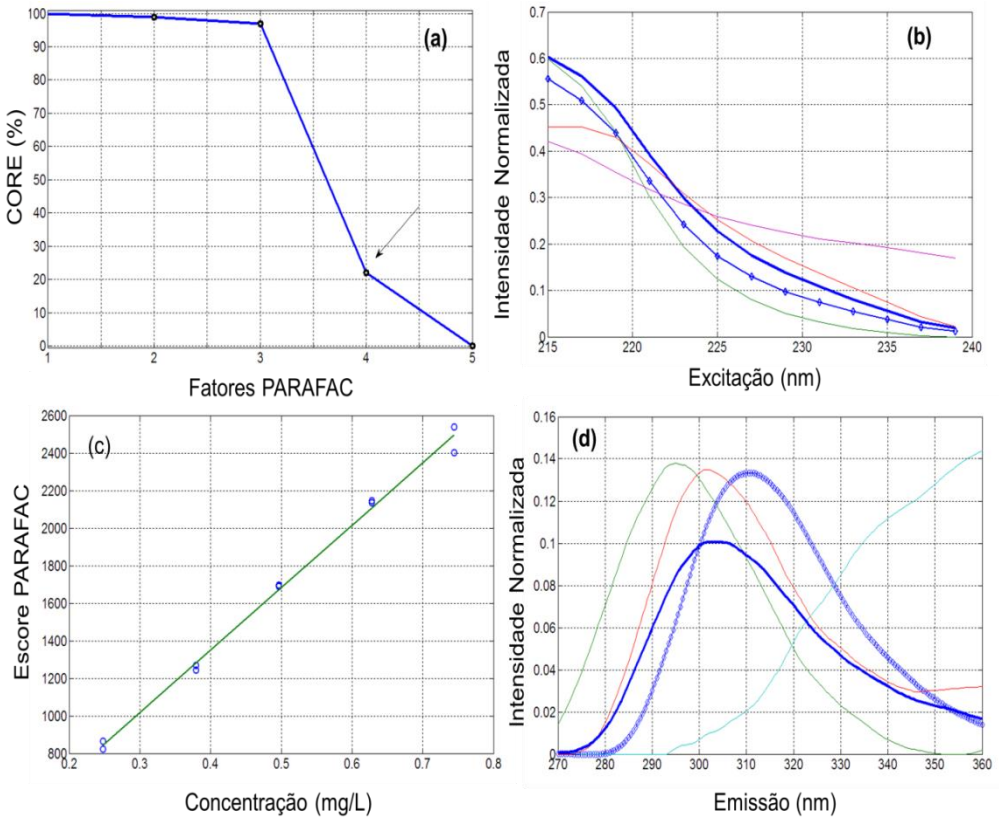


Figura 35: Resultados da modelagem PARAFAC, em (a) variação do CORE em função do número de fatores, em (b) os perfis recuperados pelo PARAFAC no modo excitação, em (c) a curva de calibração pseudo-univariada e em (d) no modo emissão. A linha azul solida é o perfil experimental da FEN e a linha azul-losango o perfil recuperado pelo PARAFAC. As demais linhas são os fatores dois (linha verde), três (linha vermelha) e quatro (linha ciano).

Observando a variação do CORE (**FIGURA 36a**) em função do número de fatores PARAFAC, é possivel ver que a trilinearidade é mantida apenas para três fatores. Com

quatro fatores o valor do CORE é de 22% apenas e para quantidades maiores de fatores o CORE se aproxima de zero. Este resultado mostra que os dados EEM perdem seu caráter trilinear devido o efeito de filtro interno. Na ausência de efeito de filtro interno, valores de CORE elevados deveriam ser obtidos para quatro fatores.

Na **FIGURA 36c** é mostrado o ajuste de calibração (curva pseudo-univariada do PARAFAC), e vemos que apesar da quebra de trilinearidade os escores PARAFAC e as concentrações nominais guardam uma relação linear. Como forma completar de avaliar o resultado do PARAFAC pode-se examinar os perfis recuperados (**FIGURAS 36b** e **36d**). Assim é possivel verificar que existe uma diferença entre perfil registrado experimentalmente (linha azul sólida) e o perfil recuperado para o analito (linha pontilhada-losango), essas diferenças são relativas a intensidade e posição dos máximos, o que mostra a deficiência do PARAFAC em lidar com dados não trilineares.

Com respeito aos modelos baseados em variáveis latentes, U-PLS (**FIGURA 37a**) e o *i*SPA-U-PLS (**FIGURA37b**), o gráfico da variação de PRESS em função do número de fatores PLS aponta para um mínimo bem localizado com dois fatores. Ao contrário dos dados simulados, neste estudo de caso a concentração de PAR, espécie que causa o efeito de filtro interno, foi mantida constante. Isso justifica o porquê dos dados simulados, quando modelados como U-PLS, requererem três fatores e neste caso apenas dois foram necessários.

Na **FIGURA 37** são apresentadas as curvas obtidas pelo ajuste entre valores preditos *versus* valores nominais para validação cruzada completa com dois fatores para os modelos U-PLS (**FIGURA 37c**) e *i*SPA-U-PLS (**FIGURA 37d**). Em ambos os casos é possível observar um ajuste satisfatório, com valores preditos muito proximos dos valores esperados.

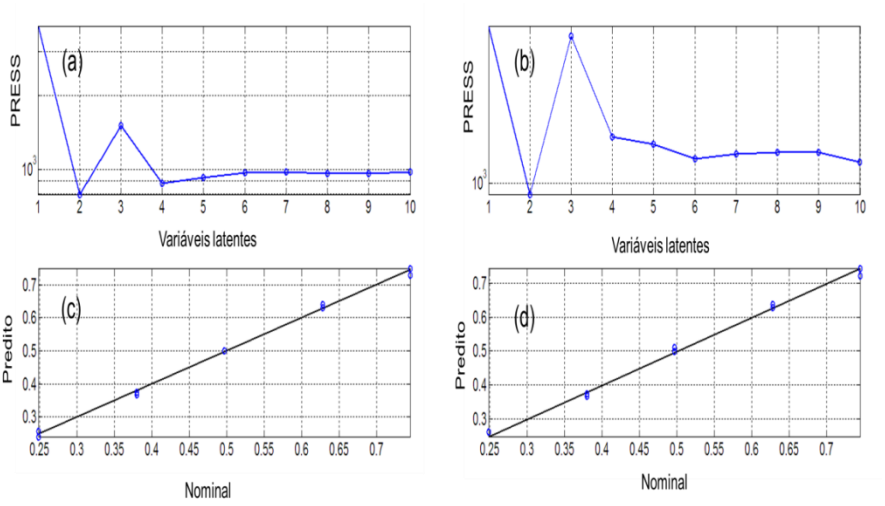


Figura 36: Resultado da validação cruzada dos modelos baseados em PLS, variação de PREES em função do número de variáveis latentes e valor nominal versus valor predito pelo modelo (a, c) para o modelo U-PLS e (b, d) para o modelo *i*SPA-U-PLS.

