"Desenvolvimento e Aplicação de Métodos Quimiométricos de Ordem Superior"

Tese de Doutorado

Fevereiro de 2002

Autor: Marlon Martins dos Reis

Orientadora: Dra. Márcia M. C. Ferreira

Depto. Físico-Química

Instituto de Química

Universidade Estadual de Campinas

UNICAMP

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO INSTITUTO DE QUÍMICA UNICAMP

R277d

Reis, Marlon Martins dos
Desenvolvimento e aplicação de métodos quimiométricos de ordem superior / Marlon Martins dos Reis. -- Campinas, SP: [s.n], 2002.

Orientadora: Márcia M. C. Ferreira.

Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Química.

1. Métodos em multi modos. 2. RBL. 3. NBRA. 4. PARAFAC-TUCKER. I. Ferreira, Márcia M.C. II. Universidade Estadual de Campinas. III. Titulo.

BANCA EXAMINADORA

Profa. Dra. Márcia Miguel Castro Ferreira (Orientadora)

Prof. Dr. Wilson Castro Ferreira Júnior (IME-UNICAMP)

Profa. Dra. Paula Fernandes de Aguiar (IQ-UFRJ)

Prof. Dr. Adalberto Bono Maurizio Sacchi Bassi (IQ-UNICAMP)

Prof. Dr. Ronei Jesus Poppi (IQ-UNICAMP)

Este exemplar corresponde à redação final da Tese de Doutorado defendida pelo aluno **MARLON MARTINS DOS REIS,** aprovada pela Comissão Julgadora em 06 de fevereiro de 2002.

Profa. Dra. Márcia Miguel Castro Ferreira

(Presidente da Banca)

- Agradecimentos –

Aos colaboradores:

Professora Silene pela amizade, pelo incentivo, prudência, precisão e interesse na discussão de manuscritos e resultados.

Professor Pessine pela disponibilidade, apoio na obtenção de dados e discussão de manuscritos.

A colega Débora pela amizade, pelo apoio, pela precisão na obtenção dos dados e discussão dos resultados e manuscritos.

Professor Age Smilde pela amizade, confiança e respeito, pela oportunidade, empenho e precisão na discussão de nosso trabalho.

Ao colega Steve pelo apoio, amizade e dedicação na elaboração do manuscrito de nossa colaboração.

Aos Colegas:

Marcelo pela duradoura amizade, apoio e grande incentivo, pela infinita paciência com as eternas lamentações filosóficas acerca de nosso trabalho.

Da Rep D'Ou, Amarildo, Cleber, Fábio (SJ), Fábio (Japonês), Gustavo e Relze, Helmut, Paulo, Thales e Wander, pela agradável convivência e grande alegria em meio ao caos de nossa república.

De grupo, Edilton pelo apoio, companherismo e precisão na discussão de nossos trabalhos, Cristiano pela inspiração, Ciriaco pelo exemplo, e pelo coleguismo a Eduardo, Fabiana, Luciana, Luci, Marcos Garcia, Marcos Guterres, Rudolf e Thais.

Da salinha, Aline, Beto, Ednalva, Fabiane, Juan, Julho, Hermes, Luis, Marcelo Miranda, Oyrton e Sayuri, pelo apoio e agradável convivência.

Do corredor, André, Ataulpa, Edilson, Max, Muftah, Pedro e Rogério.

De Amsterdã, Renger e Frans pela agradável convivência e amizade, a Johan e Sabina pelo coleguismo.

A orientadora e amiga Márcia, pela confiança, respeito, oportunidade, paciência e grande apoio.

A FAPESP pelo apoio financeiro.

E a todos aqueles que de certa forma contribuíram para a elaboração deste trabalho.

Aos meus pais, Lourdinha e Ricartinho, meus irmãos e irmã, Claudinho, Claudinha e Cacalo, pela compreensão na ausência, fundamental apoio na partida, exemplo primordial de luta, força e persistência e pelo infindável carinho. E ao Kiko, Preta, Bernado e Beatriz pelo grande carinho e apoio.

A minha querida esposa Mariza, pela eterna compreensão em ausências no desenvolvimento deste trabalho, pela paciência em tempos de relatório, apoio, pelo incentivo ao ir em frente, incomensurável carinho e zelo.

Aqui o meu muito obrigado!!!

Resumo

Dois dos métodos em multi modos apresentados neste trabalho, foram desenvolvidos na psicometria para estudos psicológicos. Para sua aplicação em dados químicos estes métodos requerem alguns refinamentos para acomodar variações inerentes aos dados químicos. Mesmo com o desenvolvimento deste tipo de refinamento, há alguns tipos de variação nestes dados que não podem ser descritos pelos métodos em multi modos. Neste sentido, o principal foco deste trabalho está voltado para a relação entre a parte dos dados que pode ser descrita e aquela que não pode ser pelos métodos em multi modos, em outras palavras, como se pode extrair uma informação útil de um conjunto de dados em multi modos. Três são os tópicos estudados: Calibração de segunda ordem; resolução de curvas; e análise exploratória em multi modos. Na calibração de segunda ordem é avaliado como o ambiente experimental influencia na estrutura em multi modos dos dados e como isto afeta os resultados da calibração. Na resolução de curvas, a variação presente nos dados, que não pode ser acomodada pelos métodos, interfere na escolha do número de curvas a serem resolvidas. Sendo assim, um procedimento de validação é apresentado para efetuar tal escolha. Em análise exploratória dois problemas são estudados. No primeiro, é sugerida uma metodologia para separar a parte dos dados que pode ser descrita daquela que não pode ser descrita pelos métodos. No segundo, um problema similar ao primeiro é estudado sendo usado um método para a suavização de curvas.

Com a atual popularidade dos instrumentos de segunda ordem (hifenados), há agora um número de técnicas quimiométricas para a chamada calibração de segunda ordem, *i.e.* quantificação de um analito de interesse na presença de um (ou mais) interferente(s). Instrumentos de segunda ordem produzem dados de complexidade variada, sendo um fenômeno particular a sobreposição de posto (ou deficiência de posto), onde o posto dos dados não é igual à soma dos valores de posto das espécies que contribuem para a medida instrumental. Uma das propostas deste trabalho é avaliar o desempenho de dois métodos de calibração de ordem superior, um baseado em uma solução de um problema de autovalores-autovetores e outro em uma aproximação por quadrados mínimos, em termos de sua habilidade de quantificar o analito de interesse e sua estabilidade, quando aplicados em dados de injeção em fluxo com sobreposição de posto. Na presença de alta colinearidade dos dados, o método baseado em uma aproximação por quadrados mínimos apresenta resultados um pouco melhores e mais estáveis. A análise de componentes em dois modos (TMCA) é empregada na investigação do porque da diferença entre os resultados destes dois métodos em termos das propriedades químicas das espécies analisadas. O sucesso da calibração de segunda ordem está, em geral, fortemente associada aos valores de pK_a dos compostos estudados.

A identificação de cromóforos em amostras biológicas demanda a separação física dos composto presentes nestas amostras. Embora várias vantagens, na identificação de substâncias, tenham sido geradas pelo uso das técnicas espectroscópicas hifenadas, misturas complexas de cromóforos que apresentam espectros sobrepostos que não podem ser identificados diretamente. Este trabalho apresenta uma aplicação quimiométrica para identificação de compostos em amostras biológicas através de uma técnica espectroscópica hifenada. O método PARAFAC, que não possui problema de rotação livre, é usado para a resolução de curvas de espectros de emissão-excitação coletados a partir de amostras de tártaro dentário humano. O PARAFAC foi aplicado sob restrição (*i.e.* unimodalidade e não-negatividade) e avaliado por um procedimento de validação. Os perfis resolvidos são do tipo porfirinico, pois apresentam bandas de excitação com máximos em 407, 416, e 431 nm, na região da banda Soret, característica deste tipo de composto.

Os métodos propostos por Ledyard R. Tucker durante a década de sessenta apresentam o problema de rotação livre, tornando sua interpretação difícil de ser efetuada. Com a proposta de fazer a análise de dados em multi modos mais aceitável, este trabalho sugere uma metodologia para a extração de informações úteis do conjunto de dados. Para tal, a metodologia proposta se baseia na decomposição do conjunto de dados em blocos em 3 modos através de Modelos Tucker. Com o objetivo de capturar em um bloco a informação similar sobre as propriedades dos dados, a decomposição proposta emprega o Modelo Tucker Restrito, onde o núcleo possui alguns elementos com valores fixados em zero. Esta metodologia foi aplicada com sucesso em dados de propriedades físicas e físico-químicas de amidos extraídos de quatro cultivares de mandioca, colhidas em oito diferentes idades no período normalmente usado na colheita para fins industriais.

Em geral, o método PARFAC emprega um espaço vetorial para descrever as matrizes resultantes da decomposição que promove. Este trabalho apresenta uma aplicação do PARAFAC onde os fatores resultantes da decomposição são considerados como sendo funções. Os objetos funcionais usados para ligar o PARAFAC e a análise funcional são as funções spline. Uma das vantagens do emprego de splines é a possibilidade de ativar a suavização dos perfis decompostos. A quantidade de suavização aplicada aos perfis é controlada, nesta metodologia, através de parâmetro de penalização, λ, ou pelo número de funções de base. Desta forma, Spline-PARAFAC requer o cálculo do parâmetro λ ou número de bases, que foram determinados através de uma validação cruzada ordinária. O Spline-PARFAC foi aplicado em um conjunto de dados formado por medidas horárias de monóxido de carbono durante os anos de 1997 e 1999 na cidade de São Paulo. O Spline-PARAFAC apresentou um bom desempenho descrevendo a maior variação diária de emissão deste gás e os efeitos sazonais ao longo do ano sobre esta variação.

Abstract

Two of the multi-way methods presented in this work were developed in psychometrics for psychological studies. On being applied in chemical data these methods demand on refinements to accommodate the variation present in the chemical data. Although the development of such kind of refinements there are some variation within the data which cannot be described by the multi-way methods. In this way, the main focus of the thesis is to describe the relationship between the parts of the data that can be described or not by the methods, in other words, how one can extract meaningful information from multi-way data set. Three topics are studied: Second order calibration; curve resolution and multi-way exploratory data analysis. In the second order calibration the influence of experimental behaviour over the multi-way structure of the data and how it affects the calibration results is evaluated. In the curve resolution, the variation of the data that cannot be accomodated by the methods, interfere on the choice of the number of curves to be resolved and a validadtion procedure is used on such choice. On the exploratory data analysis two problems are studied: In the first, it is suggested a methodology to discriminate the part of the data that can be described by the Multi Way methods from that one which cannot be described. In the second, a similar problem is treated where a smoothing method is used.

With the current popularity of second-order (or hyphenated) instruments, there now exists a number of chemometric techniques for the so-called second-order calibration problem, *i.e.* that of quantifying an analyte of interest in the presence of one (or more) interferent(s). Second-order instruments produce data of varying complexity, one particular phenomenon sometimes encountered being that of rank overlap (or rank deficiency), where the overall rank of the data is not equal to the sum of the ranks of the contributing species. One of the purpose of the present work is to evaluate the performance of two second-order calibration methods, a least squares-based and an eigenvalue-based solution, in terms of their quantitative ability and stability, as applied to flow injection analysis (FIA) data which exhibits rank overlap. In the presence of high collinearity in the data, the least squares methods is found to give a more stable solution. Two-mode component analysis (TMCA) is used to investigate the reasons for this difference in terms of the chemical properties of the species analyzed. The success of the second-order calibration in general is found to depend strongly on the associated pK_a values.

Chromophores identification in biological samples often demands on the physical separation of the compounds, which can be difficult. Although there are several advantages of hyphenated spectroscopic techniques on substances identification, complex mixtures of chromophores presenting overlapped spectra cannot be identified directly. This work presents a chemometric application to

compounds identification in biological samples by spectroscopic hyphenated techniques using a curve resolution method. The PARAllel FACtor analysis model (PARAFAC), which has no rotational indeterminacy, was used for curve resolution of excitation-emission spectra of human dental tartars. PARAFAC was applied under constraints (*i.e* unimodality and non-negativity) and evaluated with a validation procedure. The resolved profiles are porphyrinic like spectra presenting the excitation band maxima in 407, 416 and 431nm in the Soret band region (390-440nm) of these substances.

The original methods proposed by Ledyard R. Tucker during the 1960s presented the rotational freedom problem, making the interpretation of results obtained by these methods rather difficult to be carried out. On the proposal of making the multi-way data analysis more acceptable, this work suggests a methodology for extracting meaningful information from the data set. For that, the proposed methodology is based on the decomposition of the data set in 3-way blocks by using Tucker Models. With the aim of keeping in one block the similar information about data properties, it is proposed decomposition based on a Constrained Tucker Model, where the core array has some of its elements fixed to zero. This methodology is successfully applied to a data set formed by physical and physico chemical properties of starches of four cassava cultivars, harvested at different ages during the period usually taken for harvest of industrial uses.

The PARAFAC model has been used in several applications in chemistry, *e.g.* for overlapped spectra resolution, second order calibration and others. In general, the PARAFAC method uses a vector space approach by considering the resulting matrices from the decomposition as a collection of vectors. This work presents a PARAFAC application where the decomposition resulting factors are considered as being functions. The functional objects used to link PARAFAC method and functional analysis are spline functions. The methodology used to promote the Spline-PARAFAC decomposition is based on Bro-Sidiropoulos' approach for the unimodality constraint. One of the advantages of using splines is the possibility of achieving smoothing on the decomposed components. The amount of smoothing applied on the components is controlled in the presented methodology by a penalty parameter or by the number of basis functions. Thus Spline-PARFAC requires the calculation of the parameter λ and the number basis, which were found in this work by using an Ordinary Cross Validation OCV. The Spline-PARAFAC was applied to a carbon monoxide data set, which corresponds to concentrations measured every hour during the years of 1997 and 1999 in the São Paulo city in Brazil. Each data set was arranged in a Three Way Array of dimension (24 hours×5 days×52 weeks). The Spline-PARAFAC presented a good performance producing smoothed profiles which describe the dally variation of emitted gas and the seasonal effects during the year.

Súmula Curricular

Informações Pessoais

Marlon Martins dos Reis

Data de nascimento: 10/03/1974 Nacionalidade : Brasileira Naturalidade: Barbacena-MG

Estado civil : Casado Sexo: Masculino

Formação Acadêmica

- Doutor em ciências pela Universidade Estadual de Campinas.

Dept.de Físico-Química-Instituto de Química – Universidade Estadual de Campinas- UNICAMP. (Fevereiro de 2002) .Título da Tese de doutorado: "Desenvolvimento e Aplicação de Métodos Quimiométricos de Ordem Superior"

- Mestre em Quimiometria pela Universidade Estadual de Campinas.

Dept.Físico-Química-Instituto de Química – Universidade Estadual de Campinas- UNICAMP (Outubro-1997). Título da Dissertação de Mestrado: "Aplicação de Métodos Quimiométricos em Separação de Espectros e Reconhecimento de Padrões"

- Bacharel em Química Fundamental pela Universidade Federal de Juiz de Fora (Março-1996)

Lista de publicações

[1] Reis, M. M.; Bilioti, D.N.; Ferreira, M.M.C.; Pessine, F.B.T.; "PARAFAC for Curve Resolution: A Case Study Using Total Luminescence in human Dental Tartar", Applied Spectroscopy, 55 (7) (2001) 847-851.

[2] Reis, M.M; Gurden, S.P.; Smilde, A. K.; Ferreira, M. M.C.; "Calibration And Detailed Analysis Of Second-Order Flow Injection Analysis Data With Rank Overlap", Analytica Chimica Acta,422 (2000) 21-36.

[3] Sarmento, S.B.S; Reis, M.M.; Ferreira, M.M.C.; Cereda, M. P.; Penteado, M.V.C.; Anjos, C. B.; Análise Quimiométrica de Propriedades Físicas, Físico-Químicas e Funcionais de Féculas de Mandioca. Braz. J. Food Technol., 2(1,2) (1999) 131-137.

[4] Bilioti, D.N.; dos Reis, M. M.; Ferreira, M.M.C.; Pessine, F.B.T.; "Photochemical Behavior Under UVA Radiation of β-Cyclodextrin Included Parsol 1789 with Chemometric Approach", J. Molecular Structure 480-481(1999) 557-561.

[5] Reis, M. M.; Ferreira, M.M.C.; "Separação de Espectros Simulados e Luminescência Total Através do Método Generalizado de Anulação do Posto(GRAM)", Química Nova, 22 (1999) 11-17.

Trabalhos submetidos

[1] Reis, M. M.; Ferreira, M.M.C., PARAFAC with Splines: A Case Study, Journal of Chemometrics.

[2] Reis, M.M.; Ferreira, M.M.C.; Sarmento, S.B.S; A Multi-Way Analysis of Starch Cassava Properties, Chemometrics and Intell. Lab. Systems.

Outras informações biográficas.

Estágio relevante

Seis meses no grupo de quimiometria do departamento de engenharia química da Universidade de Amsterdã sob coordenação do Prof. Dr. Age K. Smilde. Neste período foi efetuado um estudo para avaliar a eficiência de um método quimiométrico (i.e. método matemático de interesse em química) empregado na quantificação de substâncias químicas na presença de interferentes. Para tal, foram analisados dados de um sistema por injeção em fluxo com gradiente de pH monitorado por espectroscopia de absorção na região do ultravioleta. Este trabalho rendeu uma publicação científica em periódico internacional.

Participações em congressos internacionais

Sétimo Congresso Escandinavo de Quimiometria (SSC7) - Dinamarca-Agosto de 2001.
Foram apresentados dois trabalhos, um em forma de apresentação oral, e outro em forma de poster.
Sexto Congresso Escandinavo de Quimiometria (SSC6) - Noruega-Agosto de 1999.
Foi apresentado um trabalho em forma poster.

Disciplinas cursadas no Mestrado e Doutorado

Planejamento e Otimização de Experimentos-A* , Calibração Multivariada-A*, Quimiometria-A*, Química Quântica-A*, Métodos Físico-Químicos em Química Orgânica (técnicas espectroscópicas)-B*, Matemática Aplicada I (Introdução a Equações Diferenciais Parciais)-A*. Obs.: * - Conceito obtido.

Apresentação de trabalhos em congressos

Congressos da Sociedade Brasileira de Química (SBQ): 2001-(AB074), 1999-(ED075), 1998-(QA133).

Simpósio Brasileiro de Química Teórica(SBQT): 1997-(P297).

European Congress on Molecular Spectroscopy (EUCMOS): 1998-(D115).

Encontro Nacional de Química Analítica (ENQA): 1997

Índice

1 Introdução

Em um dado experimento, 24 estudantes destros, 12 meninos e 12 meninas, são filmados ao desempenharem 7 tarefas, previamente determinadas, com pequenos blocos de madeira. Os filmes são analisados para a identificação de 19 tipos de movimentos das mãos, e para cada tipo de movimento é enumerado o número de vezes em que as mãos direita, esquerda ou ambas são usadas*[Harshman]*. Em outro experimento, informações acerca da qualidade da água de um rio são avaliadas em seis estações de coleta, localizadas em diferentes partes ao longo deste rio. Neste caso, são efetuadas medidas da temperatura, vazão do rio, pH, demanda química de oxigênio, demanda bioquímica de oxigênio e outras variáveis químicas, isto repetido em quatro diferentes épocas do ano *[Kiers (a)]*. Se L. Tucker[1] fosse quimiometrista e lhe fosse perguntado o que estes dois experimentos têm em comum, provavelmente ele responderia que ambos os conjuntos de dados resultantes destes experimentos possuem estrutura em multi modos e que poderiam ser analisados através de seus métodos (modelos). Estes exemplos são geralmente empregados para ilustrar os dados em multi modos, seja na psicometria ou na química, e o fazem bem, por evidenciarem as características intrínsecas dos dados.

Em ambos os casos mencionados, um problema deve ser estudado e as informações sobre este problema foram transcritas em números. A determinação da relação entre estas informações numéricas e o problema não é tarefa trivial. Dentre os vários métodos desenvolvidos na estatística e em áreas correlatas, (psicometria, econometria, biometria, etc.), a análise de fatores têm sido amplamente usada para o estudo desta relação (informações numéricas *versus* problema). Este tipo de análise objetiva a identificação de características latentes do conjunto de dados através do estudo da "covariação" das variáveis analisadas (por exemplo, a covariação entre a demanda química de oxigênio e época do ano, ou seja, se variando a época do ano a demanda química de oxigênio também sofre variação). Neste caso, a análise direta da relação entre variáveis, de sua "covariação", pode se tornar proibitiva quando o número de variáveis é grande, devido ao grande número de combinações dois a dois que é gerado. Para solucionar este problema, a análise de fatores propõe a transformação das variáveis originais em variáveis não correlacionas, também chamadas de variáveis latentes, que então são analisadas. Os métodos de ordem superior ou em multi modos são uma extensão natural da análise de fatores para dados com três ou mais modos (deve-se entender modo com sendo a característica estudada, por exemplo, um conjunto de dados em dois modos seria aquele formado por coletas realizadas em uma

[1] L. Tucker era psicometrista e desenvolveu modelos para o tratamento de dados de ordem superior *[Tucker]*.

única estação, sendo o modo 1 as quatro diferentes épocas do ano, e o modo 2 as variáveis físico-químicas).

O estudo da relação entre informações numéricas e problema através da análise de fatores é chamado, em certos casos, de análise exploratória e visa "explorar" os dados para identificar informações que possibilitem a compreensão do problema estudado.

Embora a identificação da estrutura em multi modos seja, para alguns, imediata, sua análise não o é, como nos casos de dados em dois modos. Isto ocorre, pois a análise de fatores para dados em dois modos é baseada, geralmente, em uma transformação algébrica (*i.e*. projeção ortogonal), enquanto que os métodos usados para análise de dados em multi modos fazem uso de algumas suposições acerca dos dados para realizar as transformações entre variáveis. Assim, em certos casos, os métodos em multi modos não acomodam todas as variações presentes nos dados analisados. Neste sentido, este trabalho de tese apresenta algumas análises de dados, por meio de métodos de ordem superior, prestando especial atenção àqueles desvios das suposições, acerca dos dados, admitidas com verdadeiras pelos métodos. Desta forma, o título desta tese, que a princípio se mostra ousado pela proposta de desenvolvimento de métodos, se refere àqueles desenvolvimentos necessários para a aplicação dos métodos de ordem superior em certos problemas encontrados na química, onde tais desvios não possam ser evitados.