Após a seleção do número de fatores, como sendo quatro para os modelos PARAFAC e 2 para os modelos U-PLS e *i*SPA-UPLS, foram conduzidas a análises das amostras do conjunto de teste. As amostras de teste, além de FEN e PAR, contêm dois constituintes não modelados, o IBF e o ASS (veja as estruturas na **FIGURA 38a**).

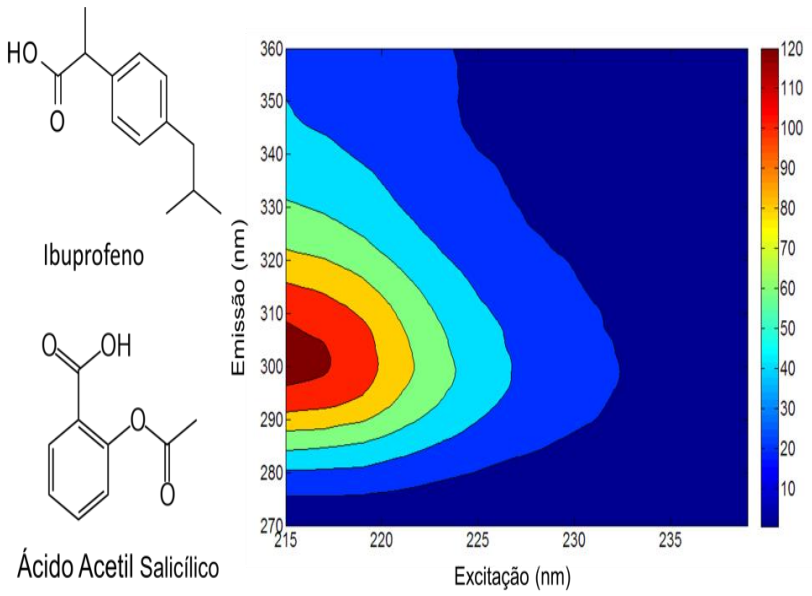


Figura 37: Detalhes do conjunto de amostras de teste em (a) é apresentado a formula estrutural dos constituintes não modelados e em (b) o típico sinal das amostras de teste na forma de superfície de contorno.

Na **FIGURA 38b** é aprersentado o típico perfil das amostras do conjunto de teste que, além da presença de filtro interno, apresentam constituintes não modelados se fazendo necessário alcançar vantagem de segunda ordem para se obter acurácia na fase de predição.

Ao contrário do PARAFAC, que possui vantagem de segunda ordem de forma intrínseca, o U-PSL precisa da etapa RBL para alcançar acuracia em presença de constituintes não modelados. A seleção do número de fatores RBL foi conduzida pela inspeção do gráfico do resíduo (s_u) em função do número de fatores RBL (N_i) como mostrado na **FIGURA 39**.

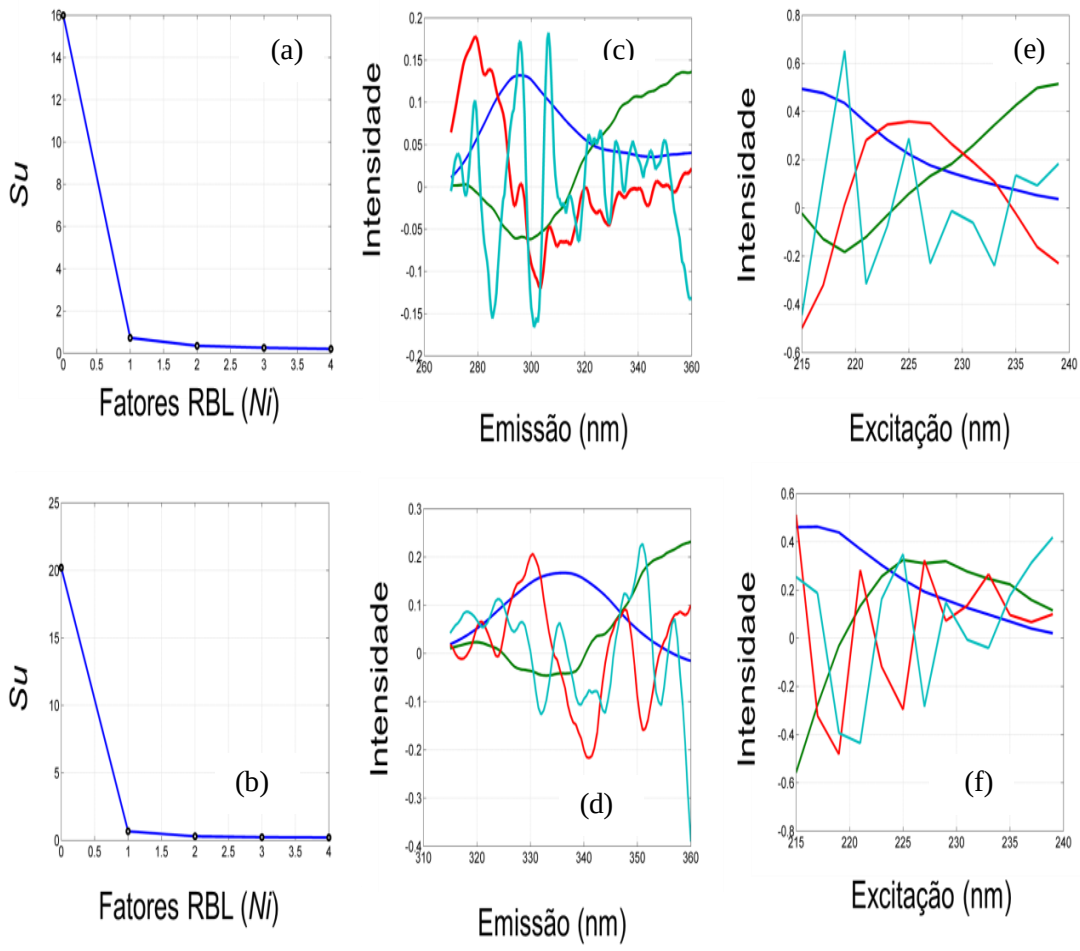


Figura 39: Resultados da etapa RBL em (a e b) típica variação do resíduo Su em função dos fatores Ni em (c e d) os perfis RBL recuperados no modo de emissão e em (e e f) no modo excitação para os modelos U-PLS/RBL e iSPA-U-PLS/RBL respectivamente. Fator 1 (linha azul), fator dois (linha verde), fator 3(linha vermelha) e fator 4 linha (ciano).