Neste trabalho de tese, três tópicos, comuns na literatura quimiométrica, são abordados: Calibração de segunda ordem; separação de curvas; e análise exploratória de dados. O primeiro, calibração de segunda ordem, objetiva a construção de modelos para a determinação da concentração de certo composto químico, através de informações instrumentais, em amostras cujas concentrações sejam desconhecidas. A calibração de segunda ordem recebe este nome por empregar a estrutura dos dados em multi modos na construção do modelo. O segundo tema, separação de espectros ou de curvas, emprega as características dos dados, estrutura em multi modos, para resolver curvas sobrepostas, ou seja, respostas instrumentais obtidas a partir de misturas de compostos químicos, que correspondem à soma das respostas individuais de cada composto destas misturas. No terceiro tópico, análise exploratória, os métodos são empregados para evidenciar uma característica latente dos dados.

A principal vantagem da calibração de segunda ordem é possibilitar a identificação de um analito de interesse na presença de interferentes, sendo esta vantagem chamada de 'vantagem de segunda ordem'. Diversos métodos têm sido sugeridos na construção de modelos de calibração, onde a principal diferença está em suas características algébricas. A primeira classe de métodos se baseia na solução de um problema de autovalores-autovetores, possuindo como principal vantagem, a eficiência

computacional e desvantagem, a 'rigidez' do modelo construído, ou seja, pouco robusto frente às variações instrumentais/experimentais. A outra classe de métodos se baseia em uma aproximação por quadrados mínimos, onde uma otimização é efetuada para minimizar a diferença entre modelo sugerido e dados. Este tipo de método tem como vantagens a flexibilidade frente às variações experimentais e a possibilidade da inclusão de informações conhecidas sobre o analito na aplicação do modelo de calibração. As principais desvantagens deste tipo de abordagem são a necessidade da convergência do método para uma solução global no processo de otimização e o desempenho computacional. Neste tipo de problema, calibração de segunda ordem, a reprodutibilidade e ruídos experimentais são fatores determinantes do desempenho dos modelos construídos. Isto, pois estas variações experimentais alteram a estrutura dos dados, suposta constante pelos métodos usados na construção dos modelos de calibração. Com o objetivo de estudar a estabilidade das duas abordagens algébricas citadas, solução de um problema de autovalores-autovetores e uma aproximação por quadrados mínimos, frente à variações experimentais, este trabalho apresenta uma calibração de segunda ordem onde dois métodos com características gerais são avaliados. O principal objetivo desta aplicação é mostrar que além das características dos dados, estrutura em multi modos, as variações daquelas suposições, feitas pelos métodos acerca dos dados, devem ser consideradas na construção do modelo de calibração.

A separação de curvas em dados de ordem superior também pode ser efetuada por métodos baseados na solução de um problema de autovalores-autovetores ou por aproximação por quadrados mínimos. Em geral, os métodos baseados em aproximação por quadrados mínimos apresentam os melhores resultados, e portanto este tipo de aproximação é usado neste trabalho. Na resolução de curvas apresentada neste trabalho, o principal problema abordado é a determinação do número de curvas a serem resolvidas, que não pode ser determinado diretamente devido às variações experimentais, como ruídos. Neste caso, um fator determinante é a resolução instrumental (por exemplo, um espectro de absorção coletado com intervalo de 1nm tem resolução maior que um coletado com intervalo de 4nm) pois está diretamente relacionada à diferenciação entre as curvas. Desta forma, para uma dada resolução experimental é necessário determinar o número de curvas possíveis de serem resolvidas. Adicionalmente, deve ser determinado se a solução do método empregado, neste caso, baseado em aproximação por quadrados mínimos e consequentemente em uma otimização, é a melhor, isto é, se é a solução global para a otimização. Neste trabalho, é empregado um processo de validação para avaliar o número de curvas e a solução global para a otimização. Enfim, o principal objetivo deste trabalho, ao apresentar uma resolução de curvas, é mostrar a importância do

processos de validação para a determinação da melhor solução por meio de um método de ordem superior.

A análise exploratória visa evidenciar, como já mencionado, as características latentes dos dados. Na prática, estas características podem não estar evidentes devido às demais variações presentes nos dados. Assim, a principal meta em uma análise exploratória é separar uma dada característica latente das demais variações, sendo este, o problema fundamental deste tipo de análise. Neste sentido, as duas análises exploratórias apresentadas neste trabalho sugerem formas para identificar se a separação proporcionada pelos métodos é apropriada para a análise dos dados. Ou seja, se aquelas variações separadas pelos métodos, dentre aquele conjunto de variações presentes nos dados, identificam o problema estudado.

Este trabalho de tese é organizado em seções, dentre as quais, quatro descrevem os problemas mencionados acima de forma independente. Uma seção de fundamentos é apresentada para facilitar a leitura dos demais tópicos. Adicionalmente, é apresentado um glossário com termos e palavras usadas ao longo do texto.

2 Fundamentos

2.1 Introdução

O objetivo desta seção é introduzir algumas formulações algébricas úteis na compreensão da descrição dos métodos de ordem superior. Esta seção em conjunto com o glossário e notação, descritos na seção anexos, foram elaborado para facilitar a leitura deste trabalho de tese.

Considere um experimento no qual uma amostra contendo o composto 1 é injetada em uma coluna cromatográfica, onde a cada tempo é coletada uma alíquota do eluente e a concentração do composto 1 nesta alíquota é quantificada, isto feito para em um total de *n* alíquotas em um dado intervalo de tempo. A soma das concentrações do composto 1 nas *n* alíquotas é proporcional à concentração do composto 1 na amostra injetada. Segundo as propriedades de difusão do composto 1, sua concentração em uma dada alíquota mantém uma relação com o respectivo tempo de eluição, que é característica para cada composto. Desta forma, a curva obtida no gráfico das concentrações do composto 1 nas *n* alíquotas versus os tempos de eluição desta alíquotas é característica para cada composto, sendo chamada de cromatograma, possuindo importante papel na quantificação e identificação de compostos. Considerando agora, a injeção de uma amostra na qual o composto 1 aparece em concentração unitária, o cromatograma, ou perfil cromatográfico, é descrito aqui pela expressão **1** onde a soma de todas as *n* alíquotas é normalizada para um (soma-se as concentrações do composto 1 em todas as *n* alíquotas e divide-se cada uma destas concentrações pelo resultado desta soma).

$$\mathbf{t} = \begin{pmatrix} t_{11} \\ t_{21} \\ \vdots \\ t_{n1} \end{pmatrix} \qquad 1$$

onde t_{11} é porção (concentração) do composto 1 coletado na alíquota 1, o mesmo sendo válido para os outros tempos.

Para os casos onde a concentração do composto 1 não é unitária, o cromatograma pode ser descrito segundo a expressão **2**

$$\mathbf{t}c_{11} = \begin{pmatrix} t_{11} \\ t_{21} \\ \vdots \\ t_{n1} \end{pmatrix} c_{11} \qquad 2$$

onde c_{11} é a concentração do composto 1 na amostra injetada.

Agora considere que o composto 1 apresenta absorção em uma dada faixa espectral e que a lei de Beer *[Atkins]* seja válida. As abstortividades molares para este composto, nos *m* comprimentos de onda usados para descrever a absorção na faixa estudada, são dadas pela expressão **3**.

$$\mathbf{s} = \begin{pmatrix} s_{11} \\ s_{21} \\ \vdots \\ s_{m1} \end{pmatrix} \qquad 3$$

onde s_{11} é a absortividade molar do composto 1 para o comprimento de onda 1, o mesmo sendo válido para o restante das absortividades molares.

Segundo a lei de Beer, a absorbância do composto no comprimento de onda 1 é proporcional à concentração deste composto, ou seja, absorbância da amostra= $c_{11}s_{11}$. Assim, se a absorbância, no comprimento de onda 1, da alíquota 1 é medida, o resultado é: absorbância da alíquota 1 = $t_{11}c_{11}s_{11}$. Se o experimento for elaborado de tal forma que para cada alíquota, em um total de *n*, é coletado um espectro de absorção em *m* comprimentos de onda, o resultado final é uma matriz de dados com *n* linhas e *m* colunas, correspondendo cada linha a um espectro de absorção coletado em *m* comprimentos de onda. Esta matriz é dada pela expressão **4**.

$$\mathbf{t}c_{11}\mathbf{s}^T = \begin{pmatrix} t_{11} \\ t_{21} \\ \vdots \\ t_{n1} \end{pmatrix} c_{11} \begin{pmatrix} s_{11} & s_{21} & \cdots & s_{m1} \end{pmatrix} \qquad 4$$

Para o caso onde a amostra injetada é formada pela mistura do composto 1 e composto 2, que também apresenta absorção na faixa espectral empregada, e eluição na faixa de tempo estudada, a matriz de dados é dada pela expressão **5**.

$$\mathbf{t}_1 c_{11} \mathbf{s}_1^T + \mathbf{t}_2 c_{12} \mathbf{s}_2^T = \begin{pmatrix} t_{11} \\ t_{21} \\ \vdots \\ t_{n1} \end{pmatrix} c_{11} \begin{pmatrix} s_{11} & s_{21} & \cdots & s_{m1} \end{pmatrix} + \begin{pmatrix} t_{12} \\ t_{22} \\ \vdots \\ t_{n2} \end{pmatrix} c_{12} \begin{pmatrix} s_{12} & s_{22} & \cdots & s_{m2} \end{pmatrix} \quad 5$$

onde t_{12}, c_{12}, s_{12} correspondem à porção relativa do cromatograma para o tempo 1, à concentração do composto 2 e à absortividade molar do composto 2 no comprimento de onda 1, respectivamente, sendo válido também para os demais comprimentos de onda e tempos de eluição.

Com o emprego de alguma manipulação algébrica a expressão **5** pode ser reescrita em termos da expressão **6** ou expressão **7**.

$$\mathbf{t}_1 c_{11} \mathbf{s}_1^T + \mathbf{t}_2 c_{12} \mathbf{s}_2^T = \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \\ \vdots & \vdots \\ t_{n1} & t_{n2} \end{pmatrix} \times \begin{pmatrix} c_{11} & 0 \\ 0 & c_{12} \end{pmatrix} \times \begin{pmatrix} s_{11} & s_{21} & \cdots & s_{m1} \\ s_{12} & s_{22} & \cdots & s_{m2} \end{pmatrix} \quad 6$$

$$\mathbf{t}_1 c_{11} \mathbf{s}_1^T + \mathbf{t}_2 c_{12} \mathbf{s}_2^T = \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \\ \vdots & \vdots \\ t_{n1} & t_{n2} \end{pmatrix} \times \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \times \begin{bmatrix} c_{11} \begin{pmatrix} s_{11} & s_{21} & \cdots & s_{m1} \\ s_{12} & s_{22} & \cdots & s_{m2} \end{pmatrix} \\ c_{12} \begin{pmatrix} s_{11} & s_{21} & \cdots & s_{m1} \\ s_{12} & s_{22} & \cdots & s_{m2} \end{pmatrix} \end{bmatrix} \quad 7$$

2.2 O produto de Kronecker "⊗"

O símbolo "⊗" representa o produto de Kronecker, também é chamado de produto tensorial , que possui propriedades bem definidas e de grande importância em formulações algébricas. Dadas a matrizes $\mathbf{A}=a_{ij}$ ($n\times m$) e $\mathbf{B}=b_{ij}$ ($p\times q$) o produto de Kronecker entre **A** e **B** é definido como:

$\mathbf{A}\otimes\mathbf{B}= a_{ij}\,\mathbf{B}(np\times mq)$.

Este produto apresenta as segunite operações:

a- $\mathbf{A}\otimes 0 = 0\otimes\mathbf{A} = 0$;

b- $\alpha\mathbf{A}\otimes\beta\mathbf{B} = \alpha\beta(\mathbf{A}\otimes\mathbf{B})$, onde α e β são constantes;

c- $(\mathbf{A}+\mathbf{B})\otimes\mathbf{C} = (\mathbf{A}\otimes\mathbf{C})+(\mathbf{B}\otimes\mathbf{C})$;

d- $\mathbf{A}\otimes(\mathbf{B}+\mathbf{C}) = (\mathbf{A}\otimes\mathbf{B})+(\mathbf{A}\otimes\mathbf{C})$;

e- $\mathbf{A}\otimes(\mathbf{B}\otimes\mathbf{C}) = (\mathbf{A}\otimes\mathbf{B})\otimes\mathbf{C}$;

f- $(\mathbf{A}\otimes\mathbf{B})^{T} = \mathbf{A}^{T}\otimes\mathbf{B}^{T}$;

g- $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = (\mathbf{AC})\otimes(\mathbf{BD})$;

h- $(\mathbf{A}\otimes\mathbf{B})^{-1} = \mathbf{A}^{-1}\otimes\mathbf{B}^{-1}$, se $\mathbf{A}^{-1}$ e $\mathbf{B}^{-1}$ existem.

Na expressão **8** é mostrado o emprego deste produto através de duas pequenas matrizes.

$$\mathbf{A}\otimes\mathbf{B} = \begin{pmatrix} a & d \\ b & e \\ c & f \end{pmatrix} \otimes \begin{pmatrix} \alpha & \delta & \varphi \\ \beta & \varepsilon & \gamma \\ \chi & \phi & \eta \end{pmatrix} = \begin{pmatrix} a\begin{pmatrix} \alpha & \delta & \varphi \\ \beta & \varepsilon & \gamma \\ \chi & \phi & \eta \end{pmatrix} & d\begin{pmatrix} \alpha & \delta & \varphi \\ \beta & \varepsilon & \gamma \\ \chi & \phi & \eta \end{pmatrix} \\ b\begin{pmatrix} \alpha & \delta & \varphi \\ \beta & \varepsilon & \gamma \\ \chi & \phi & \eta \end{pmatrix} & e\begin{pmatrix} \alpha & \delta & \varphi \\ \beta & \varepsilon & \gamma \\ \chi & \phi & \eta \end{pmatrix} \\ c\begin{pmatrix} \alpha & \delta & \varphi \\ \beta & \varepsilon & \gamma \\ \chi & \phi & \eta \end{pmatrix} & f\begin{pmatrix} \alpha & \delta & \varphi \\ \beta & \varepsilon & \gamma \\ \chi & \phi & \eta \end{pmatrix} \end{pmatrix}, \qquad 8$$

onde, por exemplo,

$$a\begin{pmatrix} \alpha & \delta & \varphi \\ \beta & \varepsilon & \gamma \\ \chi & \phi & \eta \end{pmatrix} = \begin{pmatrix} a\alpha & a\delta & a\varphi \\ a\beta & a\varepsilon & a\gamma \\ a\chi & a\phi & a\eta \end{pmatrix}. \qquad 9$$

Assim, com emprego do produto de Kronecker a expressão **7** pode ser reescrita em termos da expressão **10**.

$$\mathbf{t}_1 c_{11} \mathbf{s}_1^T + \mathbf{t}_2 c_{12} \mathbf{s}_2^T = \begin{pmatrix} t_{11}\, t_{12} \\ t_{21}\, t_{22} \\ \vdots \quad \vdots \\ t_{n1}\, t_{n2} \end{pmatrix} \times \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \times \left[\begin{pmatrix} c_{11} & c_{12} \end{pmatrix}^T \otimes \begin{pmatrix} s_{11} & s_{21} & \cdots & s_{m1} \\ s_{12} & s_{22} & \cdots & s_{m2} \end{pmatrix} \right] \quad 10$$

Com a denominação descrita nas expressões **11**, **12**, **13** e **14** a expressão **10** pode ser descrita segundo a expressão **15**.

$$\mathbf{T} = \begin{pmatrix} t_{11}\, t_{12} \\ t_{21}\, t_{22} \\ \vdots \quad \vdots \\ t_{n1}\, t_{n2} \end{pmatrix} \quad 11$$

$$\mathbf{S}^T = \begin{pmatrix} s_{11} & s_{21} & \cdots & s_{m1} \\ s_{12} & s_{22} & \cdots & s_{m2} \end{pmatrix} \quad 12$$

$$\mathbf{I}_{DS} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad 13$$

$$\mathbf{N} = \mathbf{t}_1 c_{11} \mathbf{s}_1^T + \mathbf{t}_2 c_{12} \mathbf{s}_2^T \quad 14$$

$$\mathbf{N} = \mathbf{T}\mathbf{I}_{DS} \left(\begin{pmatrix} c_{11} & c_{12} \end{pmatrix}^T \otimes \mathbf{S}^T \right) \quad 15$$

Agora um exemplo onde duas amostras são comparadas, sendo estas amostras formadas por misturas dos compostos 1 e 2 em diferentes concentrações, é discutido. A primeira amostra corresponde àquela já descrita pela expressão **14**. A segunda é dada pela expressão **16**.

$$\mathbf{M} = \mathbf{t}_1 c_{21} \mathbf{s}_1^T + \mathbf{t}_2 c_{22} \mathbf{s}_2^T \quad 16$$

onde $\mathbf{t}_1$, c_{21}, $\mathbf{s}_1$ correspondem ao perfil cromatográfico (ou cromatograma para concentração unitária), à concentração do composto 1 na amostra 2 e às absortividade molares do composto 1 para os *m* comprimentos de onda, respectivamente. De forma semelhante, $\mathbf{t}_2$, c_{22}, $\mathbf{s}_2$ correspondem ao perfil cromatográfico, à concentração do composto 2 na amostra 2 e às absortividades molares do composto 2 para os *m* comprimentos de onda, respectivamente.

A justaposição lateral (*i.e.* formar uma nova matriz através da junção de duas ou mais matrizes lateralmente), das matrizes correspondentes à amostras 1 e 2 (*i.e.* **N** e **M**) é definida pela expressão **17**.

$$(\mathbf{N} \quad | \quad \mathbf{M}) = \left(\mathbf{t}_1 c_{11} \mathbf{s}_1^T + \mathbf{t}_2 c_{12} \mathbf{s}_2^T \quad | \quad \mathbf{t}_1 c_{21} \mathbf{s}_1^T + \mathbf{t}_2 c_{22} \mathbf{s}_2^T\right), \tag{17}$$

onde " | " indica a separação "fictícia" entre as duas matrizes.

Com o emprego de uma manipulação algébrica similar àquela já discutida, $(\mathbf{N} \quad | \quad \mathbf{M})$ pode ser reescrita em termos da expressão **22** segundo **18**, **19** e **20**.

$$(\mathbf{N} \quad | \quad \mathbf{M}) = \left(\mathbf{TI}_{DS}\left(\begin{pmatrix} c_{11} & c_{12} \end{pmatrix}^T \otimes \mathbf{S}^T\right) \quad | \quad \mathbf{TI}_{DS}\left(\begin{pmatrix} c_{21} & c_{22} \end{pmatrix}^T \otimes \mathbf{S}^T\right)\right) \tag{18}$$

$$(\mathbf{N} \quad | \quad \mathbf{M}) = \mathbf{TI}_{DS}\left(\left(\begin{pmatrix} c_{11} & c_{12} \end{pmatrix}^T \otimes \mathbf{S}^T\right) \quad | \quad \left(\begin{pmatrix} c_{21} & c_{22} \end{pmatrix}^T \otimes \mathbf{S}^T\right)\right) \tag{19}$$

$$(\mathbf{N} \quad | \quad \mathbf{M}) = \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \\ \vdots & \vdots \\ t_{n1} & t_{n2} \end{pmatrix} \times \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \times \begin{bmatrix} c_{11}\begin{pmatrix} s_{11} & s_{21} & \cdots & s_{m1} \\ s_{12} & s_{22} & \cdots & s_{m2} \end{pmatrix} & c_{21}\begin{pmatrix} s_{11} & s_{21} & \cdots & s_{m1} \\ s_{12} & s_{22} & \cdots & s_{m2} \end{pmatrix} \\ c_{12}\begin{pmatrix} s_{11} & s_{21} & \cdots & s_{m1} \\ s_{12} & s_{22} & \cdots & s_{m2} \end{pmatrix} & c_{22}\begin{pmatrix} s_{11} & s_{21} & \cdots & s_{m1} \\ s_{12} & s_{22} & \cdots & s_{m2} \end{pmatrix} \end{bmatrix} \tag{20}$$

fazendo

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} \tag{21}$$

$$(\mathbf{N} \quad | \quad \mathbf{M}) = \mathbf{TI}_{DS}\left(\mathbf{C}^T \otimes \mathbf{S}^T\right) \qquad 22$$

A descrição algébrica mostrada em **22** é de grande importância, pois é a base do modelo PARAFAC a ser descrito em seções posteriores.