Nas **FIGURAS 39a** e **39b** são apresentadas típicas curvas da varição do resíduo das amostras de teste em função do aumento do número de fatores RBL. Para ambos os modelos (U-PLS/RBL e *i*SPA-U-PLS/RBL) dois fatores RBL são necessários para se obter a vantagem de segunda ordem. Após o segundo fator não ocorre mais variações pronunciadas no valor de s_u. Entretanto esta escolha não é tão óbivia como nos demais casos, e um recurso adicional foi empregado, a inspeção dos perfis recuperados ta etapa RBL.

Nas **FIGURAS 39c** e **39d** são mostrados os perfis dos interferentes recuperados na etapa RBL para os quatro primeiros fatores. Examinando estes perfis não é trivial perceber que os fatores três (linha vermelha) e quatro (linha ciano) não contribuem para modelagem, pois eles apresentam perfis suaves similar ao primeiro (linha azul) e segundo (linha verde) fator. Ao contrário do que se observa nos perfis do modo emissão, os perfis do modo excitação para o modelo U-PLS/RBL (**FIGURAS 39e**) são mais informativos com respeito à seleção do número de fatores RBL. O perfil associado ao quarto fator apresenta comportamento aleatório típico de ruído. Para o modelo *i*SPA-U-PLS/RBL (**FIGURAS 39f**) foi observado que o terceiro e o quarto fatores não possuem perfis compatíveis com informação química.

Portanto, a análise conjunta dos perfis recuperados e da variação de s_u em função de N_i justifica o uso de dois fatores RBL para alcançar a vantagem de segunda ordem, e não menos importante esse número é concordante com a composição das amostras de teste, que possui dois constituintes não modelados. Os resultados obtidos na predição das amostras de teste são mostrados na **TABELA 9**.

Tabela 9: Resumo da predição de FEN nas amostras de teste.

Modelos	Métricas de Desempenho				
	RMSEP ($\mu g\ mL^{-1}$)	SEN	γ^{-1} ($\mu g\ mL^{-1}$)	LOD ($\mu g\ mL^{-1}$)	LOQ ($\mu g\ mL^{-1}$)
PARAFAC	0,164	0,14	0,4	0,20	0,70
U-PLS/RBL	0,089	2,96	0,3	0,01	0,02
***i*SPA-U-PLS/RBL**	0,069	0,01	0,1	0,03	0,08

*Metricas de desempenho para PARAFAC calculadas de acordo com a referência [118].

Observando os valores de RMSEP da **TABELA 9** vemos um comportamento similar aos dados simulados, em que o PARAFAC não foi capaz de modelar adequadamente as EEM com efeito de filtro interno, enquanto o U-PLS/RBL obteve resultados aceitáveis o *i*SPA-U-PLS/RBL foi capaz de reduzir esse erro em aproximadamente 12%. Outra melhoria na acurácia promovida pela seleção de variáveis pode ser visualizada observando as EJCR na **FIGURA 40**.

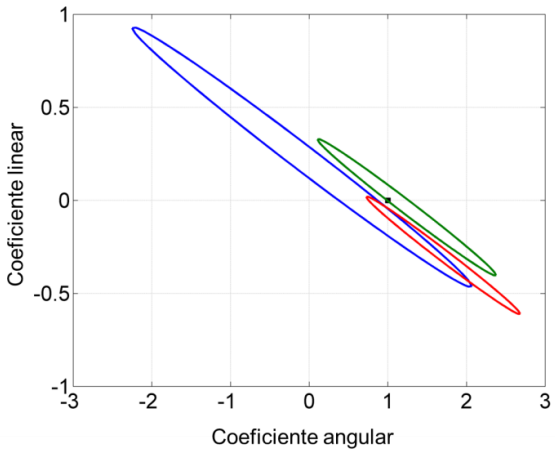


Figura 38: EJCR obtidas para os modelos (linha azul) PARAFAC, (linha vermelha) *U*-PLS/RBL e (linha verde) modelo *i*SPA-U-PLS/RBL.

Uma redução na área da região elíptica é alcançada quando se emprega o U-PLS/RBL e *i*SPA-U-PLS/RBL com relação ao PARAFAC. Este fato corrobora com a ideia de que a estrutura de dados desdobrados do U-PLS e modelagem em variáveis latentes seguida de bilinearização residual é uma ferramenta capaz de lidar um EEM afetadas por efeito de filtro interno. Em adição, o processo de seleção de variáveis é capaz de pormover melhorias nos resultdos obtidos pelo U-PLS/RBL. Observe que as elipses do PARAFAC e do U-PLS/RBL estão deslocadas para baixo do ponto ideal, sugerindo um *bias* negativo, em concordância com o filtro interno que atenua o sinal do analito. Entretanto, o único modelo para o qual o *bias* é não significativo (a 95% de confiança) é para o método proposto (*i*SPA-U-PLS/RBL).

Com relação às métricas de desempenho dos modelos baseados no cálculo da sensibilidade, é notável que a falta de acurácia do modelo PARAFAC afeta drasticamente os valores de SEN, γ, LOD e LOQ. É possível ver, como nos demais casos, que a seleção de variáveis reduz a SEN, contudo este decréscimo é compensado pelo melhor ajuste quando se emprega canais mais seletivos. Este fato pode ser visualizado no valor de γ, e consequentemente os valores de LOD e LOQ que não são afetados de forma considerável, quando se compara o "*custo beneficio*" de realizar seleção de variáveis. Este subconjuntos de canais mais seletivos selecionado pelo *i*SPA-U-PLS/RBL é mostrado na **FIGURA 41**.

Observando o intervalo selecionado pelo *i*SPA-U-PLS/RBL, vemos que no caso da modelagem de filtro interno, a melhoria obtida em termos de acurácia e parcimônia é alcançada pelo uso da faixa de sensores menos afetados pelo mesmo. Observe que a distorção do sinal de fluorescência é mais significativa para os sinais de emissão acima de 320 nm e excitação abaixo de 225 nm, esta faixa foi evitada pelo *i*SPA.

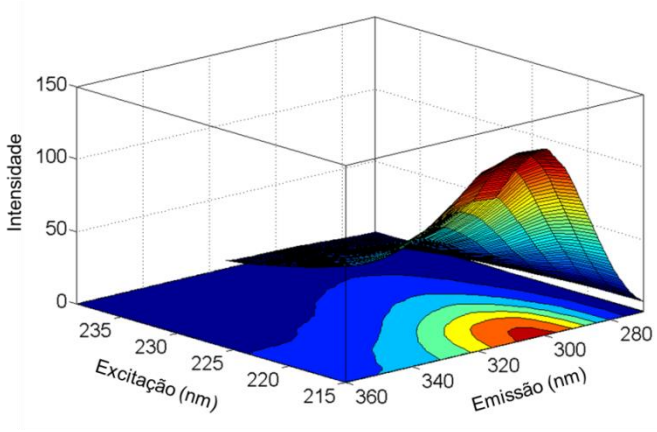


Figura 39: Superfície de contorno típica para as amostras do conjunto de teste e descolado por offset o intervalo selecionado pelo *i*SPA-U-PLS/RBL.