2.3 O operador *vec*

Outro importante operador empregado aqui é o operador *vec* ou de vetorização. Este operador transforma uma matriz em um vetor coluna, como exemplificado na expressão **23**.

$$vec\begin{pmatrix} a & d \\ b & e \\ c & f \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \\ d \\ e \\ f \end{pmatrix} \qquad 23$$

O resultado da aplicação do operador *vec* na matriz de dados da amostra 1, **N**, é mostrado na expressão **24** , que pode ser compreendida a partir da expressão **5**, onde a primeira coluna da matriz **N** é dada pela soma $\mathbf{t}_1 c_{11} s_{11} + \mathbf{t}_2 c_{12} s_{12}$.

$$vec\mathbf{N} = \begin{pmatrix} \begin{pmatrix} t_{11} \\ t_{21} \\ \vdots \\ t_{n1} \end{pmatrix} c_{11}s_{11} + \begin{pmatrix} t_{12} \\ t_{22} \\ \vdots \\ t_{n2} \end{pmatrix} c_{12}s_{12} \\ \begin{pmatrix} t_{11} \\ t_{21} \\ \vdots \\ t_{n1} \end{pmatrix} c_{11}s_{21} + \begin{pmatrix} t_{12} \\ t_{22} \\ \vdots \\ t_{n2} \end{pmatrix} c_{12}s_{22} \\ \vdots \\ \begin{pmatrix} t_{11} \\ t_{21} \\ \vdots \\ t_{n1} \end{pmatrix} c_{11}s_{m1} + \begin{pmatrix} t_{12} \\ t_{22} \\ \vdots \\ t_{n2} \end{pmatrix} c_{12}s_{m2} \end{pmatrix} \qquad 24$$

Com alguma manipulação algébrica, como a dada nas expressões **25** e **26**, a expressão **24** pode ser reescrita em termos de **27**.

$$vec\mathbf{N} = \begin{pmatrix} \begin{pmatrix} t_{11} \\ t_{21} \\ \vdots \\ t_{n1} \end{pmatrix} s_{11} & \begin{pmatrix} t_{12} \\ t_{22} \\ \vdots \\ t_{n2} \end{pmatrix} s_{12} \\ \begin{pmatrix} t_{11} \\ t_{21} \\ \vdots \\ t_{n1} \end{pmatrix} s_{21} & \begin{pmatrix} t_{12} \\ t_{22} \\ \vdots \\ t_{n2} \end{pmatrix} s_{22} \\ \vdots & \\ \begin{pmatrix} t_{11} \\ t_{21} \\ \vdots \\ t_{n1} \end{pmatrix} s_{m1} & \begin{pmatrix} t_{12} \\ t_{22} \\ \vdots \\ t_{n2} \end{pmatrix} s_{m2} \end{pmatrix} \begin{pmatrix} c_{11} \\ c_{12} \end{pmatrix} \qquad 25$$

$$vec\mathbf{N} = \begin{pmatrix} \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \\ \vdots & \vdots \\ t_{n1} & t_{n2} \end{pmatrix} s_{11} & \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \\ \vdots & \vdots \\ t_{n1} & t_{n2} \end{pmatrix} s_{12} \\ \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \\ \vdots & \vdots \\ t_{n1} & t_{n2} \end{pmatrix} s_{21} & \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \\ \vdots & \vdots \\ t_{n1} & t_{n2} \end{pmatrix} s_{22} \\ \vdots & \\ \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \\ \vdots & \vdots \\ t_{n1} & t_{n2} \end{pmatrix} s_{m1} & \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \\ \vdots & \vdots \\ t_{n1} & t_{n2} \end{pmatrix} s_{m2} \end{pmatrix} \times \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} c_{11} \\ c_{12} \end{pmatrix} \qquad 26$$

$$vec\mathbf{N} = \left(\begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \\ \vdots & \vdots \\ t_{n1} & t_{n2} \end{pmatrix} \otimes \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \\ \vdots & \vdots \\ s_{m1} & s_{m2} \end{pmatrix} \right) \times \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} \times \begin{pmatrix} c_{11} \\ c_{12} \end{pmatrix} \quad 27$$

Com um pouco mais de álgebra a expressão **27** pode ser reescrita em termos da expressão **29**, através de **28**.

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} = \left(vec\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad vec\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right) \quad 28$$

$$vec\mathbf{N} = (\mathbf{T} \otimes \mathbf{S}) \left(vec\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad vec\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right) \begin{pmatrix} c_{11} \\ c_{12} \end{pmatrix} \quad 29$$

2.4 A notação $\underline{\mathbf{X}}$

Considerando agora o caso onde duas amostras são analisadas de forma semelhante àquela descrita na expressão **22**, a resposta experimental também pode ser arranjada em uma matriz com dois vetores coluna, sendo o primeiro para a amostra 1 e o segundo para a amostra 2. Para tal, considere a adoção da notação $\underline{\mathbf{X}}$ que indica que esta matriz é formada pela justaposição lateral de várias matrizes que possuem a mesma estrutura, como dado na expressão **30** para o exemplo de duas amostras. Esta notação é usada na literatura para indicar um arranjo em multi modos .

$$\underline{\mathbf{X}} = (\mathbf{N} \mid \mathbf{M}) \quad 30$$

Neste trabalho, admite-se que o operador *vec* quando aplicado a matrizes do tipo $\underline{\mathbf{X}}$ resullte na aplicação deste operador em cada matriz justaposta (esta propriedade do *vec* é usada neste trabalho para facilitar a descrição da formulação, mas se restringe a este trabalho). Um exemplo disto é apresentado na expressão **31**.

$$vec\underline{\mathbf{X}} = \left(vec\mathbf{N} \quad | \quad vec\mathbf{M}\right) \tag{31}$$

O emprego destas notações e das expressões **32** e **33** possibilita reescrever a expressão **34** em termos de **35**.

$$\underline{\mathbf{I}_{DS}} = \left(\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\right) \tag{32}$$

$\underline{\mathbf{I}_{DS}}$ é denominado arranjo diagonal superior,

$$vec\underline{\mathbf{I}_{DS}} = \left(vec\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad vec\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\right) \tag{33}$$

$$\left(vec\mathbf{N} \quad | \quad vec\mathbf{M}\right) = \left(\mathbf{T} \otimes \mathbf{S}\right)\left(vec\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad vec\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\right)\mathbf{C}^T \tag{34}$$

$$vec\underline{\mathbf{X}} = \left(\mathbf{T} \otimes \mathbf{S}\right)vec\underline{\mathbf{I}_{DS}}\mathbf{C}^T \tag{35}$$

A expressão **35** descreve o arranjo em multi modos de fundamental importância para a elaboração do modelo PARAFAC.

A formulação geral deste tipo de estrutura é mostrada na expressão **36**, que corresponde ao Modelo Tucker, a ser descrito em seções posteriores.

$$\underline{\mathbf{X}} = \mathbf{A}\underline{\mathbf{G}}\left(\mathbf{C}^T \otimes \mathbf{B}^T\right) \tag{36}$$

onde

$$\underline{\mathbf{X}} = \left(\mathbf{X}_1 \quad | \quad \mathbf{X}_2\right) \tag{37}$$

$$\underline{\mathbf{G}} = \left(\mathbf{G}_1 \quad | \quad \mathbf{G}_2\right) = \begin{pmatrix} g_{111} & g_{112} & g_{211} & g_{212} \\ g_{121} & g_{122} & g_{221} & g_{222} \end{pmatrix}. \tag{38}$$

Em termos gerais, as matrizes $\mathbf{X}_1$, $\mathbf{X}_2$,…, $\mathbf{X}_n$ de $\underline{\mathbf{X}}$ serão chamadas de "fatias", e $\underline{\mathbf{G}}$ de núcleo, que também possui estrutura em multi modos como $\underline{\mathbf{X}}$.

2.5 Quadrados Mínimos Alternantes "QMA"

Neste trabalho o algoritmo empregado na determinação das matrizes **A**, **B**, **C**, e $\underline{\mathbf{G}}$, é chamado de Quadrados Mínimos Alternantes "QMA" (*do inglês Alternating Least Squares*) que tem por princípio a minimização da função dada na expressão **39**. (No Apêndice 2.1 é apresentada uma breve introdução do problema de Quadrados Mínimos.)

$$l\left(\mathbf{A},\mathbf{C},\mathbf{B},\underline{\mathbf{G}}\right) = \left\|\underline{\mathbf{X}} - \mathbf{A}\underline{\mathbf{G}}\left(\mathbf{C}^T \otimes \mathbf{B}^T\right)\right\|^2 \tag{39}$$

sendo $\|\cdot\|^2$ usado para indicar a soma dos quadrados dos elementos de " $\cdot$ ".

No caso do modelo PARAFAC, $\underline{\mathbf{G}}$ é substituído por $\underline{\mathbf{I}_{DS}}$, que é mantido fixo durante o processo de otimização.

O processo de otimização do QMA, a princípio, faz a busca pelo mínimo da função *l* fixando todas a direções menos uma *[Kiers(b)]*. Um exemplo deste tipo de minimização é dado na Figura 2.1, onde a função dada pela expressão **40** é minimizada.

$$f\left(x,y\right) = x^2 + y^2 \tag{40}$$

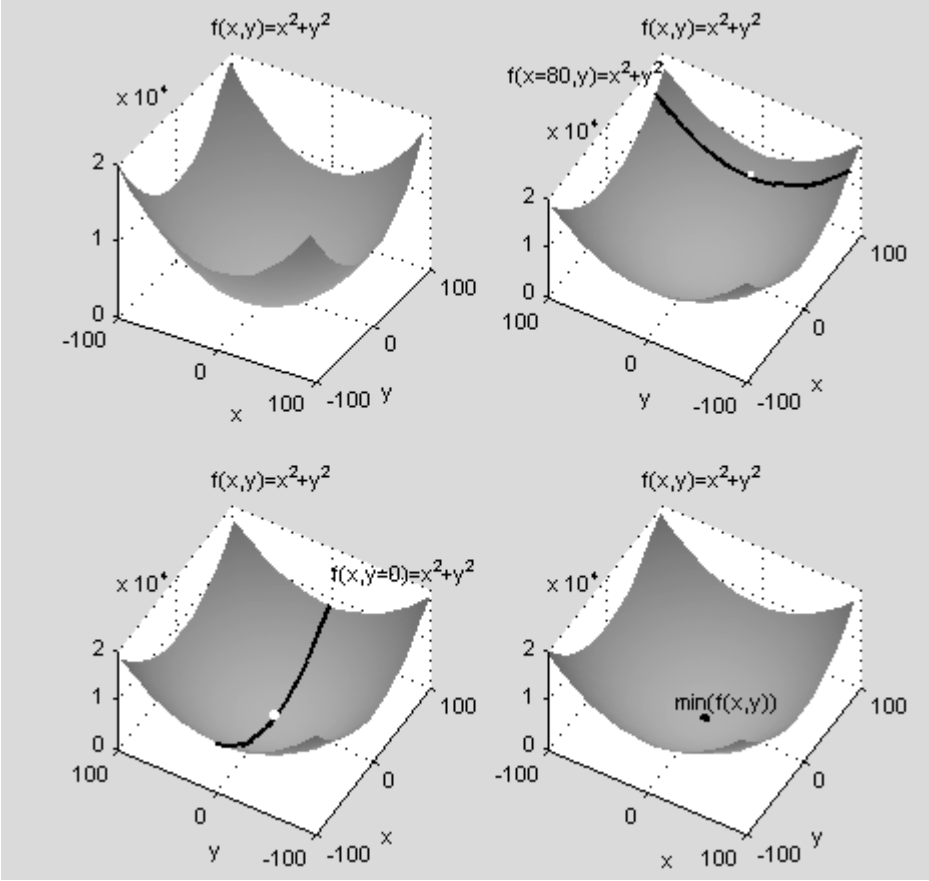


Figura 2.1 – Processo de otimização. O primeiro passo é mostrado no topo da figura, o segundo, abaixo na figura à esquerda e o último, à direita na parte de baixo da figura.

No primeiro passo deste processo, a variável *x* é mantida constante, neste caso, com o valor igual a 80 (*.i.e.* *x*=80). A derivada desta função em relação a *y*, quando *x*=80, é calculada e seu ponto de mínimo determinado. O valor onde *y* é mínimo, quando *x*=80, é usado no próximo passo, pois *y* é mantido constante. Neste último passo, a derivada de *f* em relação a *x* é calculada, *y*=0, e o ponto de mínimo determinado. Este processo segue, alternando entre *x* e *y*, até que a solução seja estável. No exemplo da Figura 2.1, o mínimo global é determinado em dois passos.

Assim, o QMA para a otimização da função dada na expressão **39** tem como primeiro passo a determinação da matriz **A**, onde a "direção" a ser fixada é dada pela expressão **41** (se for o primeiro passo do QMA, as matrizes **B**, **C** e **G** devem ser iniciadas seja com valores aleatórios ou por algum

outro critério) e solução para o problema descrito na expressão **39** é dada na expressão **42**.(*Ver* Apêndice 2.1 para breve introdução ao problema de Quadrados Mínimos)

$$\mathbf{Z} = \underline{\mathbf{G}}\left(\mathbf{C}^T \otimes \mathbf{B}^T\right) \tag{41}$$

$$\mathbf{A} = \underline{\mathbf{X}}\mathbf{Z}\left(\mathbf{Z}\mathbf{Z}^T\right)^{-1} \tag{42}$$

No passo seguinte a matriz **B** é calculada e para tal, a matriz $\underline{\mathbf{X}}$ é remodelada, como mostrado na expressão **43**, e a matriz a ser mantida fixa é mostrada na expressão **45**. A solução deste passo é descrita na expressão **46**.

$$\underline{\mathbf{X}}_B = \left(\mathbf{X}_1^T \quad | \quad \mathbf{X}_2^T\right), \tag{43}$$

onde

$$\underline{\mathbf{X}}_B = \mathbf{B}\underline{\mathbf{G}}\left(\mathbf{C}^T \otimes \mathbf{A}^T\right). \tag{44}$$

$$\mathbf{Z}_B = \underline{\mathbf{G}}\left(\mathbf{C}^T \otimes \mathbf{A}^T\right) \tag{45}$$

$$\mathbf{B} = \underline{\mathbf{X}}\mathbf{Z}_B\left(\mathbf{Z}_B\mathbf{Z}_B^T\right)^{-1} \tag{46}$$

Para o cálculo da matriz **C**, a matriz $\underline{\mathbf{X}}$ também é remodelada, como mostrado na expressão **47**, e a matriz a ser mantida fixa é dada na expressão **48**. A solução deste passo é descrita na expressão **49**.

$$vec\underline{\mathbf{X}} = \left(\mathbf{A} \otimes \mathbf{B}\right)vec\underline{\mathbf{G}}\mathbf{C}^T \tag{47}$$

$$\mathbf{Z}_C = (\mathbf{A} \otimes \mathbf{B}) vec\underline{\mathbf{G}} \quad (48)$$

$$\mathbf{C} = \left[\left(\mathbf{Z}_C^T \mathbf{Z}_C \right)^{-1} \mathbf{Z}_C^T vec\underline{\mathbf{X}} \right]^T \quad (49)$$

No último passo, para o cálculo da matriz **G**, a expressão usada no cálculo de **C** é empregada, como mostrado na expressão **47**, e as matrizes a serem mantidas fixas são a **C** e a dada em **50**. A solução deste passo é dada em **51**.

$$\mathbf{Z}_G = (\mathbf{A} \otimes \mathbf{B}) \quad (50)$$

$$vec\underline{\mathbf{G}} = \left(\mathbf{Z}_G^T \mathbf{Z}_G \right)^{-1} \mathbf{Z}_G^T vec\underline{\mathbf{X}} \mathbf{C} \left(\mathbf{C}^T \mathbf{C} \right)^{-1} \quad (51)$$

Neste último passo, caso **G** seja restrita a possuir alguns de seus elementos fixados com o valor zero, pode-se empregar a manipulação descrita no Apêndice 2.1.

Após o cálculo de todas a matrizes, **A**, **B**, **C** e **G**, o valor da função dada na expressão **39** é testado, caso a diferença entre o valor atual da função *l* e o de um passo anterior seja menor que determinado critério, o algoritmo convergiu, caso contrário, ele é reiniciado com os valores atuais das matrizes **A**, **B**, **C**, e **G**.

2.6 Validação

O processo de validação empregado neste trabalho de tese é baseado na amostragem do conjunto de dados total, ou seja, dividindo cada matriz original em novas matrizes a serem usadas na análise dos dados. Ao final todas as matrizes são analisadas e os resultados são comparados. A amostragem é efetuada de forma sistemática, um exemplo disto é descrito aqui a partir da matriz de dados **A** mostrada a seguir:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} & a_{17} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} & a_{27} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & a_{36} & a_{37} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} & a_{46} & a_{47} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} & a_{56} & a_{57} \\ a_{61} & a_{62} & a_{63} & a_{64} & a_{65} & a_{66} & a_{67} \end{pmatrix}$$

A primeira etapa neste caso, é eliminar as colunas da matriz **A**, por exemplo, 2,4 e 6, como mostrado a seguir:

$$\mathbf{A}_{col} = \begin{pmatrix} a_{11} & a_{13} & a_{15} & a_{17} \\ a_{21} & a_{23} & a_{25} & a_{27} \\ a_{31} & a_{33} & a_{35} & a_{37} \\ a_{41} & a_{43} & a_{45} & a_{47} \\ a_{51} & a_{53} & a_{55} & a_{57} \\ a_{61} & a_{63} & a_{65} & a_{67} \end{pmatrix}$$

A matriz resultante desta redução, $\mathbf{A}_{col}$, é então reduzida a matriz final, neste caso, $\mathbf{A}_{final}$ como se segue:

$$\mathbf{A}_{final} = \begin{pmatrix} a_{11} & a_{13} & a_{15} & a_{17} \\ a_{31} & a_{33} & a_{35} & a_{37} \\ a_{51} & a_{53} & a_{55} & a_{57} \\ a_{61} & a_{63} & a_{65} & a_{67} \end{pmatrix}$$

Esta matriz, $\mathbf{A}_{final}$, corresponde a uma das matrizes do conjunto de validação. As outras matrizes poderiam ser obtidas eliminando outras combinações de linhas e colunas.

2.7 Apêndice 2.1

2.7.1 Núcleo Restrito

Considere o caso onde **G** é restrita a possuir alguns de seus elementos fixados com o valor zero. Empregando a propriedade do operador *vec* dada pela expressão **52** *[Henderson]*.

$$vec(\mathbf{\Pi\Theta\Phi}) = (\mathbf{\Phi}^T \otimes \mathbf{\Pi})vec\mathbf{\Theta} \quad 52$$

e as matrizes dadas pelas expressões **53** e **54**.

$$\mathbf{Z} = (\mathbf{A} \otimes \mathbf{B}) \quad 53$$

$$\mathbf{W} = vec\underline{\mathbf{G}} \quad 54$$

a expressão **47** do texto principal pode ser reescrita em termos da expressão **55**.

$$vec\underline{\mathbf{X}} = \mathbf{ZWC}^T \quad 55$$

Aplicando o operador *vec* na expressão **55**, como mostrado na expressão **56**, e por conseguinte empregando a propriedade dada pela expressão **52**, tem-se a expressão **57**.

$$vec(vec\underline{\mathbf{X}}) = vec(\mathbf{ZWC}^T) \quad 56$$

$$vec(vec\underline{\mathbf{X}}) = (\mathbf{C} \otimes \mathbf{Z})vec\mathbf{W} \quad 57$$

Antes de continuar nesta formulação, é interessante relembrar um pouco sobre matrizes de permutação. Para tal, considere o exemplo da multiplicação dado na expressão **58**.

$$\begin{pmatrix} \alpha & \delta & \varphi \\ \beta & \varepsilon & \gamma \\ \chi & \phi & \eta \end{pmatrix} \times \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} a\alpha + b\delta + c\varphi \\ a\beta + b\varepsilon + c\gamma \\ a\chi + b\phi + c\eta \end{pmatrix} \quad 58$$

A permutação (troca da posição) dos elementos do vetor coluna $(a\ b\ c)^T$ pode ser obtida com o emprego de uma matriz de permutação como mostrado na expressão **59**.

$$\begin{pmatrix} \alpha & \delta & \varphi \\ \beta & \varepsilon & \gamma \\ \chi & \phi & \eta \end{pmatrix} \times \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \times \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \varphi & \alpha & \delta \\ \gamma & \beta & \varepsilon \\ \eta & \chi & \phi \end{pmatrix} \times \begin{pmatrix} c \\ a \\ b \end{pmatrix} \quad 59$$

onde:

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad 60$$

$$\begin{pmatrix} \varphi & \alpha & \delta \\ \gamma & \beta & \varepsilon \\ \eta & \chi & \phi \end{pmatrix} \times \begin{pmatrix} c \\ a \\ b \end{pmatrix} = \begin{pmatrix} c\varphi + a\alpha + b\delta \\ c\gamma + a\beta + b\varepsilon \\ c\eta + a\chi + b\phi \end{pmatrix} \quad 61$$

$$\begin{pmatrix} c\varphi + a\alpha + b\delta \\ c\gamma + a\beta + b\varepsilon \\ c\eta + a\chi + b\phi \end{pmatrix} = \begin{pmatrix} a\alpha + b\delta + c\varphi \\ a\beta + b\varepsilon + c\gamma \\ a\chi + b\phi + c\eta \end{pmatrix} \quad 62$$

De volta ao problema dado pela expressão **55** onde **W** possui alguns elementos que devem ser iguais a zero. Neste caso, com o emprego de um matriz de permutação e sua respectiva inversa, os elementos de *vec***W** podem ser permutados de forma a se obter a matriz dada pela expressão **64**, onde os elementos iguais a zero fiquem "separados" daqueles diferentes de zero.

$$vec(vec\underline{\mathbf{X}}) = (\mathbf{C} \otimes \mathbf{Z})\mathbf{P}^{-1}\mathbf{P}vec\mathbf{W} \quad 63$$

$$\mathbf{P}vec\mathbf{W} = \begin{pmatrix} \mathbf{\Delta}_1 \\ \mathbf{0} \end{pmatrix} \quad 64$$

Substituindo a expressão **65** em conjunto com a expressão **64** na expressão **63** têm-se a expressão **66**.

$$(\mathbf{C} \otimes \mathbf{Z})\mathbf{P}^{-1} = (\mathbf{\Psi}_1 \quad | \quad \mathbf{\Psi}_2) \quad 65$$

$$vec(vec\underline{\mathbf{X}}) = (\mathbf{\Psi}_1 \quad | \quad \mathbf{\Psi}_2) \times \begin{pmatrix} \mathbf{\Delta}_1 \\ \mathbf{0} \end{pmatrix} \quad 66$$

A determinação dos elementos diferentes de zero da matriz **W,** dados por $\mathbf{\Delta}_1$, é obtida com decomposição da expressão **66** na soma dada pela expressão **67** que tem como solução a expressão **68**. Finalmente, **G** pode ser obtida a partir de **W,** com o emprego da inversa da matriz de permutação $\mathbf{P}^{-1}$.

$$vec(vec\underline{\mathbf{X}}) = \mathbf{\Psi}_1\mathbf{\Delta}_1 + \mathbf{\Psi}_2\mathbf{0} \quad 67$$

$$\left(\mathbf{\Psi}_1{}^T\mathbf{\Psi}_1\right)^{-1}\mathbf{\Psi}_1{}^T vec(vec\underline{\mathbf{X}}) = \mathbf{\Delta}_1 \quad 68$$

2.7.2 Quadrados Mínimos

2.7.2.1 Interpretação Geométrica

Para mostrar a solução de um problema de quadrados mínimos, a sua interpretação geométrica é usada neste apêndice. Para tal, considere um problema onde os valores de uma certa propriedade são aproximados por duas medidas instrumentais. Os valores da propriedade, supostos verdadeiros, para cada uma das três amostras, são organizados em um vetor coluna, como mostrado na expressão **72**, enquanto que os valores das medidas instrumentais são organizados em uma matriz, como mostrado na expressão **69**. Um excelente livro que aborda este tema é o do Prof. Gilbert Strang *[Strang]* que também possui uma página na internet com vídeos de seu curso *[http(e)]*.

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 \end{pmatrix} \tag{69}$$

onde

$$\mathbf{x}_1 = \begin{pmatrix} x_{11} \\ x_{21} \\ x_{31} \end{pmatrix} \tag{70}$$

$$\mathbf{x}_2 = \begin{pmatrix} x_{12} \\ x_{22} \\ x_{32} \end{pmatrix} \tag{71}$$

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{21} \\ y_{31} \end{pmatrix} \tag{72}$$

O problema a ser tratado é a elaboração de um modelo linear para a determinação dos valores da propriedade de interesse, através de medidas instrumentais, em amostras futuras. Para tal construção, os coeficientes da combinação linear das medidas instrumentais, dispostos em um vetor coluna **b** em **73**, devem ser encontrados.

$$\mathbf{y} = \mathbf{Xb} \tag{73}$$

onde

$$\mathbf{b} = \begin{pmatrix} b_{11} \\ b_{21} \end{pmatrix} \tag{74}$$

Este problema colocado em termos de quadrados mínimos é apresentado na expressão **75**, onde os coeficientes de regressão, **b**, correspondentes ao ponto de mínimo da função *f*(**b**) devem ser determinados.

$$f(\mathbf{b}) = \|\mathbf{Xb} - \mathbf{y}\|^2 \tag{75}$$

sendo a função *f*(b) definida no plano real.

Em termos geométricos, este problema corresponde a representação do vetor de propriedades, **y**, no espaço descrito pelos vetores coluna da matriz de medidas instrumentais, **X**, onde o vetor diferença, entre **y** e aquele no espaço descrito pelos vetores coluna de **X**, deve possuir o menor módulo possível. A Figura 2.2 apresenta uma ilustração para este tipo de problema.

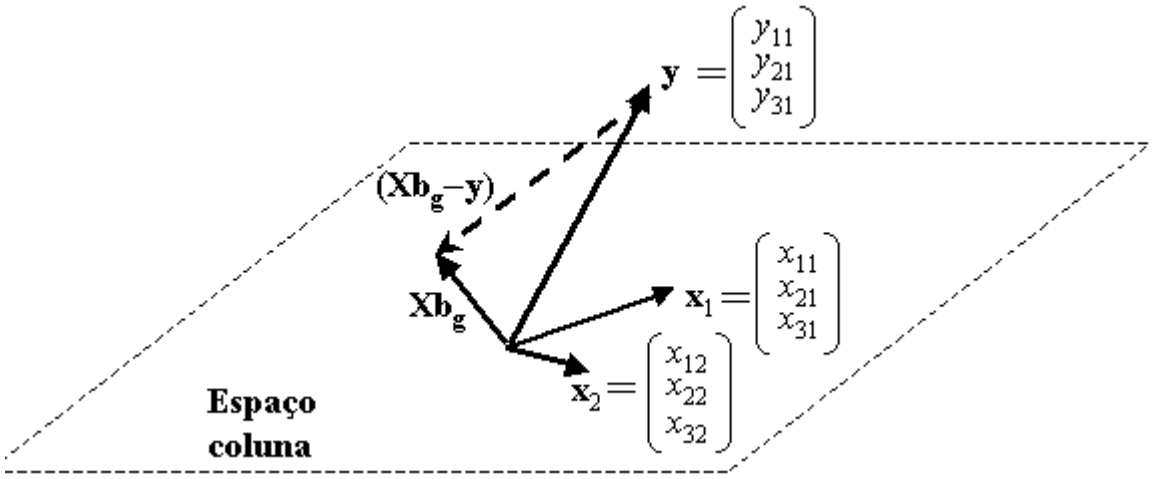


Figura 2.2 – Ilustração geométrica do problema de Quadrados Mínimos. $\mathbf{Xb_g}$ representa um vetor geral resultado da combinação linear dos vetores coluna de **X**.

A melhor solução para o problema de Quadrados Mínimos é aquela onde o vetor diferença, entre **y** e aquele no espaço descrito pelos vetores coluna de **X** neste caso **Xb**, é perpendicular ao plano dos vetores coluna da matriz **X** (ou espaço coluna de **X**), como mostrado na Figura 2.3.

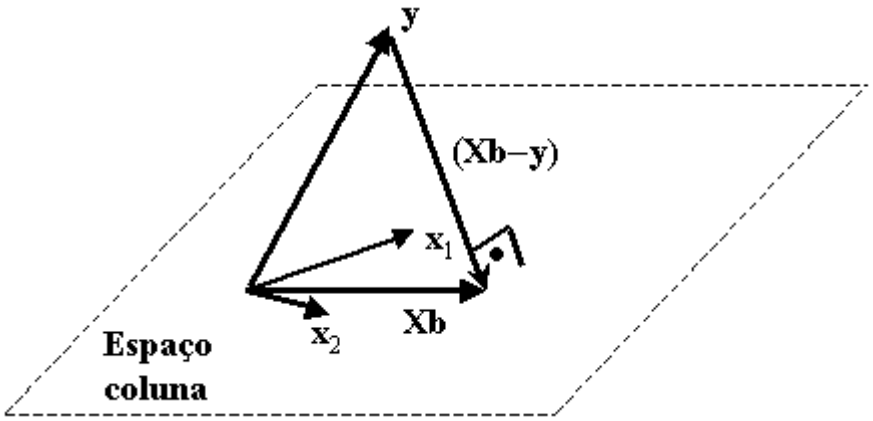


Figura 2.3 – Ilustração geométrica da solução do problema de Quadrados Mínimos.

Como mostrado na Figura 2.3 e argumentação acima, a melhor solução é aquela onde o vetor diferença é perpendicular ao plano dos vetores coluna da matriz **X**, argumento que em termos algébricos é dado pela expressão **76** (deve ser lembrado que o produto interno entre dois vetores perpendiculares é igual a zero).

$$\mathbf{X}^T(\mathbf{Xb}-\mathbf{y})=\begin{pmatrix}0\\0\end{pmatrix} \qquad 76$$

O devido desenvolvimento da expressão **76** mostra que a solução para o problema descrito em **75** é dado pela expressão **79** obtida através de **78**. É importante dizer que a solução mostrada na expressão **79** admite a matriz **X** ($n\times m$) como posto completo e que $n \geq m$.

$$\left(\mathbf{X}^T\mathbf{Xb}-\mathbf{X}^T\mathbf{y}\right)=\begin{pmatrix}0\\0\end{pmatrix} \qquad 77$$

$$\mathbf{X}^T\mathbf{Xb}=\mathbf{X}^T\mathbf{y} \qquad 78$$

$$\mathbf{b}=\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \qquad 79$$

Para casos onde várias propriedades são de interesse, o vetor **y** passa a ser representado por uma matriz **Y**, onde cada vetor coluna representa uma propriedade de interesse. Neste caso, a representação da matriz **Y** através das medidas instrumentais é descrita na expressão **80**.

$$\mathbf{Y} = \mathbf{XB} \qquad 80$$

Para este caso, a representação geométrica também é interessante, mas para sua compreensão é melhor reescrever a expressão **80**, através do operador de vetorização, como dado em **83**, obtida a partir da aplicação da propriedade dada na expressão **82** na expressão **81**, como mostrado na expressão **83**.

$$vec(\mathbf{Y}) = vec(\mathbf{XB}) \qquad 81$$

$$vec(\mathbf{\Omega\Theta}) = (\mathbf{I} \otimes \mathbf{\Omega})vec(\mathbf{\Theta}) \qquad 82$$

$$vec(\mathbf{Y}) = (\mathbf{I}_k \otimes \mathbf{X})vec(\mathbf{B}) \qquad 83$$

onde *k* é o número de propriedades de interesse.

A solução do problema de quadrados mínimos para casos onde várias propriedades de interesse estão envolvidas é mostrada na expressão **84**.

$$vec(\mathbf{B}) = \left(\mathbf{I}_k \otimes \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)vec(\mathbf{Y}) \qquad 84$$

pois

$$(\mathbf{\Theta} \otimes \mathbf{\Omega})(\mathbf{\Gamma} \otimes \mathbf{\Pi}) = (\mathbf{\Theta\Gamma} \otimes \mathbf{\Omega\Pi}) \qquad 85$$

e

$$\left(\mathbf{I}_k \otimes \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)(\mathbf{I}_k \otimes \mathbf{X}) = \left(\mathbf{I}_k\,\mathbf{I}_k \otimes \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\right) = (\mathbf{I}_k \otimes \mathbf{I}_n) = \mathbf{I}_{nk}, \qquad 86$$

sendo $\mathbf{I}_k$, $\mathbf{I}_n$, $\mathbf{I}_{nk}$ matrizes identidade de dimensões, *k*, *n*, e *nk*, respectivamente, onde *k* é o número de propriedades de interesse.

As expressões **83** e **84** mostram que nos casos onde o número de propriedades de interesse é maior que um, também é possível representar a solução a partir da determinação do vetor diferença entre propriedades de interesse e projeção, ou seja, da determinação do vetor diferença com o menor valor em módulo. Neste caso, também é possível identificar que a solução é única pois só deve existir um vetor diferença com o menor valor de módulo, isto é, aquele perpendicular ao espaço coluna de ($\mathbf{I}_k \otimes \mathbf{X}$) , *ver* expressão **83**, e conseqüentemente um vetor projeção.

Reescrevendo a expressão **84** a partir da propriedade dada na expressão **82**, como mostrado na expressão **87**, ou seja, a solução direta para a determinação de **B**, é dada pela expressão **88**.

$$vec(\mathbf{B}) = vec\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left[\mathbf{Y}\right]\right) \quad 87$$

$$\mathbf{B} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} \quad 88$$

A solução direta do problema de quadrados mínimos mostrada na expressão **88** também é única segundo a argumentação acima, fato de grande importância no processo de otimização do algoritmo Quadrados Mínimos Alternantes.

2.7.2.2 Interpretação diferencial

A determinação do vetor de coeficientes da regressão linear foi descrita primeiro através da interpretação geométrica. Nesta parte do apêndice é descrito o emprego de derivadas para a determinação dos coeficientes de regressão. Em resumo, será discutido aqui a solução do problema descrito na expressão 89 onde o vetor de coeficientes, **b**(=(b_{11} b_{21})), deve ser determinado, sendo o vetor **y** e a matriz **X** conhecidos. Neste caso, **b** é determinado através da minimização da função dada pela expressão **90**, ou seja, através de um problema de quadrados mínimos.

$$\begin{pmatrix} y_{11} \\ y_{21} \\ y_{31} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{pmatrix} \begin{pmatrix} b_{11} \\ b_{21} \end{pmatrix} \quad 89$$

$$f(\mathbf{b}) = \left\|\mathbf{y} - \mathbf{X}\mathbf{b}\right\|^2 \quad 90$$

A determinação do ponto de mínimo da função f(**b**)dada na expressão **90** pode ser efetuada através da obtenção do ponto onde a derivada da função f(**b**) se iguala a zero. Para tal, considere primeiro a função f(**b**) como função da variável b_{11}, primeiro elemento do vetor de coeficientes. Neste caso, a derivada de f(**b**) em relação à variável b_{11}, é dada pela expressão **94** obtida a partir de **91**, **92** e **93**. A mesma formulação pode ser empregada para a obtenção da derivada de f(**b**) em relação à variável b_{21}, dada em **95** *[Strang]*.

$$f(b_{11}) = \|\mathbf{y} - \mathbf{Xb}\|^2 = (y_{11} - x_{11}b_{11} - x_{12}b_{21})^2 + (y_{21} - x_{21}b_{11} - x_{22}b_{21})^2 + (y_{31} - x_{31}b_{11} - x_{32}b_{21})^2 \quad 91$$

$$\begin{aligned} \frac{\partial f}{\partial b_{11}} &= 2(y_{11} - x_{11}b_{11} - x_{12}b_{21})(-x_{11}) + \\ &+ 2(y_{21} - x_{21}b_{11} - x_{22}b_{21})(-x_{21}) + \\ &+ 2(y_{31} - x_{31}b_{11} - x_{32}b_{21})(-x_{31}) \end{aligned} \quad 92$$

$$\begin{aligned} \frac{\partial f}{\partial b_{11}} &= 2(-y_{11}x_{11} + x_{11}^2 b_{11} + x_{12}x_{11}b_{21}) + \\ &+ 2(-y_{21}x_{21} + x_{21}^2 b_{11} + x_{22}x_{21}b_{21}) + \\ &+ 2(-y_{31}x_{31} + x_{31}^2 b_{11} + x_{32}x_{31}b_{21}) \end{aligned} \quad 93$$

$$\begin{aligned} \frac{\partial f}{\partial b_{11}} &= -2(y_{11}x_{11} + y_{21}x_{21} + y_{31}x_{31}) + \\ &+ 2b_{11}(x_{11}^2 + x_{21}^2 + x_{31}^2) + \\ &+ 2b_{21}(x_{12}x_{11} + x_{22}x_{21} + x_{32}x_{31}) \end{aligned} \quad 94$$

$$\begin{aligned} \frac{\partial f}{\partial b_{21}} &= -2(y_{11}x_{12} + y_{21}x_{22} + y_{31}x_{32}) + \\ &+ 2b_{21}(x_{12}^2 + x_{22}^2 + x_{32}^2) + \\ &+ 2b_{11}(x_{12}x_{11} + x_{22}x_{21} + x_{32}x_{31}) \end{aligned} \quad 95$$

Reescrevendo as derivadas de *f*(**b**), em relação às variáveis b_{11} e b_{21}, em termos matriciais, obtém-se a expressão **98**, através de **96** e **97**.

$$\begin{pmatrix} \frac{\partial f}{\partial b_{11}} \\ \frac{\partial f}{\partial b_{21}} \end{pmatrix} = 2\left[\begin{pmatrix} x_{11}^2 + x_{21}^2 + x_{31}^2 & x_{12}x_{11} + x_{22}x_{21} + x_{32}x_{31} \\ x_{12}x_{11} + x_{22}x_{21} + x_{32}x_{31} & x_{12}^2 + x_{22}^2 + x_{32}^2 \end{pmatrix} \begin{pmatrix} b_{11} \\ b_{21} \end{pmatrix} - \begin{pmatrix} x_{11} & x_{21} & x_{31} \\ x_{12} & x_{22} & x_{32} \end{pmatrix} \begin{pmatrix} y_{11} \\ y_{21} \\ y_{31} \end{pmatrix} \right] \quad 96$$

$$\begin{pmatrix} x_{11} & x_{21} & x_{31} \\ x_{12} & x_{22} & x_{32} \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{pmatrix} = \begin{pmatrix} x_{11}^2 + x_{21}^2 + x_{31}^2 & x_{12}x_{11} + x_{22}x_{21} + x_{32}x_{31} \\ x_{12}x_{11} + x_{22}x_{21} + x_{32}x_{31} & x_{12}^2 + x_{22}^2 + x_{32}^2 \end{pmatrix} = \mathbf{X}^T\mathbf{X} \quad 97$$

$$\begin{pmatrix} \frac{\partial f}{\partial b_{11}} \\ \frac{\partial f}{\partial b_{21}} \end{pmatrix} = 2\left[\mathbf{X}^T\mathbf{X}\mathbf{b} - \mathbf{X}^T\mathbf{y}\right] \quad 98$$

O mínimo da função $f(\mathbf{b})$ ocorre onde as derivadas desta função, em relação às variáveis b_{11} e b_{21}, se igualam a zero simultaneamente, o que em termos algébricos é dado pela expressão **99**.

$$\begin{pmatrix} \frac{\partial f}{\partial b_{11}} \\ \frac{\partial f}{\partial b_{21}} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad 99$$

A aplicação desta condição mostrada na expressão **99** em **98** resulta na solução do problema de quadrados mínimos como mostra a expressão **102** obtida a partir das expressões **100** e **101**.

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = 2\left[\mathbf{X}^T\mathbf{X}\mathbf{b} - \mathbf{X}^T\mathbf{y}\right] \quad 100$$

$$\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y} \quad 101$$

$$\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \quad 102$$

É importante relembrar que a solução descrita em **102** admite a matriz $\mathbf{X}$ ($n \times m$) como posto completo e que $n \geq m$.

De forma semelhante àquela descrita na interpretação geométrica do problema de quadrados mínimos a solução dada em **102** também pode ser estendida aos casos onde $\mathbf{y}$ é substituída por uma matriz $\mathbf{Y}$ *[Strang]*.

3 Calibração de Segunda Ordem em Problema com Sobreposição de Posto

3.1 Introdução

O desenvolvimento de instrumentos hifenados, também chamados de segunda ordem *[Hirschfeld]*, como *"LC-UV"* , *"GC-MS"* e *"MS-MS"* tem proporcionado diversas vantagens para o químico analítico no que se refere, principalmente, à identificação e quantificação de compostos em sistemas de misturas. Os dados gerados por este tipo de instrumento impulsionaram o desenvolvimento de métodos quimiométricos denominados métodos de segunda ordem *[Booksh]*. Embora estes métodos tenham como proposta, em geral, a quantificação de uma analito conhecido na presença de um interferente, existem diferenças significativas entre seus algoritmos reportados na literatura. Por exemplo, em termos do tipo de dados nos quais tem sido aplicados, alguns tendo sido usados para a solução de problemas em dados de *posto-um* (também conhecido como de complexidade-um ou bilinear) e outros usados para resolver problemas em dados de *posto-maior-que-um* (também conhecidos como complexidade misturada ou não-bilinear). Em termos de suas formulações algébricas, alguns destes métodos empregam uma aproximação por quadrados mínimos e outros são baseados na solução de um problema de autovalores-autovetores. Dentre aqueles baseados em uma aproximação por quadrados mínimos estão o *Residual Bilinearization (RBL) [Öhman]*, PARAFAC *[Bro (b), Harshman]* e os Modelos Tucker Restrito *[Smilde (a)]*, e dentre os baseados em um problema de autovalores-autovetores estão o *Rank Annihilation Factor Analysis (RAFA) [Ho]*, o *Generalized Rank Annihilation Method (GRAM) [Sanchez]* e o *Non-Bilinear Rank Annihilation (NBRA) [Wilson, Wang(a)]*.

Instrumentos de segunda ordem podem produzir dados de várias complexidades dependendo da natureza das técnicas analíticas empregadas *[Kiers(c)]*. Um fenômeno particular é a sobreposição de posto (deficiência no posto), onde o posto total da matriz dos dados não é igual à soma dos valores de posto das matrizes dos dados das espécies em separado, *i.e.* $\mathbf{X} = \mathbf{A} + \mathbf{B}$ mas o posto($\mathbf{X}$) < posto($\mathbf{A}$) + posto($\mathbf{B}$). Exemplos deste tipo de dados ocorrem em análise por injeção em fluxo na presença de um gradiente de pH *[Nørgaard]*; análise por injeção em fluxo onde uma cinética de decomposição ocorre *[Saurina]*; no emprego de sensores químicos, baseados na reação de Fujiwara, para a determinação de moléculas halogenadas *[Smilde (b)]*.

A proposta desta parte do trabalho de tese é avaliar a eficiência de dois métodos de calibração de segunda ordem, um baseado em solução de um problema de autovalores-autovetores e outro por uma aproximação por quadrados mínimos, isto em termos de sua habilidade e estabilidade quando empregados em dados com sobreposição de posto. O problema de sobreposição de posto foi escolhido, por ser este tipo de dados encontrado na prática e representar um ambiente não ideal (*i.e.* onde as variações experimentais afetam as suposições aceitas como verdadeiras pelos métodos), que, em geral, não pode ser evitado. Em certos casos, a sobreposição de posto é intrínseca da técnica usada, por exemplo, o emprego de solutos semelhantes em análise por injeção em fluxo resulta em propriedades semelhantes de dispersão (difusão) e, como resultado, perfis de eluição praticamente iguais para os solutos. A sobreposição de posto também pode ocorrer quando os componentes apresentam espectros semelhantes ou até mesmo iguais, um exemplo disto é o emprego de sensores químicos para acompanhar cinéticas de reação com caminhos distintos mas que resultam no mesmo produto final, ou seja, um mesmo espectro, o do produto final, mas para diferentes perfis de reação. Na presença de baixa reprodutibilidade e/ou uma baixa razão sinal ruído, uma alta colinearidade entre perfis também pode gerar sobreposição de posto.

Os métodos escolhidos para comparação são o RBL, baseado em uma aproximação por quadrados mínimos, e o NBRA, baseado em na solução de um problema de autovalores-autovetores. Estes métodos podem ser considerados como gerais, pois não empregam nenhuma informação intrínseca do conjunto de dados no processo de calibração. Uma comparação, através de dados simulados, foi efetuada por Wang *et alli [Wang]* que considerou estes dois métodos equivalentes, entretanto possuindo diferentes formas de propagação de erros. Os dados empregados nesta parte do trabalho correspondem à dados reais de análise por injeção em fluxo (FIA-*do inglês Flow Injection Analysis*) de isômeros de hidroxibenzadeídos *[Nørgaard]*, que na presença de um gradiente de pH são dissociados em suas formas ácida e básica.

Deve ser notado que resultados melhores podem ser obtidos através de métodos baseados em aproximação por quadrados mínimos que utilizam de vantagens como restrições (por exemplo, não-negatividade, unimodalidade) aplicadas aos perfis estimados, informações a respeito do analito (espectros e perfis de tempo encontrados em análises prévias) e técnicas que permitem a redução da influência da colinearidade na eficiência do método *[Kiers (f)]*. Estes métodos, no entanto, requerem algum conhecimento químico a respeito do sistema, o que nem sempre é possível.

3.2 Dados

O sistema da análise por injeção em fluxo (FIA) empregado neste estudo foi descrito na literatura *[Nørgaard]* e é mostrado esquematicamente na Figura 3.1. Tubos de Polipropileno (0,70cm de diâmetro interno) foram empregados. O fluido carregador empregado é uma solução tamponada Britton-Robinson com um pH de 4,5 e o fluido reagente uma solução tamponada Britton-Robinson com pH 11,4.

Figura 3.1- Esquema de um sistema de análise por injeção em fluxo. C = fluido carregador (tampão Britton-Robinson, pH = 4,5); A = Amostra (77µl); e R = Reagente (770µl, tampão Britton-Robinson, pH = 11,4).