Capítulo VI

Conclusões

6. CONCLUSÕES

Neste trabalho foi desenvolvida uma nova estratégia de seleção de variáveis em calibração de ordem superior. A nova estratégia fez uso do Algoritmo das Projeções Sucessivas como ferramenta de seleção de intervalos combinado com regressão por Mínimos Quadrados Parciais em modo desdobrado e multidimensional ambos com etapa de bilinearização residual. Os métodos propostos foram avaliados em quatro estudos de casos divididos em duas partes.

Na primeira parte, o *i*SPA foi combinado com o N-PLS/RBL e avaliado em dois estudos de casos: dados simulados e na quantificação de ofloxacina em amostras de água. Os resultados obtidos foram comparados ao método N-PLS/RBL sem seleção de variáveis e ao GA-N-PLS/RBL.

O método proposto permitiu a construção de modelos empregando apenas um subconjunto de variáveis mais informativas e seletivas. Como resultado, foi observada melhor acurácia quando comparado ao método N-PLS/RBL sem seleção de intervalos. Foi observado também que para todos os casos, os modelos relativos ao método proposto foram isentos de *bias* significativos.

Em comparação com o GA-N-PLS/RBL, o método proposto mostrou melhor resultado na maioria dos casos, indicando que seleção de variáveis na forma de intervalos é a estratégia mais adequada para modelos PLS.

Na segunda parte, o *i*SPA foi combinado com o U-PLS/RBL e avaliado em dois estudos de casos: dados simulados e na quantificação de fenilefrina em amostras de água. Em ambos os estudos de casos foi empregados dados do tipo EEM com

ocorrência de filtro interno. Os resultados obtidos foram comparados ao método U-PLS/RBL sem seleção de variáveis e ao PARAFAC.

Em ambos os estudos de caso foi observado que o efeito de filtro interno nas EEM promove a quebra da trilinearidade e compromete o ajuste do modelo PARAFAC. Por outro lado, a estrutura flexível da modelagem em variáveis latentes do U-PLS/RBL mostrou-se capaz de modelar o efeito de filtro interno. Em adição o processo de seleção de variáveis mostrou ser capaz de melhorar o ajuste dos modelos U-PLS/RBL, bem como produzir modelos isentos de *bias* significativos.

Por conseguinte, o *i*SPA combinado com o N-PLS e U-PLS pode ser considerado uma nova estratégia válida para fazer seleção de intervalos em calibração de ordem superior.

6.1 Continuidade do Trabalho

Com propostas de continuidade deste trabalho as seguintes etapas poderão ser desenvolvidas:

- Determinações de espécies químicas em matrizes complexas.
- Generalização do *i*SPA para dados de altas ordens.
- Adaptação do *i*SPA em modelagem de segunda e alta ordem em calibração não linear (*i*SPA-Kernel/U-PLS/RBL).
- Avaliação do método proposto em aplicações envolvendo dados de outras técnicas analíticas.
- Adaptação do *i*SPA em de dados intrinsecamente não bilinear.
- Refinamento de interface gráfica do *i*SPA.

REFERÊNCIAS

[1] G. Lespes, J. Gigault. *Hyphenated analytical techniques for multidimensional characterisation of submicron particles: A review*. **Anal. Chim. Acta**, 692 (2011) 26–41.

[2] H. Parastar, J.R. Radovic, M. Jalali-Heravi, S. Diez, J.M. Bayona, R. Tauler. *Resolution and Quantification of Complex Mixtures of Polycyclic Aromatic Hydrocarbons in Heavy Fuel Oil Sample by Means of GC GC-TOFMS Combined to Multivariate Curve Resolution*. **Anal. Chem**. 83 (2011) 9289–9297.

[3] I. Lavilla, V. Romero, I. Costas, C. Bendicho. *Greener derivatization in analytical chemistry*. **Trends Anal. Chem**. 61 (2014) 1–10.

[4] N. Kumar, A. Bansal, G.S. Sarma, R. K. Rawal. *Chemometrics tools used in analytical chemistry: An overview*. **Talanta** 123 (2014) 186 – 19 9.

[5] D. D. S. Fernandes, A. A. Gomes, G. B. Costa, G.W. B. Silva, G. Véras. *Determination of biodiesel content in biodiesel/diesel blends using NIR and visible spectroscopy with variable selection*. **Talanta** 87 (2011) 30– 34.

[6] J. Burger, P. Geladi. *Hyperspectral NIR imaging for calibration and prediction: a comparison between image and spectrometer data for studying organic and biological samples*. **Analyst** 131 (2006) 1152–1160.

[7] D.S. Ferreira, O.F. Galão, J.A.L. Pallone, R.J. Poppi. *Comparison and application of near-infrared (NIR) and mid-infrared (MIR) spctroscopy for determination of quality parameters in soybean samples*. **Food Control** 35 (2014) 227–232.

[8] G. M. Escandar, P. C. Damiani, H. C. Goicoechea, A. C. Olivieri. *A review of multivariate calibration methods applied to biomedical analysis*. **Microchem. J.** 82 (2006) 29 – 42.

[9] H. Wu, Y. Li, R. Yu, Hai-Long Wu, Yong Li and Ru-Qin Yu. *Recent developments of chemical multiway calibration methodologies with second-order or higher-order advantages*. **J. Chemom.** 2014 (28) 476–489.

[10] J. A. Arancibia, P. C. Damiani, G. M. Escandar, G. A. Ibañez, A. C. Olivieri. *A review on second and third-order multivariate calibration applied to chromatographic data*. **J. Chromat. B**, 910 (2012) 22–30

[11] H. Parastar, R. Tauler. *Multivariate Curve Resolution of Hyphenated and Multidimensional Chromatographic Measurements: A New Insight to Address Current Chromatographic Challenges*. **Anal. Chem**. 19 (2013) 286-297.

[12] M.C. Hurtado-Sánchez, I. Durán-Merás, M.I. Rodríguez-Cáceres, A. Jiménez-Girón, A.C. Olivieri. *Comparison of the predictive ability of several second-order*

multivariate methods in the simultaneous determination of two therapeutic drugs in human urine. **Talanta** 88 (2012) 609–616.

[13] R. L. Pérez, G. M. Escandar. *Liquid chromatography with diode array detection and multivariate curve resolution for the selective and sensitive quantification of estrogens in natural Waters*. **Anal. Chim. Acta** 835 (2014) 19–28.

[14] M. P. Godoy-Caballero, M. J. Culzoni, T. Galeano-Díaz, M. I. Acedo-Valenzuela. *Novel combination of non-aqueous capillary electrophoresis and multivariate curve resolution-alternating least squares to determine phenolic acids in virgin olive oil*. **Anal. Chim. Acta** 763 (2013) 11–19.

[15] A.D. Walmsley. *Improved variable selection procedure for multivariate linear regression.* **Anal. Chim. Acta** 354 (1997) 225–232.