A amostra foi injetada por uma autobureta ABU 80 (0,375ml min^{-1}) entre o carregador e o reagente como mostrado na Figura 3.1. Como o volume da amostra é pequeno (77µL) se comparado com o do carregador e do reagente (770µL), um suave gradiente de pH é criado no decorrer da análise devido à difusão do carregador (pH baixo) sobre o reagente (pH alto) .

A amostra é medida em uma célula de fluxo de 8µL e a medida feita através de espectrofotômetro com um arranjo de fotodíodos. As medidas das amostras são feitas durante 88 segundos com um segundo de intervalo, após 20 segundos do início da injeção e na faixa espectral de 254 a 450nm com intervalos de 2nm. Os dados de segunda ordem obtidos para cada amostra possuem dimensões de 89×99, por exemplo, a resposta do soluto 2-HBA, *ver* descrição a seguir, é mostrado na Figura 3.2. Este dados estão disponíveis em *[http(b)]*

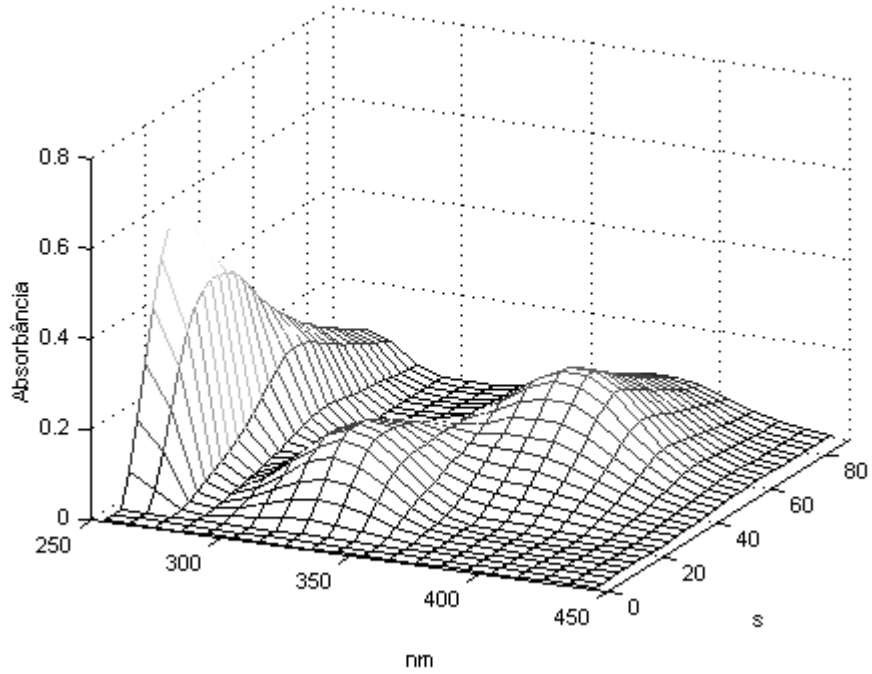


Figure 3.2- Exemplo de um conjunto de dados de uma análise por injeção em fluxo (padrão 2-HBA).

A duração da detecção e o gradiente de pH são suficientes para que ambas as formas, ácida e básica, sejam identificadas ao longo do experimento. Soluções de etanol-água foram usadas na preparação do carregador e padrões de forma que as soluções finais estivessem na relação 1:9 etanol-água(v/v). O tampão Britton-Robinson contendo ácido cítrico, bifosfato de potássio, ácido bórico, e tri-(hidroximetila) amino-metano(TRIS), segundo Perrin e Dempsey *[Perrin]*. TRIS foi usado para prevenir possíveis absorções, do tampão, na região do ultravioleta. A concentração do tampão foi 1,788mM e o pH da solução do reagente foi ajustada com hidróxido de sódio.

As soluções teste com os solutos 2-hidroxibenzadeído(2-HBA), 3- hidroxibenzadeído (3-HBA) e 4-hidroxibenzadeído (4-HBA) apresentam diferentes espectros de absorção dependendo da forma em que se encontram, ou seja, se na forma ácida ou na básica. A Figura 3.3 mostra os espectros de absorção de 2-HBA, 3-HBA e 4-HBA para suas formas ácida e básica.

Teoricamente não há separação dos constituintes de uma amostra, uma vez que o FIA é um sistema de transporte e não um sistema cromatográfico. A forma do perfil de concentração de um soluto específico é a mesma para a amostra, entretanto devido ao gradiente de pH, a primeira parte da amostra é dominada pela forma protonada e a última parte pela forma desprotonada. Dependendo do

pK_a de um dado soluto, este aparece com diferentes perfis de concentração para as formas ácida e básica no zona amostral, como mostrado na Figura 3.4, para os 3 solutos. Os valores do pK_a de 2-HBA, 3-HBA e 4-HBA são: 8,37; 8,98 e 7,61, respectivamente *[Serjeant]*.

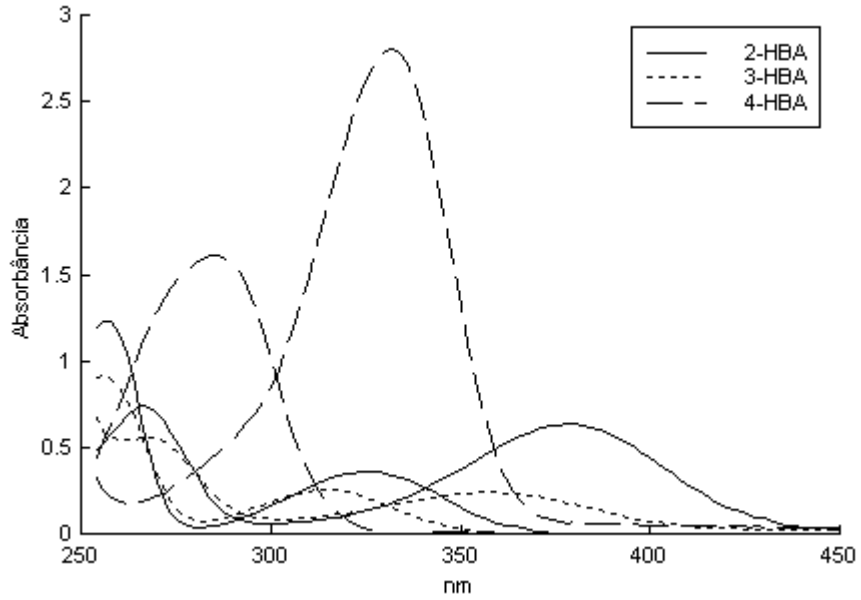


Figure 3.3- Espectros de 2-HBA, 3-HBA e 4-HBA em suas formas ácida e básica.

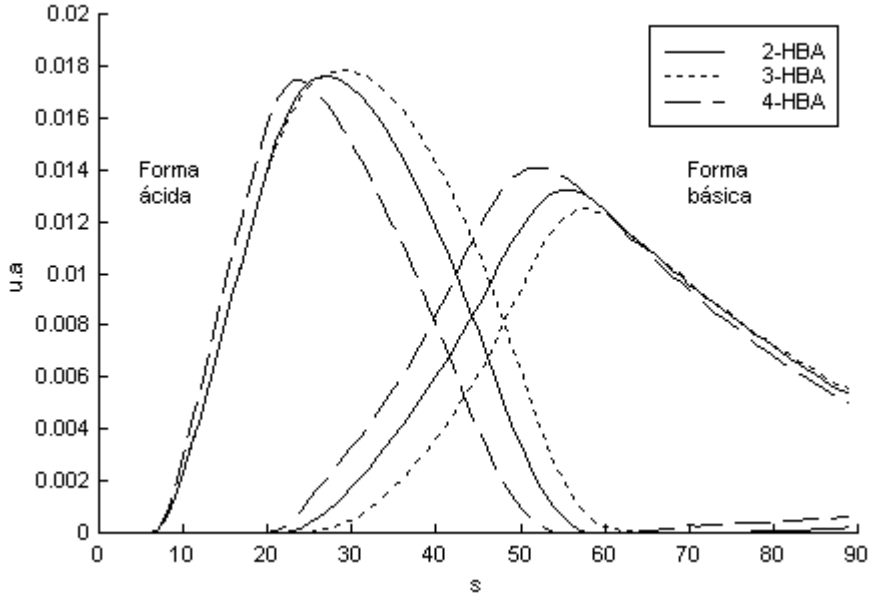


Figure 3.4 Perfis de tempo para a forma ácida (à esquerda) e básica para 2-HBA, 3-HBA e 4-HBA.

Os espectros das formas ácidas de 2-HBA, 3-HBA e 4-HBA são descritos como: $\mathbf{sa_2}$, $\mathbf{sa_3}$ e $\mathbf{sa_4}$ respectivamente e os espectros das formas básicas são descritos como: $\mathbf{sb_2}$, $\mathbf{sb_3}$ e $\mathbf{sb_4}$, respectivamente. O subscrito '2', '3' ou '4' se refere a 2-HBA, 3-HBA ou 4-HBA e a extensão 'a' ou 'b' às formas ácida e básica, respectivamente. Estes espectros são mostrados na Figura 3.3. De forma semelhante, os perfis de concentração das formas ácida de 2-HBA, 3-HBA e 4-HBA são descritos como: $\mathbf{ca_2}$, $\mathbf{ca_3}$ e $\mathbf{ca_4}$, e os perfis de concentração das formas básica como: $\mathbf{cb_2}$, $\mathbf{cb_3}$ e $\mathbf{cb_4}$. Estes perfis são mostrados na Figura 3.4. Estes espectros e perfis de concentração se referem à concentração unitária dos solutos, a não ser que estabelecido de forma diferente.

As respostas medidas dos solutos puros 2-HBA, 3-HBA e 4-HBA são descritas como $\mathbf{N_{2\text{-}HBA}}$, $\mathbf{N_{3\text{-}HBA}}$ e $\mathbf{N_{4\text{-}HBA}}$, respectivamente. Estas matrizes possuem dimensões de 89×99, ou seja, 89 tempos e 99 comprimentos de onda. Assumindo que a Lei de Beer *[Atkins]* é válida, as medidas, respostas do soluto puro de 2-HBA em concentração unitária, por exemplo, podem ser escritas como se segue:

$$\mathbf{N}_{2-\mathbf{HBA}} = \mathbf{ca}_2 \cdot \mathbf{sa}_2^{\mathrm{T}} + \mathbf{cb}_2 \cdot \mathbf{sb}_2^{\mathrm{T}} + \mathbf{E}_{\mathbf{2\text{-}HBA}} \qquad 103$$

onde $\mathbf{E_{2\text{-}HBA}}$ é o erro experimental das medidas.

3.3 Teoria

3.3.1 A calibração

O objetivo desta calibração é quantificar um analito de interesse na presença de um ou mais interferentes. Neste trabalho, apenas a quantificação de um analito na presença de um interferente é estudada, ou seja, apenas misturas binárias serão avaliadas. Para tal, duas matrizes são empregadas: o padrão, que contém as informações sobre o analito de interesse (em concentração conhecida) , e uma mistura binária, analito de interesse em concentração desconhecida (a ser determinada pelo método de calibração) e um interferente.

3.3.2 A sobreposição de posto

No decorrer da análise, o perfil de concentração total do composto (forma ácida mais básica) é dado por **ca**+**cb**. Isto resulta em três perfis de concentração total: $\mathbf{ctot_2}(=\mathbf{ca_2}+\mathbf{cb_2})$, $\mathbf{ctot_3}(=\mathbf{ca_3}+\mathbf{cb_3})$ e $\mathbf{ctot_4}(=\mathbf{ca_4}+\mathbf{cb_4})$ para 2-HBA, 3-HBA e 4-HBA, respectivamente. Estes perfis são mostrados na Figura 3.5. A forma destes perfis é determinada pelo processo de difusão que ocorre no FIA *[Buguera]*, como neste caso, os três solutos possuem, praticamente, a mesma estrutura tridimensional e conseqüentemente o mesmo valor para do coeficiente de difusão *[Nørgaard]*, seus perfis de concentração total são praticamente os mesmos (*i.e.* $\mathbf{ctot_2} = \alpha \cdot \mathbf{ctot}_3 = \beta \cdot \mathbf{ctot}_4$, onde α e β são constantes). Este fenômeno restringe a calibração e resulta assim, em uma sobreposição de posto.

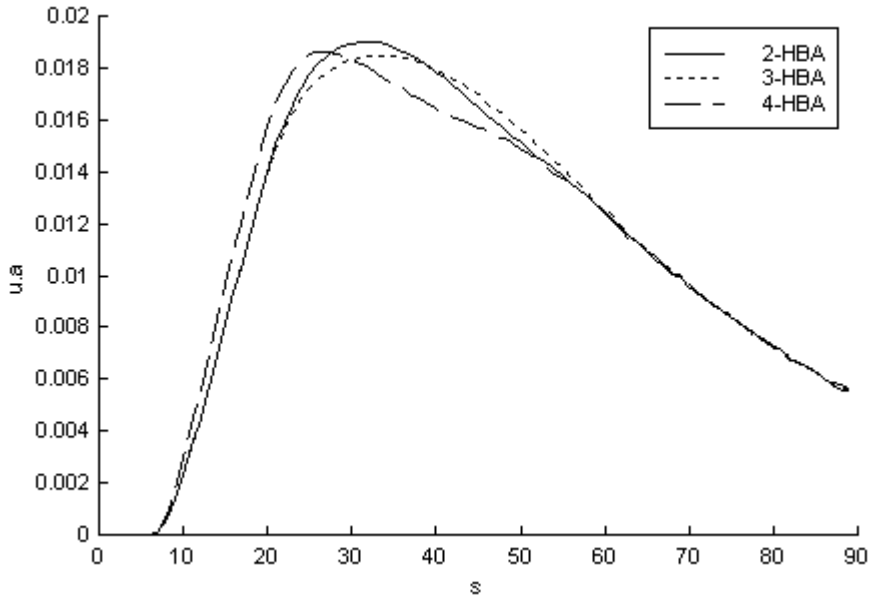


Figure 3.5- Perfis de concentração total para 2-HBA, 3-HBA e 4-HBA.

Para explicar o modelo de calibração e descrever algebricamente o problema de sobreposição de posto, é empregado como exemplo a calibração de 2-HBA, o analito de interesse, na presença de um interferente desconhecido (*e.g.* 3-HBA ou 4-HBA). A mistura pode ser descrita como mostrado a seguir:

$$\begin{aligned}\mathbf{M} &= \gamma \cdot \mathbf{ca}_2 \cdot \mathbf{sa}_2^{\mathrm{T}} + \gamma \cdot \mathbf{cb}_2 \cdot \mathbf{sb}_2^{\mathrm{T}} + \mathbf{ca}_\mathbf{u} \cdot \mathbf{sa}_\mathbf{u}^{\mathrm{T}} + \mathbf{cb}_\mathbf{u} \cdot \mathbf{sb}_\mathbf{u}^{\mathrm{T}} + \mathbf{E}_\mathbf{M} \\ &= \hat{\mathbf{M}} + \mathbf{E}_\mathbf{M}\end{aligned} \qquad 104$$

onde $\mathbf{M}$ é a matriz da resposta experimental na presença de ruídos, $\hat{\mathbf{M}}$ é a matriz da resposta experimental para o caso ideal sem erros de medidas, $\mathbf{E_M}$ é a matriz de ruídos, γ é a concentração do

analito na mistura e o subscrito 'u' indica o interferente desconhecido. A concentração é absorvida em $\mathbf{ca_u}$ **e** $\mathbf{cb}_u$ sem perda de generalidade.

Considerando que o perfil de concentração total do analito 2-HBA e do interferente desconhecido possuem a mesma forma, como colocado na formulação algébrica dada nas expressões **105** e **106**,

$$\mathbf{ctot}_2 = \mathbf{ca}_2 + \mathbf{cb}_2 \text{, e} \tag{105}$$

$$\alpha \cdot \mathbf{ctot}_2 = \mathbf{ca}_\mathbf{u} + \mathbf{cb}_\mathbf{u}, \tag{106}$$

onde α é um escalar, então as respostas do padrão e da mistura como descritos nas expressões **103** e **104** podem ser reescritas com a eliminação de $\mathbf{ca_2}$ e $\mathbf{cb_u}$ através das expressões **105** e **106** como se segue:

$$\begin{aligned} \mathbf{N} &= (\mathbf{ctot}_2 - \mathbf{cb}_2)\mathbf{sa}_2^T + \mathbf{cb}_2\mathbf{sb}_2^T + \mathbf{E}_\mathbf{N} \\ &= \mathbf{ctot}_2\mathbf{sa}_2^T - \mathbf{cb}_2\mathbf{sa}_2^T + \mathbf{cb}_2\mathbf{sb}_2^T + \mathbf{E}_\mathbf{N} \\ &= \mathbf{ctot}_2\mathbf{sa}_2^T + \mathbf{cb}_2(\mathbf{sb}_2^T - \mathbf{sa}_2^T) + \mathbf{E}_\mathbf{N} \end{aligned} \tag{107}$$

$$\begin{aligned} \mathbf{M} &= \gamma \cdot \mathbf{ctot}_2\mathbf{sa}_2^T + \gamma \cdot \mathbf{cb}_2(\mathbf{sb}_2^T - \mathbf{sa}_2^T) + \alpha \cdot \mathbf{ctot}_2\mathbf{sb}_\mathbf{u}^T + \mathbf{ca}_\mathbf{u}(\mathbf{sa}_\mathbf{u}^T - \mathbf{sb}_\mathbf{u}^T) + \mathbf{E}_\mathbf{M} \\ &= \mathbf{ctot}_2(\gamma \cdot \mathbf{sa}_2^T + \alpha \cdot \mathbf{sb}_\mathbf{u}^T) + \gamma \cdot \mathbf{cb}_2(\mathbf{sb}_2^T - \mathbf{sa}_2^T) + \mathbf{ca}_\mathbf{u}(\mathbf{sa}_\mathbf{u}^T - \mathbf{sb}_\mathbf{u}^T) + \mathbf{E}_\mathbf{M} \\ &= [\mathbf{ctot}_2 \mid \mathbf{cb}_2 \mid \mathbf{ca}_\mathbf{u}] \times \begin{bmatrix} \gamma \cdot \mathbf{sa}_2^T + \alpha \cdot \mathbf{sb}_\mathbf{u}^T \\ \gamma \cdot (\mathbf{sb}_2^T - \mathbf{sa}_2^T) \\ \mathbf{sa}_\mathbf{u}^T - \mathbf{sb}_\mathbf{u}^T \end{bmatrix} + \mathbf{E}_\mathbf{M} \end{aligned} \tag{108}$$

A expressão **108** mostra que o pseudo-posto de **M**, que corresponde ao posto de $\hat{\mathbf{M}}$ (posto(**M**) > pseudo-posto(**M**) devido ao ruído), é igual a três, pois **M** pode ser descrita em termos de uma matriz de posto três. **M** é dita como tendo uma sobreposição de posto um, pois, esta matriz é resultado da soma de duas outras de pseudo-posto 2 e seu pseudo-posto é 3. Neste caso, é dito também que a aditividade linear do sistema não é válida.

3.3.3 Métodos para calibração de segunda ordem

Os dois métodos usados para quantificar o analito de interesse na presença de um interferente foram escolhidos a partir de suas características algébricas, um baseado na solução de um problema de autovalores-autovetores (NBRA) e o outro em uma aproximação por quadrados mínimos (RBL). O

objetivo da avaliação destes dois métodos é verificar se estas características algébricas conferem estabilidade nas estimativas de seus modelos de calibração frente a um ambiente não ideal no qual os dados foram coletados (*i.e.* sobreposição de posto, colinearidade, reprodutibilidade variável, ruídos nos espectros). Para tal, os métodos devem tratar os dados de forma geral, ou seja, sem o emprego de informações sobre os dados na calibração. Neste sentido, a única informação necessária para o NBRA é que o posto da mistura seja conhecido e para o RBL que o posto do interferente seja conhecido. Naqueles casos onde o posto da mistura não é conhecido, técnicas de análise de posto podem ser usadas para a determinação deste valor de posto *[Faber]*. No caso do RBL, isto é, quando o posto do interferente não é conhecido, este valor pode ser determinado no próprio algoritmo deste método*[Öhman]*. Neste trabalho os valores de posto do interferente e da mistura são conhecidos e usados, pois o objetivo principal é avaliar as características dos métodos e não suas implementações.

3.3.3.1 Método de Anulação do Posto Não Bilinear-"*NBRA*"

O NBRA será explicado neste trabalho considerando o caso no qual o pseudo-posto da mistura é igual a 3 e do analito, igual a 2, sem perda de generalidade do método que pode ser aplicado a misturas com diferentes números de interferentes. Seguindo a representação geral para o problema de sobreposição de posto sugerida por Kiers e Smilde *[Kiers(c)]*, onde as expressões **107** e **108** podem ser colocadas em termos de **112** e **113** *via* **109**, **110**, e **111**.

$$\mathbf{X_r} = [\mathbf{ctot}_2 \mid \mathbf{cb}_2] \tag{109}$$

$$\mathbf{Y_r} = [\mathbf{sa}_2 \mid (\mathbf{sb}_2 - \mathbf{sa}_2)] \tag{110}$$

$$\mathbf{D_r} = \begin{bmatrix} \gamma & 0 \\ 0 & \gamma \end{bmatrix} \tag{111}$$

$$\mathbf{N} = \mathbf{X_r}\mathbf{Y_r}^{\mathrm{T}} + \mathbf{E_N} \tag{112}$$

$$\mathbf{M} = \mathbf{X_r}\mathbf{D_r}\mathbf{Y_r}^{\mathrm{T}} + \mathbf{X_s}\mathbf{Y_s}^{\mathrm{T}} + \mathbf{X_t}\mathbf{Y_t}^{\mathrm{T}} + \mathbf{E_M} \tag{113}$$

onde $\mathbf{X_r}$ ($I \times 2$) e $\mathbf{Y_r}$ ($J \times 2$) são os perfis para o analito; $\mathbf{D_r}$ corresponde à razão da concentração do analito na mistura com aquela encontrada no padrão; $\mathbf{X_s}$ ($I \times 1$) = $\mathbf{ctot_2}$ e $\mathbf{Y_s}$ ($J \times 1$) = $\alpha \cdot \mathbf{sb_u}$ descrevem os perfis do analito que apresentam sobreposição de posto com o interferente (a coluna de $\mathbf{X_s}$ é igual à

primeira coluna de $\mathbf{X_r}$); finalmente, $\mathbf{X_t}$ $(I \times 1)$ = $\mathbf{ca_u}$ e $\mathbf{Y_t}$ $(J \times 1)$ = $\mathbf{sa_u} - \mathbf{sb_u}$ representam os perfis para o analito que não possuem sobreposição de posto com o interferente.