[16] M. B. Seasholtz, B. Kowalski. *The parsimony principle applied to multivariate calibration*. **Anal. Chim. Acta** 277 (1993) 165-177.

[17] D. A. E Skoog J. J. Leary. *Principles of instrumental analysis*. 6. ed. New York : Saunders College Publishing, 1992.

[18] G. M. Escandar, A. C. Olivieri. *Pratical Three-way Calibration*. **Elsevier** 2014.

[19] B. E. Wilson, B.R. Kowalski. *Quantitative Analysis in the Presence of Spectral Interferents Using Second-Order Nonbilinear Data*. **Anal. Chem**. 61 (1989) 2277-2284.

[20] A. Smilde, R. Bro, P. Geladi. *Multi-way Analysis: Applications in the Chemical Sciences*. **Wiley**, 2004.

[21] A. C. Olivieri. *Recent advances in analytical calibration with multi-way data*. **Anal. Methd**. 4 (2012) 1876–1886.

[22] R. G. Brereton, *Introduction to multivariate calibration in analytical chemistry*. **Analyst** 125 (2000) 2125–2154.

[23] B. Barros Neto, M. F. Pimentel, M. C. U. Araújo. *Recomendações para calibração em química analítica- Parte I: Fundamentos e calibração com um componente (calibração univariada)*. **Quim. Nova**. 25 (2002) 856–865.

[24] B. Barros Neto, M. F. Pimentel. *Calibração: Uma revisão para químicos analíticos*. **Quim. Nova**. 19 (1996) 268–277.

[25] M. F. Pimentel, R. K. H. Galvão. *Recomendações para calibração em química analítica- Parte II: Calibração Multianalito*. **Quim. Nova**.31 (2008) 462-467.

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

[27] T. C. B. Saldanha, B. Barros Neto, M. C. U. Araújo. *Análise multicomponente simultânea por espectrometria de absorção molecular UV-Vis*. **Quim. Nov**a 22 (1999) 847–853.

[28] S. Weisberg. *Applied linear regression*. **Wiley** (series in probability and statistics 3 ed.) 1947.

[29] M. Forina, S. Lanteri, M. C. C. Oliveros, C. P. Millan. *Selection of useful predictors in multivariate calibration*. **Anal Bioanal Chem** 380 (2004) 397–418.

[30] S. Wold, M. Høy, H. M.artens, J. Trygg, F. Westad, J. MacGregor, B. M. Wise. *The PLS model space revisited*. **J. Chemom**. 23 (2009) 67–68.

[31] R. B. Keithley, M. L. Heien, R. M. *Wightman, Multivariate concentration determination using principal component regression with residual analysis*. **Trend. Anal. Chem**. 28 (2009) 1128–1136.

[32] M. M. C. Ferreira, C. A. Montanari, A. C. Guadio. *Seleção de variáveis em QSAR*. **Quim. Nov**a 25 (2002) 439–448.

[33] F. Allegrini, A. C. Olivieri. *Analytical Figures of Merit for Partial Least-Squares Coupled to Residual Multilinearization*. **Anal. Chem.** 84 (2010) 10823 – 10830.

[34] R. Bro. *PARAFAC. Tutorial and applications*. **Chemom. Intell. Lab. Systm**. 38 (1997) 149-171.

[35] A. de Juan, S. C. Rutan, R. Tauler. *Two-Way Data Analysis: Multivariate Curve Resolution –Iterative Resolution Methods* in. S. D. Brown, R. Tauler, .B. Walczak. Comprehensive Chemometrics Chemical and Biochemical Data Analysis. **Elsevier** 2 (2009) 325–344.

[36] A. C. Olivieri. *On a versatile second-order multivariate calibration method based on partial least-squares and residual bilinearization: Second-order advantage and precision properties*. **J. Chemom**. 19 (2005) 253–265.

[37] R. Bro. *Multiway calibration: Multilinear PLS*. **J. Chemom**. 10 (1996) 47–61.

[38] M. M. Sena, M. G. Trevisan, R. J. Poppi. *PARAFAC: Uma ferramenta quimiometrica para tratamento de dados multidimensionais. Aplicação na determinação direta de fármacos em plasma*. **Quim. Nov**. 28 (2005) 910–920.

[39] C. M. Andersen, R. Bro. *Pratical aspects of PARAFAC modeling of fluorescence excitation-emission data*. **J. Chemom**. 17 (2003) 200–215.

[40] R. Bro. *Multi-way Analysis in the Food Industry: Models, Algorithms, and Applications*. **Tese de doutorado**. Denemarken, 1998 (disponível em http://www.models.life.ku.dk/go?filename=Thesis_Rasmus_Bro.pdf).

[41] K. R. Murphy, C. A. Stedmon, D. Graeber, R. Bro. *Fluorescence spectroscopy and multi-way techniques.PARAFAC*. **Anal. Methd**. 5 (2013) 6557 – 6566.

[42] M. Kompany-Zareh, Y. Akhlaghi, R. Bro. *Tucker core consistency for validation of restricted Tucker3 models*. **Anal. Chim. Acta** 723 (2012) 18–26.

[43] R. Bro, H. A. L. Kiers. *A new efficient method for determining the number of components in PARAFAC models*. **J. Chemom.** 17 (2003) 274–286.

[44] R. Bro, N. Viereck, M. Toft, H. Toft, P. I. Hansen, S. B. Engelsen. *Mathematical chromatography solves the cocktail party effect in mixtures using 2D spectra and PARAFAC*. **Trends Anal. Chem**. 29 (2010) 281–284.

[45] M. Bahram, R, Bro. *A novel strategy for solving matrix effect in three-way data using parallel profiles with linear dependencies*. **Anal. Chim. Acta** 584 (2007) 397–402.

[46] S. Wold, M. Sjöström, L. Eriksson. *PLS-regression: a basic tool of chemometrics*. **Chemom. Intell. Lab. Syst**. 58 2001 109–130.

[47] M. Andersson. *A comparison of nine PLS1 algorithms*. **J. Chemom**. 23 (2009) 518–529.

[48] R. Ergon. *Re-interpretation of NIPALS results solves PLSR in consistency problem.* **J. Chemom.** 23 (2009) 72–75.

[49] A. M. K. Pedro, M. M. C. Ferreira. *Simultaneously calibrating solids, sugars and acidity of tomato products using PLS2 and NIR spectroscopy*. **Anal. Chim. Acta** 595 (2007) 221–227.

[50] T. Mehmood, K. H. Liland, L. Snipen, S. Sæbø. *A review of variable selection methods in Partial Least Squares Regression.* **Chemom. Intell. Lab. Syst**. 118 (2012) 62 – 69.

[51] S. A. Bartolato, J. A. Arancibia, G. M. Escandar, A. C. Olivieri. *Improvement of residual bilinearization by particle swarm optimization for achieving the second-order advantage with unfolded partial least-squares*. **J. Chemom**. 20 (2007) 1–10.

[52] K. Baumann. *Cross-validation as the objective function for variable-selection techniques*. **Trends Anal. Chem.** 22 (2003) 395–406.

[53] Q. Xu, Y. Lian, Y. Du. *Monte Carlo cross-validation for selecting a model and estimating the predict ion error in multivariate calibration*. **J. Chemom**. 18 (2004) 112–120.

[54] M. J. Culzoni, H. C. Goicoechea, A. P. Pagani, M. A. Cabezón, A. C. Olivieri. *Evaluation of partial least-squares with second-order advantage for the multi-*

way spectroscopic analysis of complex biological samples in the presence of analyte–background interactions. **Analyst** 131 (2006) 718–723.

[55] K. Calimag-Williams, G. Knobel, H.C. Goicoechea, A.D. Campiglia. *Achieving second order advantage with multi-way partial least squares and residual bi-linearization with total synchronous fluorescence data of monohydroxy–polycyclic aromatic hydrocarbons in urine samples*. **Anal. Chim. Acta** 811 (2014) 60–69.

[56] S. A. Bortolato, J. A. Arancibia, G. M. Escandar. *Chemometrics-Assisted Excitation–Emission Fluorescence Spectroscopy on Nylon Membranes.Simultaneous Determination of Benzo[a]pyrene and Dibenz\[a,h]anthracene at Parts-Per-Trillion Levels in the Presence of the Remaining EPA PAH Priority Pollutants As Interferences*. **Anal. Chem**. 80 (2008) 8276–8286.

[57] F. Alarcón, M. E. Báez, M. Bravo, P. Richter, G. M. Escandar, A. C. Olivieri, E. Fuentes. *Feasibility of the determination of polycyclic aromatic hydrocarbons in edible oils via unfolded partial least-squares/residual bilinearization and parallel factor analysis of fluorescence excitation emission matrices*. **Talanta** 103 (2013) 361–370.

[58] G. N. Piccirilli, G. M. *Partial least-squares with residual bilinearization for the spectrofluorimetric determination of pesticides. A solution of the problems of inner-filter effects and matrix interferents*. **Analyst** 131 (2006) 1012–1020.

[59] D. B. Gil, A. M. Peña. J. A. Arancibia, G. M. Escandar, A. C. Olivieri. *Second-Order Advantage Achieved by Unfolded-Partial Least-Squares/Residual Bilinearization Modeling of Excitation-Emission Fluorescence Data Presenting Inner Filter Effects*. **Anal. Chem**. 78 (2006) 8051–8058.

[60] G. E. Rovati. *A versatile implementation of the Gauss-Newton minirrdzation algorithm using MATLAB for Macintosh microcomputers*. **Computer Methods and Programs in Biomedicine**, 32 (1990) 161–167.

[61] A. C. Silva. *Um modelo de calibração de segunda ordem para determinação espectrofluorimétrica de hidrocarbonetos policíclicos aromáticos em bebidas destiladas*. **Dissertação de Mestrado**. João Pessoa-PB. (Disponível em: http://www.quimica.ufpb.br/posgrad/dissertacoes.html).

[62] Z. X. Wang, Q. P. He, J. Wang. *Comparison of variable selection methods for PLS-based soft sensor modeling*. **J. Process Control** 26 (2015) 56–72.

[63] Y. Yun, Y. Liang, G. Xie, H. Li, D. Cao, S. Xu. *A perspective demonstration on the importance of variable selection in inverse calibration for complex analytical systems*. **Analyst** 138 (2013) 6412 – 6421.

[64] B. K. Alsberg, D. B. Kell, R. Goodacre. *Variable Selection in Discriminant Partial Least-Squares Analysis*. **Anal. Chem**. 70 (1998) 4126-4133.

[65] M. N. Martins, R. K. H. Galvão, M. F. Pimentel. *Multivariate Calibration Transfer Employing Variable Selection and Subagging*. J. Braz. Chem. Soc. 21(2010) 127-134.

[66] M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad, M. Akhond. *Ant colony optimisation: a powerful tool for wavelength selection*. **J. Chemom**. 20 (2006) 146–157.

[67] A. Rinnan, M. Andersson, C. Ridder, S. B. Engelsen. *Recursive weighted partial least squares (rPLS): an efficient variable selection method using PLS*. **J. Chemom**. 28 (2014) 439–447.

[68] A.D. Walmsley. *Improved variable selection procedure for multivariate linear regression*. **Anal. Chim. Acta** 354 (1997) 225–232

[69] J. Gauchi, P. Chagnon. *Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data*. **Chemom. Intell. Lab. Syst.** 58 (2001) 171–193

[70] R. Leardi, A. L. Gonzalez. *Genetic algorithms applied to feature selection in PLS regression: how and when to use them*. **Chemom. Intell. Lab. Syst**. 41 (1998) 195–207.

[71] J. Ghasemi, A. Niazi, R. Leardi. *Genetic-algorithm-based wavelength selection in multicomponent spectrophotometric determination by PLS: application on copper and zinc mixture*. **Talanta** 59 (2003) 311–317.

[72] M. Goodarzi, M. P. Freitas, R. Jensen. *Ant colony optimization as a feature Chemometrics and Intelligent Laboratory Systems 98 (20 09) 123 –12 9selection method in the QSAR modeling of anti-HIV-1 activities of 3-(3,5-dimethylbenzyl) uracil derivatives using MLR, PLS and SVM regressions*. **Chemom. Intell. Lab. Syst**. 98 (20 09) 123 –129.

[73] J. A. Hageman, M. Streppel, R. Wehrens, L . M. C. Buydens. *Wavelength selection with Tabu Search*. **J. Chemom**. 17 (2003) 427–437.

[74] 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*. **Chemom. Intell. Lab. Syst**. 57 (2001) 65–73

[75] H. Martens, M. Martens. *Modied Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR)*. **Food Qual. Prefer.** 11 (2000) 5–16.

[76] M. Forina, C. Casolino, C. P. Millan. *Iterative predictor weighting (IPW): A technique for the elimination of useless predictor in regression problem*. **J. Chemom**. 13 (1999) 165–184.

[77] F. Allegrini, A. C. Olivieri. *An integrated approach to the simultaneous selection of variables, mathematical pre-processing and calibration samples in partial least square*. **Talanta** 115 (2013) 755 –760.

[78] S. Sæbø, T. Almøy, J. Aarøe, A.H. Aastveit. *St-pls: a multi-dimensional nearest shrunken centroid type classifier via pls*. **J. Chemom**. 20 (2007) 54–62.

[79] G. G. Andersson, B. K. Dable, K. S. Booksh. *Weighted parallel factor analysis for calibration of HPLC-UVrVis spectrometers in the presence of Beer's law deviations.* **Chemom. Intell. Lab. Syst.** 49 (1999) 195–213.