O algoritmo do NBRA usado neste trabalho foi formulado em termos do GRAM descrito por Sanchez e Kowalski *[Sanchez]*, onde o seguinte problema de autovalores-autovetores é resolvido:

$$\overline{\mathbf{U}}^{\mathrm{T}}\mathbf{M}\overline{\mathbf{V}}\overline{\mathbf{S}}^{-1}\mathbf{Z} = \mathbf{Z}\boldsymbol{\Lambda} \qquad 114$$

onde $\overline{\mathbf{U}}$, $\overline{\mathbf{S}}$ e $\overline{\mathbf{V}}$, em sua forma truncada com 3 componentes principais (*ver* glossário), são o resultado da decomposição em valores singulares (SVD) de **W**, onde

$$\mathbf{W} = \mathbf{M} + \mathbf{N} \qquad 115$$

e

$$\overline{\mathbf{W}} = \overline{\mathbf{U}}\overline{\mathbf{S}}\overline{\mathbf{V}}^{\mathrm{T}} \qquad 116$$

A solução deste problema de autovalores-autovetores é mostrada no Apêndice 3.1. Neste caso, três autovalores são encontrados, sendo que, o menor é aquele empregado na determinação da concentração do analito na mistura (*ver* argumento no Apêndice 3.1).

Resumo do NBRA:

1. Resolva o problema de autovalores-autovetores dado na expressão **114** – Apenas o pseudo-posto de **M** deve ser conhecido.
2. Use o menor autovalor para encontrar γ, e então concentração do analito(*ver* expressão **142**,b).

3.3.3.2 RBL

O RBL é um método baseado em uma aproximação por quadrados mínimos que considera o resíduo (*i.e.* após a subtração do analito) como tendo uma estrutura bilinear. Neste caso, o resíduo corresponde à soma do sinal devido ao interferente(s) e o ruído experimental. O objetivo do RBL é minimizar a função perda descrita na expressão **117**.

$$E = \left\| \mathbf{M} - \gamma \mathbf{X_r}\mathbf{Y_r}^{\mathrm{T}} - \overline{\mathbf{P}}\,\overline{\mathbf{Q}}^{\mathrm{T}} \right\|^2 \qquad 117$$

onde $\gamma \mathbf{X_r}\mathbf{Y_r}^{\mathrm{T}}$ corresponde ao sinal do analito na concentração γ, $\overline{\mathbf{P}}$ e $\overline{\mathbf{Q}}$ são as matrizes de escores e *loadings* (*ver* glossário para definição de escores e *loadings*), respectivamente, resultantes da

decomposição em componentes principais de ($\mathbf{M} - \gamma \mathbf{X_r Y_r}^T$). O símbolo "$\bar{}$" indica que **P** e **Q** estão truncadas com o pseudo-posto do interferente, ou seja, só as primeiras colunas da matriz **P** e **Q** são usadas (o número de colunas é igual ao pseudo-posto do interferente). O posto do interferente, quando não conhecido a *priori* , pode ser avaliado aumentando o posto do interferente, a partir de um, até que a variância dos resíduos seja igual à do ruído experimental, como sugerido no artigo original deste método*[Öhman]*. Neste trabalho, o posto do interferente é conhecido, ou seja, igual a 2. O algoritmo empregado é dado no Apêndice 3.1 e converge quando os componentes principais **P** descrevem o espaço coluna do interferente e os componentes **Q** o espaço linha do interferente, *i.e.*

$$\mathbf{P}\ gera([\mathbf{X}_s\ \mathbf{X}_t]) \quad e \quad \mathbf{Q}\ gera([\mathbf{Y}_s\ \mathbf{Y}_t]) \qquad 118$$

3.3.3.3 Validação

A validação da calibração empregada nesta parte do trabalho de tese é baseada em um procedimento de reamostragem *[Efron]*. Neste caso, subconjuntos das matrizes originais 89×99 foram gerados sistematicamente, gerando 16 submatrizes (67×75). A primeira submatriz dos 16 subconjuntos corresponde aos comprimentos de onda 2,3,4,6,7,8,… (tendo sido os comprimentos de onda 1,5,9,… eliminados) e os tempos 2,3,4,6,7,8,… (tendo sido os tempos 1,5,9,… deixados fora). (*Ver seção Fundamentos para exemplo simplificado*). Para cada conjunto de comprimentos de onda escolhido, quatro novos sub conjuntos, de diferentes combinações de tempo, são gerados e validados. No total, 16 valores de concentração foram calculados para cada mistura calibrada, sendo a concentração final o valor médio daqueles 16.

3.4 Resultados da Calibração

Os resultados de calibração para as seis misturas formadas através da combinação dos três isômeros (2-HBA, 3-HBA e 4-HBA) são mostradas na Tabela 3.1 para o NBRA e na Tabela 3.2 para o RBL. Nestas tabelas o erro relativo (%) da calibração é calculado a través da seguinte fórmula: *erro%* = 100 × (c_{est}–c_r)/c_r, onde c_{est} é a concentração estimada e c_r a concentração real. Note que nestas tabelas, cada linha apresenta os resultados de duas calibrações. Por exemplo, a linha 1 para mistura 1 considera ambos os casos, onde 3-HBA é o analito e 4-HBA é o interferente e vice-versa.

Tabela 3.1 Calibração de 6 misturas binárias (unidade da concentração - μM) de 2-HBA, 3-HBA e 4-HBA usando o método NBRA.

No.	Mistura			Erro relativo %		
	2HBA	**3HBA**	**4HBA**	**2HBA**	**3HBA**	**4HBA**
1	0	100	40	-	-1,09 %	10,94 %
2	0	100	60	-	15,50 %	-3,83 %
3	50	0	60	-0,54 %	-	-5,97 %
4	100	0	60	2,22 %	-	-7,26 %
5	50	100	0	79,94 %	68,86 %	-
6	100	50	0	15,53 %	288,59 %	-

O valor negativo indica que a concentração predita é menor que o seu valor real.

Tabela 3.2 Calibração de 6 misturas binárias (unidade da concentração - μM) de 2-HBA, 3-HBA e 4-HBA usando o método RBL.

No.	Mistura			Erro relativo %		
	2HBA	**3HBA**	**4HBA**	**2HBA**	**3HBA**	**4HBA**
1	0	100	40	-	-0,87 %	13,32 %
2	0	100	60	-	17,39 %	-5,50 %
3	50	0	60	-13,09 %	-	7,84 %
4	100	0	60	0,86 %	-	12,47 %
5	50	100	0	20,29 %	19,44 %	-
6	100	50	0	13,55 %	74,54 %	-

O valor negativo indica que a concentração predita é menor que o seu valor real.

Cada estimativa corresponde ao valor médio de 16 concentrações determinadas na validação. O desvio padrão com relação à média destes 16 valores variou entre 0,05% e 0,87% do valor da concentração predita para todas as mistura exceto para uma, sendo 1,43% para a calibração de 3-HBA na presença de 2-HBA através do NBRA e 4,44% para mesma mistura quando o método empregado foi o RBL. Os baixos valores dos desvios padrão mostram que o ruído nos espectros apresenta pouca influência nos resultados, pois o resultado é praticamente o mesmo para diferentes conjuntos de pontos dos espectros.

Os resultados para a calibração de 4-HBA como analito na presença de 2-HBA ou 3-HBA são geralmente bons e similares para ambos os métodos. Para a calibração de 2-HBA ou 3-HBA, na

presença de 4-HBA como interferente, os resultados não se diferem muito entre os métodos, sendo bons também. Entretanto, para a calibração das misturas de 2-HBA e 3-HBA, os resultados são ruins sendo os do RBL um pouco melhores.

3.4.1 Análise exploratória

A proposta desta análise exploratória é investigar porque as misturas de 2-HBA e 3-HBA apresentam resultados ruins na calibração. Para tal, os perfis de tempo para as formas ácida e básica de 2-HBA e 3-HBA para ambos os padrões e mistura foram encontrados através de uma análise de componentes em dois modos (*Two Mode-Component Analysis [Magnus pag. 361]* (TMCA)) onde a sobreposição de posto também é considerada.

O princípio da TMCA é decompor a matriz **L** em três matrizes de posto completo, **W**, **K** e **Z**, como mostrado na Figura 3.6. Note que um caso especial de TMCA é a Decomposição em Valores Singulares (SVD), onde **W** e **Z** são matrizes ortogonais e **K** é uma matriz diagonal com elementos não negativos. Para os dados de FIA, a TMCA é descrita como sendo a decomposição da matriz com as respostas espectrais do experimento (tempo *versus* comprimentos de onda) em uma matriz com perfis tempos, uma matriz de concentrações e outra de perfis espectrais.

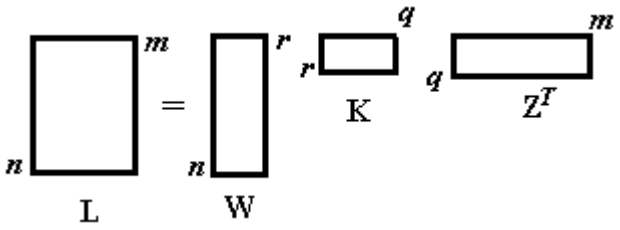


Figura 3.6- Análise de componentes em dois modos (*Two Mode Component Analysis* TMCA).

Os perfis de tempo do padrão, por exemplo 2-HBA, são encontrados através da solução do seguinte problema de quadrados mínimos:

$$\min \left\| \mathbf{N}_{\mathbf{2-HBA}} - \mathbf{T}_{\mathbf{2-HBA}} \mathbf{C}_{\mathbf{2-HBA}} \mathbf{Y}_{\mathbf{2-HBA}}^{\mathrm{T}} \right\|^2 \qquad 119$$

onde

$$\mathbf{T}_{\mathbf{2-HBA}} = \left[\mathbf{ca}_2 \mid \mathbf{cb}_2 \right] \qquad 120$$

$$\mathbf{C}_{\mathbf{2-HBA}} = \begin{bmatrix} \gamma_{2-\mathrm{HBA}} & 0 \\ 0 & \gamma_{2-\mathrm{HBA}} \end{bmatrix} \quad 121$$

sendo $\gamma_{2\text{-HBA}}$ a concentração atual de 2-HBA, e

$$\mathbf{Y}_{\mathbf{2-HBA}} = [\mathbf{sa}_2 \mid \mathbf{sb}_2] \quad 122$$

Para a mistura de, por exemplo, 2-HBA e 3-HBA, os perfis de tempo poderiam ser encontrados com a solução do seguinte problema:

$$\min \left\| \mathbf{M} - \mathbf{TCY}^{\mathrm{T}} \right\|^2 \quad 123$$

onde

$$\mathbf{T} = [\mathbf{ca}_2 \mid \mathbf{cb}_2 \mid \mathbf{ca}_3 \mid \mathbf{cb}_3] \quad 124$$

$$\mathbf{C} = \begin{bmatrix} \gamma_{2-\mathrm{HBA}} & 0 & 0 & 0 \\ 0 & \gamma_{2-\mathrm{HBA}} & 0 & 0 \\ 0 & 0 & \gamma_{3-\mathrm{HBA}} & 0 \\ 0 & 0 & 0 & \gamma_{3-\mathrm{HBA}} \end{bmatrix} \quad 125$$

e

$$\mathbf{Y} = [\mathbf{sa}_2 \mid \mathbf{sb}_2 \mid \mathbf{sa}_3 \mid \mathbf{sb}_3] \quad 126$$

Os espectros de absorção na região do ultravioleta dos compostos usados na TMCA, **Y**, foram obtidos em um experimento auxiliar efetuado para cada soluto nos diferentes valores de pH correspondentes às condições onde aparecem as formas ácida e básica no experimento *[Smilde (c)]*. Os valores das concentrações dos solutos em **C** também são conhecidos.

Através da TMCA, os perfis de tempo para 2-HBA e 3-HBA para ambos, padrão e mistura (número 6 nas Tabelas 3.1 e 3.2), foram calculados. Estes perfis são mostrados na Figura 3.7 em termos do perfil de concentração total (*i.e.* os perfis das formas ácida e básica somados). A forma das curvas encontradas para os perfis decompostos, para os padrões, concorda com a aquela esperada para um experimento FIA. Entretanto, as formas das curvas encontradas para a decomposição da mistura não apresentam significado físico, em termos do sistema FIA.

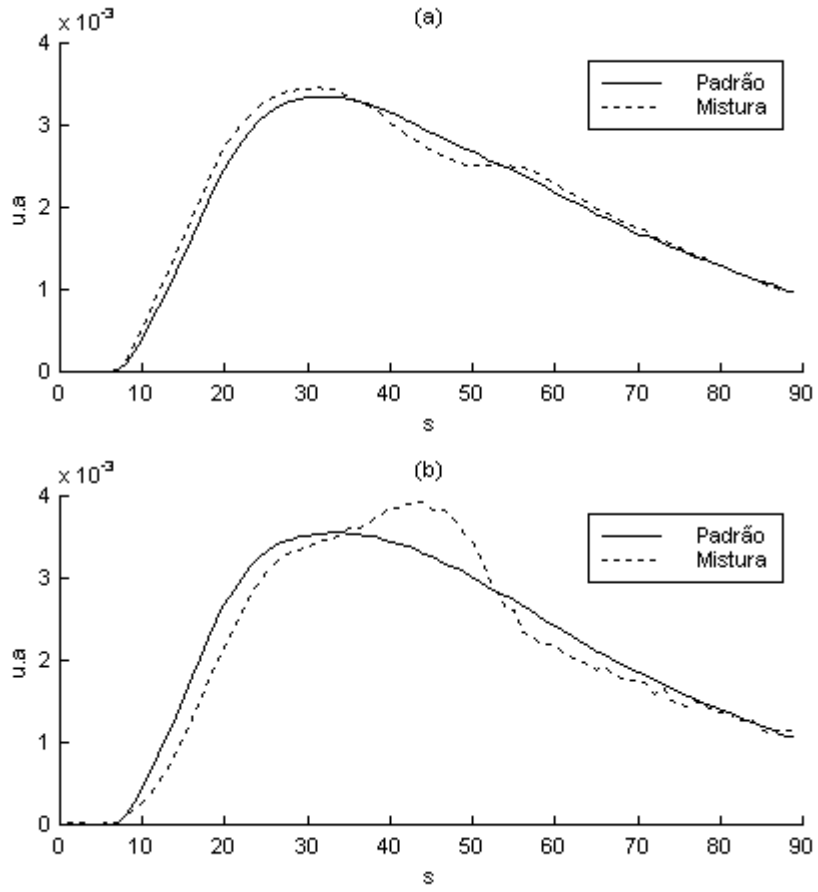


Figura 3.7- Perfis de concentração total para um padrão e uma mistura obtidos pela TMCA para (a) 2-HBA ($\mathbf{ca}_2$+$\mathbf{cb}_2$) e (b) 3-HBA ($\mathbf{ca}_3$+$\mathbf{cb}_3$).

A razão disto, ou seja, predição de uma curva que não possui significado físico, é devido ao fato de que a determinação da matriz **T** obtida resolvendo-se a expressão **123** resulta em uma modelagem de ruídos. A matriz **M** possui valor de posto maior que 3 devido ao ruído (posto(**M**) > pseudo-posto(**M**)). Empregando as matrizes **C** e **Y**, que são matrizes de posto completo (e possuem, então, posto 4) para resolver o problema de quadrado mínimos dados pela expressão **123** implica na estimativa de **T,** também, com posto igual 4. Entretanto, o valor correto para o posto de **T** é 3 (de forma semelhante ao pseudo-posto de **M** - *ver* expressão **108**), assim, a determinação de **T** com o valor de posto 4 resulta na modelagem de ruídos.

A forma correta de formular o modelo para TMCA deve levar em consideração a sobreposição de posto presente nos dados. Assim, a expressão **108** pode ser reformulada como se segue:

$$\mathbf{M} = \left[\mathbf{ctot}_2 \mid \mathbf{cb}_2 \mid \mathbf{ca}_\mathbf{u}\right] \times \begin{bmatrix} \gamma \cdot \mathbf{sa}_2^{\mathrm{T}} + \alpha \cdot \mathbf{sb}_\mathbf{u}^{\mathrm{T}} \\ \gamma \cdot \left(\mathbf{sb}_2^{\mathrm{T}} - \mathbf{sa}_2^{\mathrm{T}}\right) \\ \mathbf{sa}_\mathbf{u}^{\mathrm{T}} - \mathbf{sb}_\mathbf{u}^{\mathrm{T}} \end{bmatrix} + \mathbf{E}_\mathbf{M}$$

$$= \left[\mathbf{ctot}_2 \mid \mathbf{cb}_2 \mid \mathbf{ca}_\mathbf{u}\right] \times \begin{bmatrix} \gamma & 0 & 0 & \alpha \\ -\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \times \begin{bmatrix} \mathbf{sa}_2^{\mathrm{T}} \\ \mathbf{sb}_2^{\mathrm{T}} \\ \mathbf{sa}_\mathbf{u}^{\mathrm{T}} \\ \mathbf{sb}_\mathbf{u}^{\mathrm{T}} \end{bmatrix} + \mathbf{E}_\mathbf{M}$$

$$= \tilde{\mathbf{T}}\tilde{\mathbf{C}}\mathbf{Y} + \mathbf{E}_\mathbf{M} \qquad 127$$

Note-se que as matrizes **T** e **C** foram alteradas, como indicado por ' ~ '. Assim, $\tilde{\mathbf{T}}$ possui posto igual a 3, pois $\tilde{\mathbf{T}}$ possui 3 vetores colunas linearmente independentes, ao contrário de **T**, que possui 4 colunas linearmente independentes.

O modelo TMCA é estimado através dos espectros conhecidos, **Y**, e das concentrações também conhecidas, γ. Isto resulta em $\tilde{\mathbf{T}}$ e α. A Figura 3.8 mostra os novos perfis de concentração total, que são praticamente iguais quando calculados a partir do padrão ou da mistura. Isto mostra que a aditividade linear, segundo a Lei de Lambert-Beer, é válida para os dados.

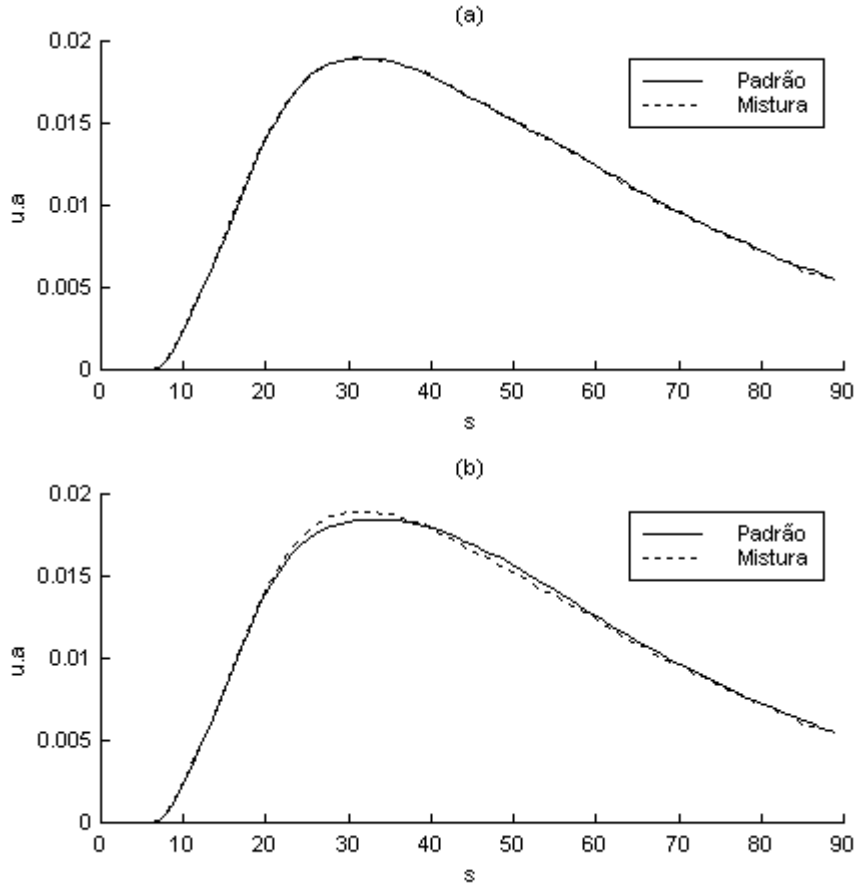


Figura 3.8- Perfil de concentração total para um padrão e uma mistura encontrados pela TMCA reformulada para 2-HBA (**ca**$_2$+**cb**$_2$) em (a) e 3-HBA (**ca**$_3$+**cb**$_3$) em (b).

3.5 Discussão

Os resultados da TMCA mostram que os perfis da forma básica de 2-HBA e 3-HBA na mistura apresentam grande colinearidade, como mostrado na Figura 3.9(b). O estudo da relação entre colinearidade entre os perfis de tempo das espécies dissociadas e das constantes de dissociação dos isômeros, considera a dissociação do isômero representada pela Reação (3.1) e as constantes de equilíbrio $k_{a\text{-}2}$ de 2-HBA e $k_{a\text{-}3}$ de 3-HBA dadas pelas expressões **128** e **129** *[Atkins]*.

$$AH + H_2O \overset{k_a}{\rightleftharpoons} A^- + H_3O^+$$ Reação (3.1)

$$k_{a-2} = \frac{\left[\mathrm{A}^-\right]_{2,T}\left[\mathrm{H_3O^+}\right]_T}{\left[\mathrm{AH}\right]_{2,T}} \quad 128$$

$$k_{a-3} = \frac{\left[\mathrm{A}^-\right]_{3,T}\left[\mathrm{H_3O^+}\right]_T}{\left[\mathrm{AH}\right]_{3,T}} \quad 129$$

onde $[A^-]_{2,T}$ e $[A^-]_{3,T}$ são as concentrações da forma básica de 2-HBA e 3-HBA medidas no tempo *t* (ou, alternativamente, em certo ponto do canal FIA) e $[AH]_{2,T}$ e $[AH]_{3,T}$ são as concentrações da forma ácida de 2-HBA e 3-HBA medidas no tempo *t*.