[80] W. Wu, Q. Guo, D.L. Massart, C. Boucon, S. Jong. *Structure preserving feature selection in PARAFAC using a genetic algorithm and Procrustes analysis*. **Chemom. Intell. Lab. Syst**. 65 (2003) 83 – 95

[81] L. Stordrange, T. Rajalahti, F. O. Libnau. *Multiway methods to explore and model NIR data from a batch process*. **Chemom. Intell. Lab. Syst**. 70 (2004) 137 – 145.

[82] S. Gourvénec, X. Capron, D.L. Massart. *Genetic algorithms (GA) applied to the orthogonal projection approach (OPA) for variable selection*. **Anal. Chim. Acta** 519 (2004) 11–21.

[83] Y. Chu, Y. Lee, C. Han. *Improved Quality Estimation and Knowledge Extraction in a Batch Process by Bootstrapping-Based Generalized Variable Selection*. **Ind. Eng. Chem. Res**. 43 (2004) 2680–2690.

[84] K. Choi, Y. Zeng, J. Qin. *Using sequential Floating Forward Selection Algorithm to detect epileptic seizure in EEG signals*. **IEEE Proceedings** 978 (2012) 1637–1640.

[85] R. L. Carneiro, J. W.B. Braga, C. B.G. Bottoli, R. J. Poppi. *Application of genetic algorithm for selection of variables for the BLLS method applied to determination of pesticides and metabolites in wine*. **Anal. Chim. Acta** 595 (2007) 51–58.

[86] A. Conesa, J. M. Prats-Montalbán, S. Tarazona, M. J. Nueda, A. Ferrer. *A multiway approach to data integration in systems biology based on Tucker3 and N –PLS*. **Chemom. Intell. Lab. Syst.** 104 (2010) 101 –111.

[87] L. W. Hantao, B. R. Toledo, F. A. L. Ribeiro, M. Pizetta, C. G. Pierozzi, E. L. Furtado. F. Augusto. *Comprehensive two-dimensional gas chromatography combined to multivariate data analysis for detection of disease-resistant clones of Eucalyptus*. **Talanta** 116 (2013) 1079–1084.

[88] H. Zhu, W. Guo, Z. Shen, Q. Tang, W. Ji, L. Jia. *QSAR models for degradation of organic pollutants in ozonation process under acidic condition*. **Chemosphere** 119 (2015) 65–71.

[89] S. F. C. Soares, A. A. Gomes, A. R. Galvão Filho, M. C. U. Araújo, R. K. H. Galvão. *The successive projections algorithm*. **Trnds Anal. Chem**. 42 (2013) 84–98.

[90] R. K. H. Galvão, M. C. U. Araújo, W. D. Fragoso, E. C. Silva, G. E. José, S. F. C. Soares, H. M. Paiva. *A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm*. **Chemom. Intell. Lab. Syst**. 92 (2008) 83–91.

[91] X. Chen, H. Li, D. Wu, X. Lei, X. Zhu, A. Zhang. *Application of a hybrid variable selection method for the classification of rapeseed oils based on ^{1}H NMR spectral analysis*. **Eur Food Res. Technol**. 230 (2010) 81–988.

[92] A. S. Soares, A. R. Galvão Filho, R. K. H. Galvão, M. C. U. Araújo. *Improving the Computational Efficiency of the Successive Projections Algorithm by using a Sequential Regression Implementation: A Case Study Involving NIR Spectrometric Analysis of Wheat Samples*. **J. Braz. Chem. Soc.** 21 (2010) 760–763.

[93] R. K. H. Galvão, M. C. U. Araújo, E. D. Silva, G. E. José, S. F. C. Soares, H. M. Paiva. *Cross-Validation for the Selection of Spectral Variables Using the Successive Projections Algorithm*. **J. Braz. Chem.** 18 (2007) 1580–1584.

[94] M. J. C. Pontes, R. K. H. Galvão, M. C. U. Araújo, P. N. T. Moreira, O. D. pessoa Neto, G. M. José, T. C. B. Saldanha. *The successive projections algorithm for spectral variable selection in classification problems*. **Chemom. Intell. Lab. Syst**. 78 (2005) 11–18.

[95] F. A. Honorato, B. Barros Neto, M. N. Martins, R. K. H. Galvão, M. F. Pimentel. *Transferência de Calibração em métodos multivariados*. **Quim. Nov.** 30 (2007) 1301-1312.

[96] H. A. Dantas Filho, R. K. H. Galvão, M. C. U. Araújo, E. D. 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*. **Chemom. Intell. Lab. Syst.** 72 (2004) 83–91.

[97] S. F. C. Soares, R. K. H. Galvão, M. C. U. Araújo, E. D. Silva, C. F. Pereira, S. I. E. Andrade, F. L. carvalho. *A modification of the successive projections algorithm for spectral variable selection in the presence of unknown interferents*. Anal. Chim. Acta 689 (2011) 22–28.

[98] S. Ye, D. Wang, S. Min. *Successive projections algorithm combined with uninformative variable elimination for spectral variable selection*. **Chemom. Intell. Lab. Syst**. 91 (2008) 194– 199

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

[100] J. Li, C. Zhao, W. Huang, C. Zhang, Y. Peng. *A combination algorithm for variable selection to determine soluble solid content and firmness of pears*. **Anal. Method.** 6 (2014) 2170–2180.

[101] M. Kompany-Zareh, Y. Akhlaghi. *Correlation weighted successive projections algorithm as a novel method for variable selection in QSAR studies: investigation of anti-HIV activity of HEPT derivatives*. **J. Chemom**. 21 (2007) 239–250.

[102] M. C. Breitkreitz, I. M. Raimundo Jr., J. J. R. Rohwedder, C. Pasquini, H. A. Dantas Filho, G. E. José, M. C. U. Araújo. *Determination of total sulfur in diesel fuel employing NIR spectroscopy and multivariate calibration*. **Analyst** 128 (2003) 1204–1207.

[103] F. A. C. Sanches, R. B. Abreu, M. J. C. Pontes, F. L. Carvalho, D. J. E. Costa, R. K. H. Galvão, M. C. U. Araújo. *Near-infrared spectrometric determination of dipyrone in closed ampoules*. **Talanta** 92 (2012) 84–86.

[104] M. S.D. Nezio, M. F. Pistonesi, W. D. Fragoso, M. J. C. Pontes, H. C. Goicoechea, M. C. U. Araújo, B. S. F. Band. *Successive projections algorithm improving the multivariate simultaneous direct spectrophotometric determination of five phenolic compounds in sea water*. **Microchem. J**. 85 (2007) 194– 200.

[105] A. F. C. Pereira, M. J. C. Pontes, F. F. Gambarra Neto, S. R. B. Santos, R. K. H. Galvão, M. C. U. Araújo. *NIR spectrometric determination of quality parameters in vegetable oils usingi PLS and variable selection*. **Food Res. Internat**. 41 (2008) 341–348.