Usando o fato de que as concentrações totais de 2-HBA e 3-HBA são proporcionais,

$$ct_T = \left[\mathrm{A}^-\right]_{2,T} + \left[\mathrm{AH}\right]_{2,T} \quad 130$$

$$\alpha \cdot ct_T = \left[\mathrm{A}^-\right]_{3,T} + \left[\mathrm{AH}\right]_{3,T} \quad 131$$

onde ct_T é a concentração total medida no tempo *t* e α é um escalar constante, a relação entre os perfis da forma básica de 2-HBA e 3-HBA podem ser determinados como se segue:

$$\left[\mathrm{H_3O^+}\right]_T = \frac{k_{a-2}\left(ct_T - \left[\mathrm{A}^-\right]_{2,T}\right)}{\left[\mathrm{A}^-\right]_{2,T}}$$

$$\left[\mathrm{H_3O^+}\right]_T = \frac{k_{a-3}\left(\alpha \cdot ct_T - \left[\mathrm{A}^-\right]_{3,T}\right)}{\left[\mathrm{A}^-\right]_{3,T}}$$

$$\frac{k_{a-2}\left(ct_T - \left[\mathrm{A}^-\right]_{2,T}\right)}{\left[\mathrm{A}^-\right]_{2,T}} = \frac{k_{a-3}\left(\alpha \cdot ct_T - \left[\mathrm{A}^-\right]_{3,T}\right)}{\left[\mathrm{A}^-\right]_{3,T}}$$

$$\left[\mathrm{A}^{-}\right]_{3,T} = \frac{(\alpha \cdot ct_T)\cdot \left[\mathrm{A}^{-}\right]_{2,T} k_{a-3}}{k_{a-2} ct_T + \left[\mathrm{A}^{-}\right]_{2,T} (k_{a-3} - k_{a-2})} \qquad 132$$

A expressão **132** mostra que a diferença entre os perfis da forma básica dos isômeros depende dos valores das constantes de dissociação , $k_{a\text{-}2}$ e $k_{a\text{-}3}$. Naqueles casos onde os valores de k_a para os dois isômeros são muito parecidos, o termo '$k_{a\text{-}3} - k_{a\text{-}2}$' na expressão **132** é próximo de zero e os perfis de 2-HBA e 3-HBA se torna colineares (note que a mesma análise é válida para os perfis das formas ácida). Considerando os valores de k_a para os isômeros ($k_{a\text{-}2} = 4{,}27\times10^{-9}$, $k_{a\text{-}3} = 1{,}05\times10^{-9}$ e $k_{a\text{-}4} = 24{,}55\times10^{-9}$), é possível explicar a alta colinearidade entre os perfis de 2-HBA e 3-HBA na mistura, em outras palavras, as misturas de 2-HBA e 3-HBA são muito mais difíceis de se calibrar devido à alta colinearidade entre os perfis de tempo do analito e do interferente, o que é atribuído à similaridade entre os valores de k_a de 2-HBA e 3-HBA.

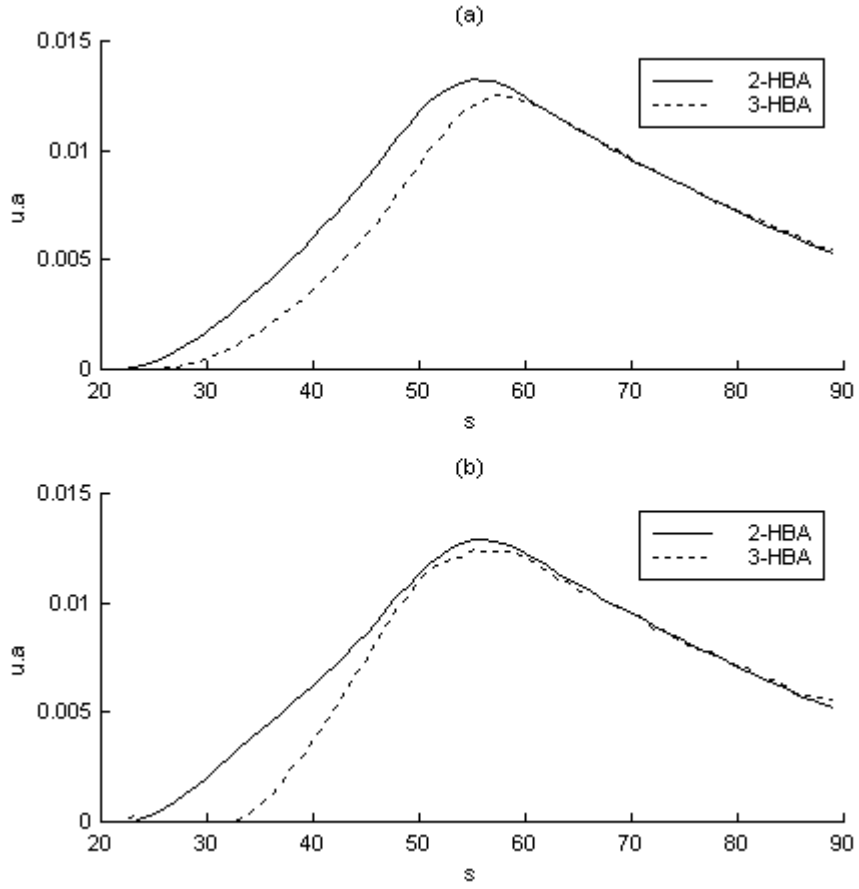


Figura 3.9- Resultados da TMCA para perfis de tempo para as formas básicas de 2-HBA e 3-HBA no padrão em (a) e em (b) para 2-HBA e 3-HBA na mistura.

A Figura 3.9(b) mostra um deslocamento, para a esquerda, da posição do perfil da forma básica do 3-HBA na mistura, isto tendo a posição do perfil, desta mesma forma e isômero, quando obtido a partir do padrão (compare *Figura* 3.9 (a) e (b)). Este deslocamento do perfil de tempo é explicado aqui através de um exemplo de dissociação mostrado na Reação (3.1), onde a concentração da espécie básica no tempo *t* é calculada em termos da constante de dissociação dada pela expressão **129**, a concentração total no tempo *t* dada pela expressão **131** e o pH no tempo *t* dado por

$$\mathrm{pH}_T = -\log_{10}\left(\left[\mathrm{H_3O^+}\right]_T\right) \tag{133}$$

A curva do gradiente de pH presente depende do comprimento da zona amostral no sistema FIA. As zonas amostrais para uma duplicata do experimento FIA para uma mesma amostra são mostradas na Figura 3.10, onde w_1 é o comprimento da zona amostral para a primeira duplicata e w_2 para a segunda ($w_1 > w_2$). Considerando que não é possível reproduzir o comprimento exato de uma zona amostral, a curva do gradiente de pH induzirá uma pequena variação entre as duas duplicatas como ilustrado na Figura 3.11. Esta variação do pH no tempo *t* para as duas duplicatas é representado por

$$\left[\mathrm{H_3O^+}\right]_{\mathrm{dupl-2},T} = \left[\mathrm{H_3O^+}\right]_{\mathrm{dupl-1},T} + \delta_T \tag{134}$$

onde os subscritos ‘dupl-1’ e ‘dupl-2’ indicam a primeira e a segunda duplicatas, respectivamente, e δ_T a diferença entre as concentrações de $[\mathrm{H_3O^+}]$ nas duplicatas no tempo *t*.

Novas expressões para a concentração no tempo *t* da forma 3-HBA para as duas duplicatas podem ser reescritas:

$$\left[\mathrm{A^-}\right]_{3,\mathrm{dupl-1},T} = \frac{\alpha \cdot ct_T}{10^{\mathrm{pK}_a - \mathrm{pH_T}} + 1} \tag{135}$$

$$\left[\mathrm{A^-}\right]_{3,\mathrm{dupl-2},T} = \frac{\alpha \cdot ct_T}{10^{\mathrm{pK}_a - \mathrm{pH_T}} + \delta_t 10^{\mathrm{pK}_a} + 1} \tag{136}$$

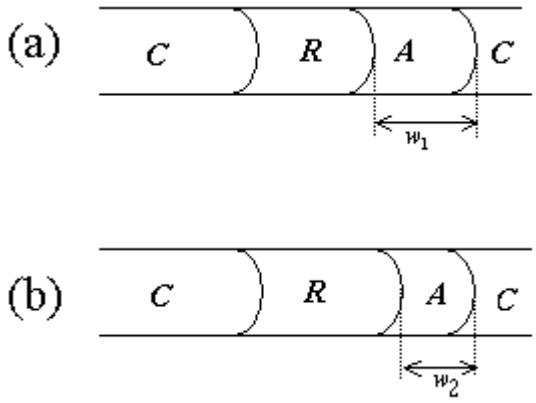


Figura 3.10- (a) Duplicata um do experimento FIA onde w_1 é o comprimento da zona amostral;(b) duplicata dois do experimento FIA onde w_2 é o comprimento da zona amostral; C = fluido carregador (tampão Britton-Robinson, pH = 4,5); A = Amostra (77µl); e R = Reagente (770µl, tampão Britton-Robinson, pH = 11,4).

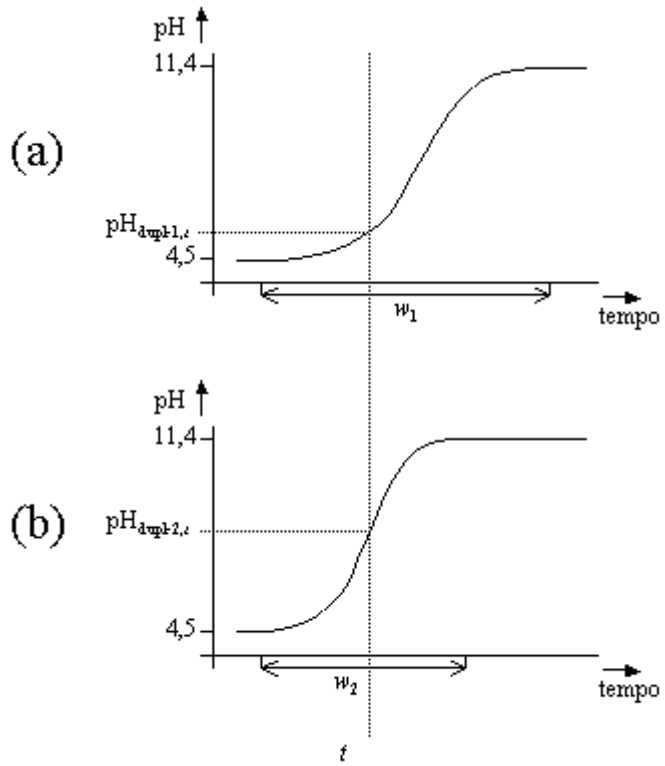


Figura 3.11 Curva do gradiente de pH e janela no tempo do FIA para medidas da (a) duplicata um do experimento FIA, onde w_1 é o comprimento da zona amostral (b)) duplicata dois do experimento FIA onde w_2 é o comprimento da zona amostral.

As expressões **135** e **136** mostram que o deslocamento do perfis de tempo devido à pequenas mudanças nas curvas do gradiente de pH depende dos valores de pK_a dos isômeros. Assim, quanto

maior for o valor do pK_a maior será o deslocamento nos perfis de tempo. Considerando o problema de se reproduzir, em duplicata por exemplo, exatamente o comprimento da zona amostral, é aparente que, quanto maior forem os valores de pK_a, maior será o erro, em termos de reprodutibilidade. Considerando os valores de pK_a dos três (2-HBA = 8,37, 3-HBA = 8,98 e 4-HBA = 7.61), é possível explicar porque a calibração de 3-HBA, que possui o maior pK_a dentre os três isômeros, geralmente resulta nos piores resultados de predição.

3.6 Conclusões

Ambos os métodos de calibração de segunda ordem, NBRA e RBL, se mostraram aplicáveis aos casos de sobreposição de posto presente nos dados. Entretanto, o método baseado na solução de um problema de autovalores-autovetores, NBRA, resultou nos piores resultados para os casos onde os perfis no tempo, do analito e interferente, apresentaram alta colinearidade. Isto ocorre pois a concentração a ser determinada depende do menor autovalor que é sensível à disposição espacial dos autovetores, que por sua vez, depende da independência dos perfis do tempo. A colinearidade entre perfis de tempo do analito e interferente resulta em subespaço instável definido pelos autovetores e, desta forma, o subespaço é mais sensível ao erro experimental. A aproximação por quadrados mínimos, neste caso o RBL, parece ser mais estável frente a colinearidades, o que é um aspecto importante, pois a colinearidade não pode ser eliminada dos dados experimentais. Estes resultados estão em acordo com o trabalho de Wang *et alli [Wang(a)]* que sugere o RBL como tendo, em geral, melhores propriedades na filtragem de ruídos se comparado com o NBRA.

A TMCA mostrou a presença de colinearidade, explicado pela similaridade entre os isômeros, e deslocamento dos perfis de tempo para as formas ácida e básica do 3-HBA. Nestas condições, ambos os métodos foram testados e apresentaram bons resultados para aqueles casos onde a colinearidade entre perfis e deslocamentos não foram significativos. No entanto, nos casos onde colinearidade e deslocamento foram significativos o método baseado em uma aproximação por quadrados mínimos, RBL, foi mais estável. Por outro lado, naqueles casos onde o deslocamento nos perfis de tempo é acentuado, o RBL também resulta resultados ruins de calibração.

Este trabalho mostra que, em geral, a solução por quadrados mínimos é um pouco mais "flexível" frente à similaridade dos perfis de tempo dos compostos. Assim, para aqueles casos onde alta colinearidade entre perfis é esperada, métodos empregando uma aproximação por quadrados mínimos são sugeridos, pois como já mencionado, se mostram mais estáveis e adicionalmente podem contar com

restrições e transformações para a redução da influência da alta colinearidade na análise numérica *[Kiers(f)]*. Por outro lado, os métodos baseados em uma solução de problemas de autovalores-autovetores são computacionalmente muito mais eficientes apresentando bons resultados para experimentos com boa reprodutibilidade.

A proposta inicial desta parte do trabalho de tese, ou seja, comparar as características algébricas de dois métodos, levou ao estudo da relação entre reprodutibilidade, variações experimentais e métodos, pois o tratamento geral, em que se baseiam os métodos, mostrou que os grandes erros de predição estavam associados a variações experimentais (reprodutibilidade). Neste estudo, chamado de análise exploratória, foi evidenciada a importância da restrição de sobreposição do posto para evitar a modelagem de ruídos. Enfim, as variações experimentais detectadas nestes dados requerem o emprego de métodos robustos para construção de modelos de calibração, como por exemplo, aqueles usados para reduzir efeitos de colinearidades *[*Kiers*(f)]*, no entanto, deve se pagar o preço do tempo computacional.

3.7 Apêndice 3.1

3.7.1 Método Generalizado Não Bilinear de Anulação do Posto- "Non-Bilinear Rank Annihilation (NBRA)"

A solução do problema de autovalores-autovetores do NBRA, como descrito no texto principal, tem como primeiro passo, expressar as matrizes **M** e **W** como se segue:

$$\begin{aligned}\mathbf{M} &= \mathbf{X_r D_r Y_r}^{\mathrm{T}} + \mathbf{X_s Y_s}^{\mathrm{T}} + \mathbf{X_t Y_t}^{\mathrm{T}} + \mathbf{E_M} \\ &= \mathbf{X_r}\left(\mathbf{Y_r D_r} + \left[\mathbf{Y_s} \mid \mathbf{0}\right]\right)^{\mathrm{T}} + \mathbf{X_t Y_t}^{\mathrm{T}} + \mathbf{E_M}\end{aligned} \qquad 137$$

onde a coluna de $\mathbf{X_s}$ é igual à primeira coluna de $\mathbf{X_r}$.

$$\begin{aligned}&\mathbf{W} = \mathbf{M} + \mathbf{N} \\ &= \mathbf{X_r}\left(\mathbf{Y_r D_r} + \left[\mathbf{Y_s} \mid \mathbf{0}\right]\right)^{\mathrm{T}} + \mathbf{X_t Y_t}^{\mathrm{T}} + \mathbf{E_M} + \mathbf{X_r Y_r}^{\mathrm{T}} + \mathbf{E_N} \\ &= \mathbf{X_r}\left(\mathbf{Y_r}\left(\mathbf{D_r} + \mathbf{I}\right) + \left[\mathbf{Y_s} \mid \mathbf{0}\right]\right)^{\mathrm{T}} + \mathbf{X_t Y_t}^{\mathrm{T}} + \mathbf{E_M} + \mathbf{E_N}\end{aligned} \qquad 138$$

A solução do problema de autovalores-autovetores mostrada na expressão **139**-*a* é encontrada com a solução do determinante dado na expressão **139**-*b*, que pode ser reescrito em termos de **139**-*c* através de $\overline{\mathbf{S}} = \overline{\mathbf{U}}^{\mathrm{T}}\overline{\mathbf{W}}\overline{\mathbf{V}}$ derivado da expressão **116**.

$$\begin{aligned}&\overline{\mathbf{U}}^{\mathrm{T}}\mathbf{M}\overline{\mathbf{V}}\overline{\mathbf{S}}^{-1}\mathbf{Z} = \mathbf{Z}\boldsymbol{\Lambda} \qquad (a) \\ &\left|\overline{\mathbf{U}}^{\mathrm{T}}\mathbf{M}\overline{\mathbf{V}} - \lambda\overline{\mathbf{S}}\right| = 0 \qquad (b) \\ &\left|\overline{\mathbf{U}}^{\mathrm{T}}\left(\mathbf{M} - \lambda\overline{\mathbf{W}}\right)\overline{\mathbf{V}}\right| = 0 \qquad (c)\end{aligned} \qquad 139$$

O determinante mostrado em **139**-*c* é reduzido à forma mostrada na expressão **141**. Isto feito, primeiro substituindo **M** e **W** na expressão **139**-*c* pelas expressões **137** e **138**, o que resulta em **140**.

$$\left|\mathbf{U}^{\mathrm{T}}\left[\mathbf{X_r} \mid \mathbf{X_t}\right]\left[\mathbf{Y_r}\left(\mathbf{D_r} - \lambda\left(\mathbf{D_r} + \mathbf{I}\right)\right) + \left[\left(1-\lambda\right)\mathbf{Y_s} \mid \mathbf{0}\right] \mid \left(1-\lambda\right)\mathbf{Y_t}\right]^{\mathrm{T}}\mathbf{V}\right| = 0 \qquad 140$$

Como a matriz quadrada $\mathbf{U}^{\mathrm{T}}(\mathbf{X_r}|\mathbf{X_t})$ é não singular, ela pode ser removida da expressão **140**, que é reduzida a **141**.

$$\left|\left[\mathbf{Y_r}\left(\mathbf{D_r} - \lambda\left(\mathbf{D_r} + \mathbf{I}\right)\right) + \left[\left(1-\lambda\right)\mathbf{Y_s} \mid \mathbf{0}\right] \mid \left(1-\lambda\right)\mathbf{Y_t}\right]^{\mathrm{T}}\mathbf{V}\right| = 0 \qquad 141$$

O determinante dado na expressão **141** é zerado por três autovalores da expressão **139**-*a*, podendo dois destes três, serem encontrados igualando as colunas da matriz

$\left[\mathbf{Y_r}\left(\mathbf{D_r}-\lambda\left(\mathbf{D_r}+\mathbf{I}\right)\right)+\left[\left(1-\lambda\right)\mathbf{Y_s}\,|\,\mathbf{0}\right]\,|\,\left(1-\lambda\right)\mathbf{Y_t}\right]$ a zero. Estes resultados são mostrados nas expressões **142**-*b* e **143**.

$$\mathbf{y}_2\left(\gamma-\lambda_2\left(\gamma+1\right)\right)=\mathbf{0} \quad (a)$$
$$\lambda_2=\frac{\gamma}{\gamma+1} \quad (b) \qquad 142$$

$$\lambda_3=1 \qquad 143$$

onde $\mathbf{y}_1$ e $\mathbf{y}_2$ são as colunas de $\mathbf{Y_r}$ e $\mathbf{y}_3$ é a coluna de $\mathbf{Y_s}$.

O autovalor λ_2 da expressão **142** é usado para encontrar o valor de γ, a concentração do analito na mistura. Em *[Wilson]* é sugerido que λ_2 é o menor autovalor dentre os três resultantes da solução do NBRA. O autovalor λ_2 é menor que 1 (*ver expressão* **142**-*b*), ou seja, ele é o menor ou aquele de valor intermediário. Em uma série de simulações foi verificado que, de fato, λ_2 é sempre o menor autovalor para este tipo de problema, desta forma, o autovalor λ_2 foi considerado como sendo o menor. A Tabela 3.3 mostra os autovalores encontrados pelo NBRA e os respectivos valores esperados para aquele autovalor relacionado à concentração, λ_2. Nesta Tabela é possível verificar que o menor autovalor determinado pelo NBRA é sempre aquele mais próximo do valor esperado, mesmo naqueles casos onde o NBRA falha.

Tabela 3.3 - Autovalores obtidos pelo NBRA

Mistura	Valor esperado do autovalor λ_2	Menor autovalor	Outros autovalores	
3HBA (100)	0,0909	0,0899	0,1433	0,6856
4HBA (40)	0,0385	0,0423	0,0518	0,0518
3HBA (100)	0,0909	0,1038	0,1573	0,8603
4HBA (60)	0,0566	0,0545	0,0681	0,0681
2HBA (50)	0,0476	0,0464	0,1152	0,1774
4HBA (60)	0,0566	0,0548	0,0548	0,0596
2HBA (100)	0,0909	0,0939	0,1168	0,1982
4HBA (60)	0,0566	0,0529	0,0631	0,0631
2HBA (50)	0,0476	0,0813	0,0813	0,1166
3HBA (100)	0,0909	0,1411	0,1411	0,1442
2HBA (100)	0,0909	0,1038	0,1038	0,1245
3HBA (50)	0,0476	0,1547	0,1547	0,1860

3.7.2 Bilinearização Residual "Residual Bilinearization (RBL)"

O algoritmo do RBL corresponde a um procedimento de quadrados mínimos alternantes, onde a seguinte função é minimizada:

$$\min_{\gamma,\mathbf{P},\mathbf{Q}} \left\| \mathbf{M} - \gamma \mathbf{X}_\mathbf{r} \mathbf{Y}_\mathbf{r}^{\mathrm{T}} - \overline{\mathbf{P}}\,\overline{\mathbf{Q}}^{\mathrm{T}} \right\|^2 . \qquad 144$$

γ é encontrado empregando uma regressão a partir das matrizes vetorizadas, $\overline{\mathbf{P}}$ e $\overline{\mathbf{Q}}$ são, respectivamente, as matrizes de escores e *loadings* determinadas através de uma Decomposição em Valores Singulares SVD ou por algoritmo NIPALS *[Vandeginste]* e "$\bar{}$" indica que $\mathbf{PQ}^{\mathrm{T}}$ estão truncadas com o pseudo-posto do interferente, ou seja, só as primeiras colunas da matriz **P** e **Q** são usadas (o número de colunas é igual ao pseudo-posto do interferente). Neste caso, o produto $\mathbf{X}_\mathbf{r}\mathbf{Y}_\mathbf{r}^{\mathrm{T}}$ corresponde à matriz de respostas do padrão.