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

[107] M. Goodarzi, W. Saeys, R. K. H. Galvão, M. C. U. Araújo, Y. V. Heyden. *Binary classification of chalcone derivatives with LDA or KNN based on their antileishmanial activity and molecular descriptors selected using the Successive Projections Algorithm feature-selection technique*. **Europ. J. Pharmac. Sci**. 51 (2014) 189–195.

[108] N. Goodarzi, M. Goodarzi, R. K. H. Galvão, M. C. U. Araújo. *QSPR Modeling of Soil Sorption Coefficients (K_{OC}) of Pesticides Using SPA-ANN and SPA-MLR*. J**. Agric. Food Chem.** 57 (2009) 7153–7158.

[109] S. K. B. Freitas, E. C. L. Nascimento, A. G. G. Dionizio, A. A. Gomes, R. K. H. Galvão, M. C. U. Araújo. *A flow-batch analyzer using a low cost aquarium pump for classification of citrus juice with respect to brand*. **Talanta** 107 (2013) 45–48.

[110] G. Véras, A. A. Gomes, A. C. Silva, A. L. B. Brito, P. B. A. Almeida, E. P. Medeiros. *Classification of biodiesel using NIR spectrometry and multivariate techniques*. **Talanta** 83 (2010) 565–568.

[111] M. Insausti, A. A. Gomes, F. V. Cruz, M. F. Pistonesi, M. C. U. Araújo, R. K. H. Galvão, C. F. Pereira, B. S. F. Band. *Screening analysis of biodiesel feedstock using UV–vis, NIR and synchronous fluorescence spectrometries and the successive projections algorithm*. **Talanta** 97 (2012) 579–583.

[112] M. J. C. Pontes, J. R. K. H. Galvão, C. Pasquini, M. C. U. Araújo, Cortez, R. M. Coelho, M. K. Chiba, M. F. Abreu, B. E. Madari. *Classification of Brazilian soils by using LIBS and variable selection in the wavelet domain*. **Anal. Chim. Acta** 642 (2009) 12–18.

[113] G. B. Costa, D. D. S. Fernandes, V. E. Almeida, T. S. P. Araújo, J. P. Melo, P. H. G. D. Diniz, G. Veras. *Digital image-based classi fication of biodiesel*. **Talanta** 139 (2015) 50– 55

[114] H. M. Paiva, S. F. C. Soares, R. K. H. Galvão, M. C. U. Araújo. *A graphical user interface for variable selection employing the Successive Projections Algorithm*. **Chemom. Intell. Lab. Syst.** 118 (2012) 260– 266.

[115] A. Mendonça, A. C. Rocha, A. C. Duarte, E. B. H. Santos. *The inner filter effects and their correction in fluorescence spectra of salt marsh humic matter*. **Anal. Chim.Acta** 788 (2013) 99–107.

[116] A. C. Olivieri, H. Wu, R. Yu. *MVC2: A MATLAB graphical interface toolbox for second-order multivariate calibration*. **Chemom. Intell. Lab. Syst**. 96 (20 09) 246–251.

[117] R.W. Kennard, L.A. Stone. *Computer Aided Design of Experiments*. **Technometrics** 11 (1969) 137.

[118] A. C. Olivieri. Analytical Figures of Merit: *From Univariate to Multiway Calibration*. **Chem. Rev.** 114 (2014) 5358–5378.

Anexo -1: Métricas de desempenho

1- Sensibilidade

$$SEN = \frac{\sigma_x}{\sigma_y} \quad \text{(A1.1)}$$

$$SEN_j = [{}^{var(x)}\big/_{var(y)}]^{\frac{1}{2}} = \{\boldsymbol{v}^T[\,\boldsymbol{P}^T(\boldsymbol{I} - \boldsymbol{Z_{int}}\boldsymbol{Z_{int}^{+}})\boldsymbol{P}]^{-1}\boldsymbol{v}\}^{-1} \quad \text{(A1.2)}$$

$$\boldsymbol{Z_{int}} = [\boldsymbol{I}_c \otimes \boldsymbol{b}_{int\,1} | \boldsymbol{c}_{int\,1} \otimes \boldsymbol{I_b}] \quad \text{(A1.3)}$$

SEN é a sensibilidade definida como a razão entre das incertezas (σ_x)sinal/(σ_y)concentração. O sobrescrito "*j*" indica o uso de matrizes Jacobiana. $\boldsymbol{v}$ são os coeficientes de regressão, **P** é a matriz dos pesos de calibração. **I** é uma matriz identidade com dimensão JK×JK e $\mathbf{Z_{int}}$ contem a informação dos constituintes não calibrados. $\mathbf{I}_c$ e $\mathbf{I}_b$ são matrizes unitárias com dimensões J×J e K×K respectivamente. $\mathbf{b_{int1}}$ e $\mathbf{c_{int1}}$ contem informação dos constituintes não calibrados. $\otimes$ é o porduto de Kronecker.

2- Sensibilidade Analítica

$$\gamma = \frac{\mathrm{SEN}}{\sigma_x} \quad \text{(A1.4)}$$

3- Limites de detecção e quantificação

$$\mathrm{LOD} = 3.3\ \mathrm{SD}\ (y) \quad \text{(A1.5)}$$

$$\mathrm{LOQ} = 10\ \mathrm{SD}(y) \quad \text{(A1.6)}$$

SD(y) é a incerteza da concentração do banco ou amostra com baixo teor do analito. SD é estimado com base em teoria de propagação de erros. Detalhes acerca das métricas de desempenho podem ser encontrados na referência [118] deste trabalho.

Apêndice 1

Produção científica da tese

Resumos em eventos

[1] GOMES, A. A.; GOICOECHEA, H. C.; ARAÚJO, M. C. U. *Algoritmo das Projeções Sucessivas para seleção de intervalos em calibração de segunda ordem: iSPA-U-PLS/RBL*. In: **I Escola de Inverno de Quimiometria**, São Carlos -SP. Anais da I EIQ, 2013.

[2] GOMES, A. A. ; SCHENONE, A. V. ; GOICOECHEA, H. C. ; ARAÚJO, M.C.U. *Algoritmo das Projeções Sucessivas para seleção de intervalos em PLS trilinear: iSPA-N-PLS*. In: **7º CAQA**, Mendoza. ANAIS 7 CAQA, 2013

Artigos

[1] GOMES, A. A., ALCARZ, M. R., GOICOECHAE, H. C., ARAÚJO, M. C. U.; *The Successive Projections Algorithm for interval selection in trilinear partial least-squares with residual bilinearization*. Anal. Chim. Acta 811 (2014) 13–22.

[2] GOMES, A. A.; SCHENONE, A. V.; GOICOECHEA, H. C.; ARAÚJO, M.C.U. *Unfolded partial least squares/residual bilinearization combined with the Successive Projections Algorithm for interval selection: enhanced excitation-emission fluorescence data modeling in the presence of the inner filter effect*. Anal. Bioanal. Chem. DOI 10.1007/s00216-015-8745-8.