Início:

O procedimento é iniciado com a seguinte estimativa de γ:

$$\gamma_0 = \text{vec}\left(\mathbf{X_r}\mathbf{Y_r}^{\text{T}}\right)^{+}\text{vec}(\mathbf{M}) \tag{145}$$

onde o operador de vetorização é dado por

$$\text{vec}\left(\mathbf{x}_1 \ldots \mathbf{x}_i \ldots \mathbf{x}_I\right) = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_i \\ \vdots \\ \mathbf{x}_I \end{pmatrix} \tag{146}$$

sendo $\mathbf{x}_i$ um vetor coluna e $\text{vec}\left(\mathbf{X_r}\mathbf{Y_r}^{\text{T}}\right)^{+} = \left[\text{vec}\left(\mathbf{X_r}\mathbf{Y_r}^{\text{T}}\right)^{T}\text{vec}\left(\mathbf{X_r}\mathbf{Y_r}^{\text{T}}\right)\right]^{\text{T}}\text{vec}\left(\mathbf{X_r}\mathbf{Y_r}^{\text{T}}\right)^{T}$

Etapa principal:

$\overline{\mathbf{P}}$ e $\overline{\mathbf{Q}}$ podem ser estimados com o algoritmo NIPALS ou por meio de uma SVD, alternando com a estimativa de γ, até que a convergência seja atingida.

$$\overline{\mathbf{P}}_k\overline{\mathbf{Q}}_k^{\text{T}} = \mathbf{M} - \gamma_k\mathbf{X_r}\mathbf{Y_r}^{\text{T}} \tag{147}$$

onde k indica o passo do algoritmo e neste caso, "$\bar{}$" indica que $\mathbf{P}_k\mathbf{Q}_k^{\text{T}}$ estão truncadas com o pseudo-posto do interferente, ou seja, só as primeiras colunas da matriz **P** e **Q** são usadas (o número de colunas é igual ao pseudo-posto do interferente).

$$\gamma_{k+1} = \mathbf{vec}\left(\mathbf{X_r}\mathbf{Y_r}^{\text{T}}\right)^{+}\left(\text{vec}(\mathbf{M}) - \text{vec}\left(\overline{\mathbf{P}}_k\overline{\mathbf{Q}}_k^{\text{T}}\right)\right) \tag{148}$$

Se (γ_{k+1} - γ_k) for maior que determinado valor, dado pelo critério de convergência, então $k = k + 1$ e uma nova interação é feita, caso contrário, pare.

4 Separação de Espectros de Luminescência Total de Amostras de Tártaro através do PARAFAC

4.1 Introdução

A identificação de cromóforos em sistemas biológicos requer, em muitos casos, a separação física dos compostos, o que, em geral, é difícil de ser obtida. Embora o desenvolvimento de técnicas espectroscópicas, como métodos hifenados *[Hirschfeld]*, tenha gerado diversas vantagens na identificação de compostos, os conjuntos de dados gerados por estas técnicas, em geral, são complicados de se trabalhar devido à quantidade de informação numérica neles presente. Nestes casos, a identificação direta de cromóforos (*i.e.* apenas com o emprego de técnicas espectroscópicas) depende da similaridade entre os cromóforos, em outras palavras, se estes não possuem espectros sobrepostos a identificação direta é possível. Por outro lado, se os espectros são sobrepostos, o problema se torna bastante complicado e a quimiometria tem desenvolvido métodos para lidar com este tipo de problema. Esta parte do trabalho de tese apresenta uma aplicação quimiométrica na identificação de compostos de amostras biológicas, através de dados espectroscópicos obtidos a partir de uma técnica hifenada, com o emprego de um método de separação de curvas.

Em trabalho anterior, foi verificado a presença de fluorescência na região do vermelho em amostras de tártaro dentário de felinos e caninos quando irradiados com luz na região do ultra violeta devido à presença de compostos porfirinicos *[Ferreira (a)]*. Em trabalho posterior, a análise de espectros de luminescência total de amostras humanas também mostrou que o mesmo tipo de compostos porfirinicos estava presente em uma amostra de tártaro humano *[Reis (a)]*, o que é confirmado nesta parte deste trabalho de tese, onde três novas amostras de tártaro humano são analisadas *[Reis (c)]*.

Os dados gerados, pela técnica espectroscópica hifenda de fluorescência, para cada amostra de tártaro corresponde a um tipo de dados em dois modos, onde um intervalo de comprimentos de ondas é varrido na excitação produzindo para cada comprimento de onda de excitação um espectro de emissão. Como resultado, uma superfície de intesidades de emissão é gerada, correspondendo uma dimensão aos comprimentos de onda da excitação e outra aos de emissão. A análise dos valores singulares, obtidos a partir das matrizes dos dados produzidos para três amostras, mostrou que, pelo menos, três cromóforos deveriam estar presentes, apresentando absorção nas regiões espectrais usadas para excitação e emissão. Com o intuito de efetuar a identificação dos perfis de emissão e de excitação, que são

proporcionais aos devidos espectros, para cada espécie presente nas amostras e com absorção na região estudada, uma separação de curvas é apresentada através do método PARAFAC (*do inglês PARAllel FACtor analysis*) *[Harshman]*. Este método é uma importante ferramenta tanto para separação de curvas quanto para calibração de ordem superior, ou seja, na identificção e quantificação de fluoróforos em amostras biológicas *[Ross, Bro(a), Bro(b)]*, especialmente em casos onde vários espectros estão sobrepostos, o que torna a identificação direta e quantificação difícil e, em vários casos, até mesmo impossível.

O ambiente não-ideal do experimento faz com que os dados sofram desvios daquele modelo teórico sugerido para eles (*i.e.* interferência de compostos com absorção em faixa espectral próxima àquela usada, espalhamento de luz e ruído experimental) gerando dificuldades para a separação de curvas. Devido a estes desvios, o PARAFAC foi escolhido para a separação de curvas, pois este método, permite a utilização de informações acerca dos dados, como por exemplo, a não-negatividade dos espectros, que são empregadas sob forma de restrições no processo de otimização usado pelo método para efetuar a separação de curvas *[Bro(b)]*.

Os resultados a serem apresentados mostram que o PARAFAC apresentou um bom desempenho com uma solução estável, o que foi verificado em procedimento de validação. Finalmente, um perfil resolvido é devido à interferência de compostos com absorção em faixa espectral próxima a daquela usada e os outros três perfis são atribuídos a perfis de emissão e excitação de três espécies porfirinicas, pois a região de excitação corresponde à da banda Soret (390-400nm), que é característica de transições eletrônicas de compostos porfirinicos *[Falk]*.

4.2 Dados

Cada amostra, em um total de três, de tártaro dentário humano foi dissolvida em ácido clorídrico 1:1(v/v). Os espectros de emissão foram coletados em uma região espectral entre 460 e 750nm, com um intervalo de 1nm, em um espectrofluorímetro SLM-AMINCO (SPF-500C), cuja fonte de radiação é uma lâmpada de Xe (250W). Estes espectros foram monitorados em faixa de excitação (390-450nm, com intervalo de 2nm), o que produziu um arranjo bidimensional de dados para cada amostra, onde cada linha, da matriz dos dados, corresponde a um espectro de emissão e cada coluna a um de excitação. O experimento foi efetuado em temperatura ambiente (*i.e* ~25^{o}C). Este experimento foi conduzido pela Doutoranda Débora Nakai Biloti do grupo do Prof. Dr Francisco Benedito Teixeira

Pessine e as amostras de tártaro foram cedidas pelo Prof. Dr Gustavo M. Teixeira da Universidade Camilo Castelo Branco.

O experimento descrito acima resulta em um conjunto de dados no qual a intensidade de emissão, μ, para cada fluoróforo, na concentração c_k, coletada em um dado comprimento de onda, λ_j^{em}, quando excitado no com luz do comprimento de onda, λ_i^{ex}, pode ser descrita por um modelo trilinear *[Ross]*:

$$\mu_{ijk} = \varepsilon_i \pi_j c_k \,, \tag{149}$$

onde ε_i é o coeficiente de extinção do fluoróforo para o comprimento de excitação λ_i^{ex}, π_j é relativo à emissão detectada no comprimento de onda λ_j^{em} e c_k é a concentração do fluoróforo. Se F fluoróforos contribuem para a emissão, então a intensidade de emissão (μ) pode ser escrita como:

$$\mu_{ijk} = \sum_{1}^{F} \varepsilon_{if} \pi_{jf} c_{kf} \quad . \tag{150}$$

A aplicação da expressão **150** requer uma baixa absorbância da espécie, ou que a amostra esteja bastante diluída, e que a excitação não seja transferida entre os cromóforos *[Ross]*.

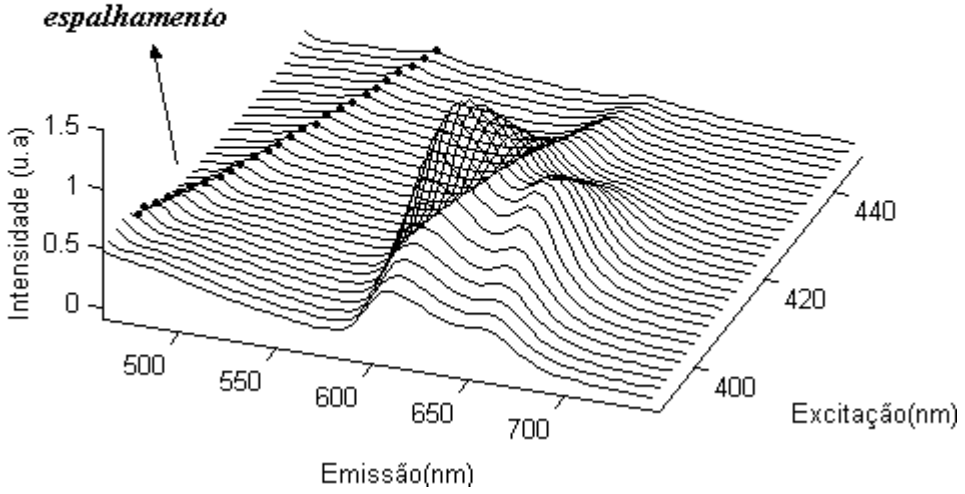


Figura 4.1 – Superfície de intensidades de emissão para a amostra 1. u.a. – unidade arbitrária.

A Figura 4.1 mostra os espectros de emissão para a amostra 1. A emissão detectada no início da faixa de comprimentos de onda empregada no experimento original apresenta uma banda cujo máximo varia com o comprimento de onda na excitação devido ao espalhamento Raman *[Andre]*. Este mesmo tipo de espalhamento é observado nas outras amostras como pode ser observado na Figura 4.2. Considerando que a banda de excitação é aquela a ser usada na identificação de fluoróforo, a faixa de emissão empregada na separação de curvas foi restrita entre 580 e 749nm, para reduzir, desta forma, a influência do espalhamento Raman. A Figura 4.2 mostra os espectros de emissão das três amostras para toda a região e para aquela empregada na análise.

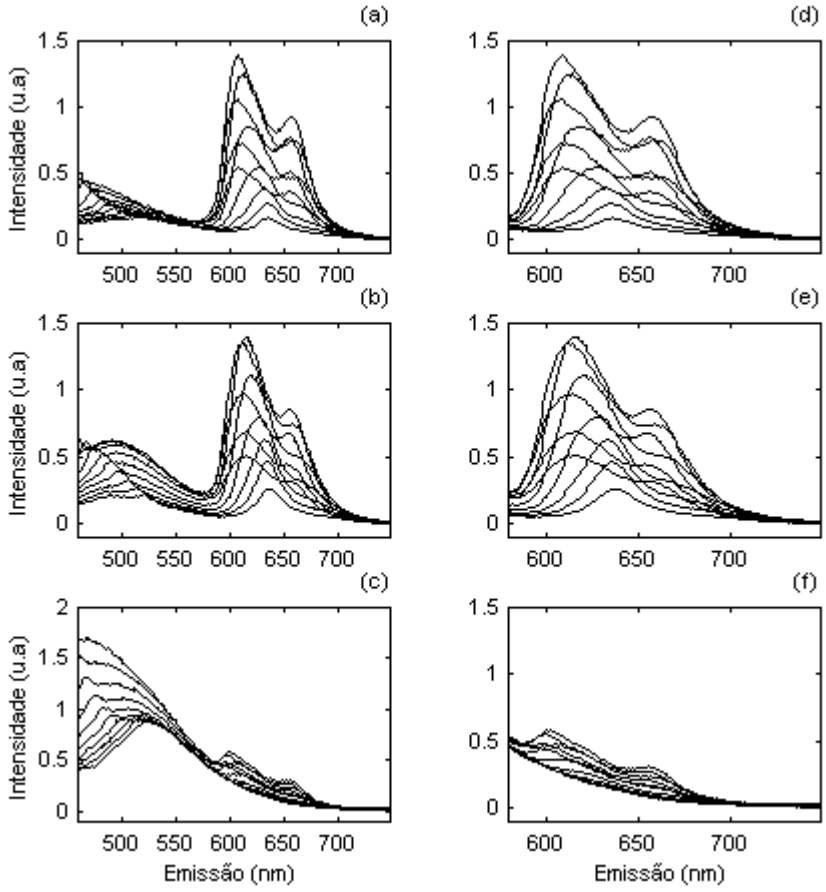


Figura 4.2 – Espectros de intensidade de emissão para diferentes comprimentos de onda na excitação. (a), (b), (c) correspondem à região espectral varrida pelo experimento. (d), (e), e (f) mostra a região espectral usada para a separação de curvas. (a,d) para amostra 1, (b,e) para amostra 2, (c,f) para amostra 3. u.a. – unidade arbitrária.

É importante notar que os espectros da amostra 3, na região espectral usada para a separação de curvas, apresentam uma grande influência, provavelmente, devido à emissão de compostos cujo espectro possui pico em faixa espectral abaixo daquela usada neste trabalho, tornando assim, a separação de curvas um tanto quanto complicada, pois estes compostos podem variar entre as amostras. Considerando este ambiente experimental, o objetivo deste trabalho é efetuar uma separação de curvas dos perfis de excitação e emissão através de um modelo trilinear.

4.3 Métodos

4.3.1 Modelo PARAFAC

O modelo de decomposição para um conjunto de dados com estrutura trilinear empregado pelo PARAFAC é dado por:

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk}, \qquad 151$$

onde x_{ijk} é o elemento (*i*, *j*, *k*) original do conjunto de dados trilinear, e a_{if}, b_{jf} , e c_{kf} são os elementos das matrizes componentes resultantes da decomposição do conjunto de dados com estrutura em três modos. e_{ijk} corresponde à parte dos dados que não segue um modelo trilinear.

A expressão **151** pode ser escrita em termos matriciais como dado por **152** (*ver seção Fundamentos*).

$$\underline{\mathbf{X}} = \mathbf{A}\underline{\mathbf{I}_{DS}}\left(\mathbf{C}^T \otimes \mathbf{B}^T\right) + \underline{\mathbf{E}} \qquad 152$$

onde $\underline{\mathbf{I}_{DS}}$ ($F\times[F\cdot F]$) e $\underline{\mathbf{X}}$ ($M\times[N\cdot R]$) correspondem à forma matricial do arranjo diagonal superior e do arranjo do dados em três modos, respectivamente. A matriz $\underline{\mathbf{X}}$ é construída pela justaposição horizontal de *R* matrizes de dimensões (*M*×*N*), que são chamadas *fatias*. *F* representa o número de componentes trilineares decompostos pelo PARAFAC. O arranjo diagonal superior em três modos é construído de forma semelhante à $\underline{\mathbf{X}}$ onde cada fatia corresponde a uma matriz quadrada *F*, que possui apenas um elemento diferente de zero e igual a um, isto é, o elemento *f*,*f* da diagonal da matriz quadrada *F*, sendo *f* o número da fatia. **A**, **B**, e **C** são as matrizes componentes de dimensões (*M*×*F*), (*N*×*F*) e (*R*×*F*), respectivamente *[Bro(c)]*.

As matrizes componentes **A**, **B** e **C**, com *F* vetores coluna correspondentes aos elementos a_{if}, b_{jf}, e c_{kf} respectivamente, são encontradas em um algoritmo de Quadrados Mínimos Alternantes "QMA" onde a função (*l*), (*ver seção Fundamentos*):

$$l(a,b,c) = \left\| x_{ijk} - \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} \right\|^2 , \qquad 153$$

é minimizada *[Bro(b)]*.

4.3.1.1 PARAFAC para dados com estrutura trilinear

O conjunto de dados de fluorescência pode ser descrito por um modelo PARAFAC em 3- Modos, onde as matrizes componentes **A** (*n*×*F*), **B** (*m*×*F*) e **C** (*r*×*F*) correspondem aos coeficientes de extinção de *n* comprimentos de onda usados na excitação para *F* fluoróforos, à emissão relativa detectada em *m* comprimentos de onda para *F* fluoróforos e concentrações de *r* amostras contendo *F* fluoróforos, respectivamente. Para o conjunto de dados de tártaro humano, a matriz **A** corresponde aos perfis de extinção, **B** aos perfis de emissão e **C** às concentrações relativas dos *F* fluoróforos (neste caso, a concentração encontrada é relativa, pois a escala correta para os coeficientes de extinção e emissão relativa não é conhecida). O conjunto de dados completo forma um Arranjo em Três Modos com dimensões 30×171×3 (30 comprimento de excitação $\times$ 171 comprimentos de onda de emissão $\times$ 3 amostras).

4.3.1.2 Restrições

O ambiente experimental não-ideal (por exemplo, interferência de compostos com absorção em região próxima àquela estudada, dados com ruídos) representa dificuldades no processo de otimização (*i.e.* mínimos locais) usado para efetuar a decomposição segundo o modelo PARAFAC. Nestes casos, restrições baseadas em informações a respeito dos dados (por exemplo, não-negatividade dos espectros), são usadas na otimização da função *l* (*ver* expressão **153**) para se obter soluções estáveis na decomposição trilinear. Testes preliminares com modelos sem restrições resultaram em perfis de excitação e emissão com alguns valores negativos. Embora estes valores negativos não afetem a identificação dos cromóforos, pois não alteram a forma e a posição das bandas de excitação, as restrições de não-negatividade e unimodalidade (*i.e.* o perfil deve apresentar apenas uma banda) foram usadas para certificar que o resultado final apresentasse significado físico. Neste trabalho, os valores

usados para dar início ao processo de otimização no algoritmo QMA foram valores aleatórios e a restrição de não negatividade foi imposta aos três modos (comprimentos de onda na excitação e emissão e concentrações relativas). Os perfis resultantes desta primeira decomposição foram usados para dar início à decomposição final, onde a restrição de não negatividade foi aplicada aos três modos e adicionalmente a restrição de unimodalidade foi aplicada ao modo dos comprimentos de onda da excitação. É importante notar que a restrição de unimodalidade foi imposta após a análise dos resultados de modelos para os quais, apenas a não-negatividade foi imposta, onde foi verificado que apenas uma banda era encontrada para cada perfil de excitação.

4.3.1.3 Validação

A etapa de validação é fundamental no PARAFAC para identificar a presença de mínimos locais na otimização da função *l*. O modelo PARAFAC foi validado, nesta parte do trabalho, através de um procedimento de reamostragem. Para tal, cada matriz original (30×171, *i.e* 30 linhas e 171 colunas) foi dividida em 35 novas matrizes com dimensões variando entre: (9×24), (9×25), (10×24) ou (10×25), dependendo do conjunto de validação. A primeira matriz foi gerada a partir da matriz original incluído as linhas 1, 4, 7,..., 28 (excluído as linhas 2, 3, 5, 6,..., 29, 30) e as colunas 1, 8, 15,..., 169 (eliminando as colunas 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14,..., 170, 171) o que resulta em uma matriz de dimensão (10×25). A segunda matriz foi derivada da matriz original incluindo as linhas 1, 4, 7,..., 28 (eliminando as linhas 2, 3, 5, 6,..., 29, 30) e as colunas 2, 9, 16,..., 170 (excluído as colunas 1, 3, 4, 5, 6, 7,8, 10, 11, 12, 13, 14,15,..., 171) e assim por diante. Este procedimento foi desenvolvido de tal forma que as linhas da última matriz correspondessem às linhas 5, 8, 11,…,29 da matriz original (tendo sido excluído as linhas 1,2, 3, 4, 6,..., 30) e às colunas de índices: 7, 14, 21, 28, 35,…, 154, 161, 168 da matriz original (com a eliminação das colunas 1,2, 3, 4, 5, 6, 8,9, 10, 11, 12, 13,..., 169,170, 171) e resultando desta forma, em uma matriz de dimensão (9×24). Assim, para cada conjunto de linhas 7 conjuntos de colunas foram derivados. Ou seja, a matriz original foi dividida em 5 matrizes (5 conjuntos de linhas), sendo cada uma destas 5 novas matrizes dividida em mais 7 novas matrizes (7 conjuntos de colunas), em resumo, cada matriz original gerou 35 novas matrizes. Com este procedimento o arranjo em três modos original gerou 35 novos arranjos. Estes 35 conjuntos de dados foram decompostos pelo PARAFAC para verificar a qualidade da decomposição (*i.e.* a presença de mínimos locais). (*Ver seção Fundamentos para exemplo simplificado*).

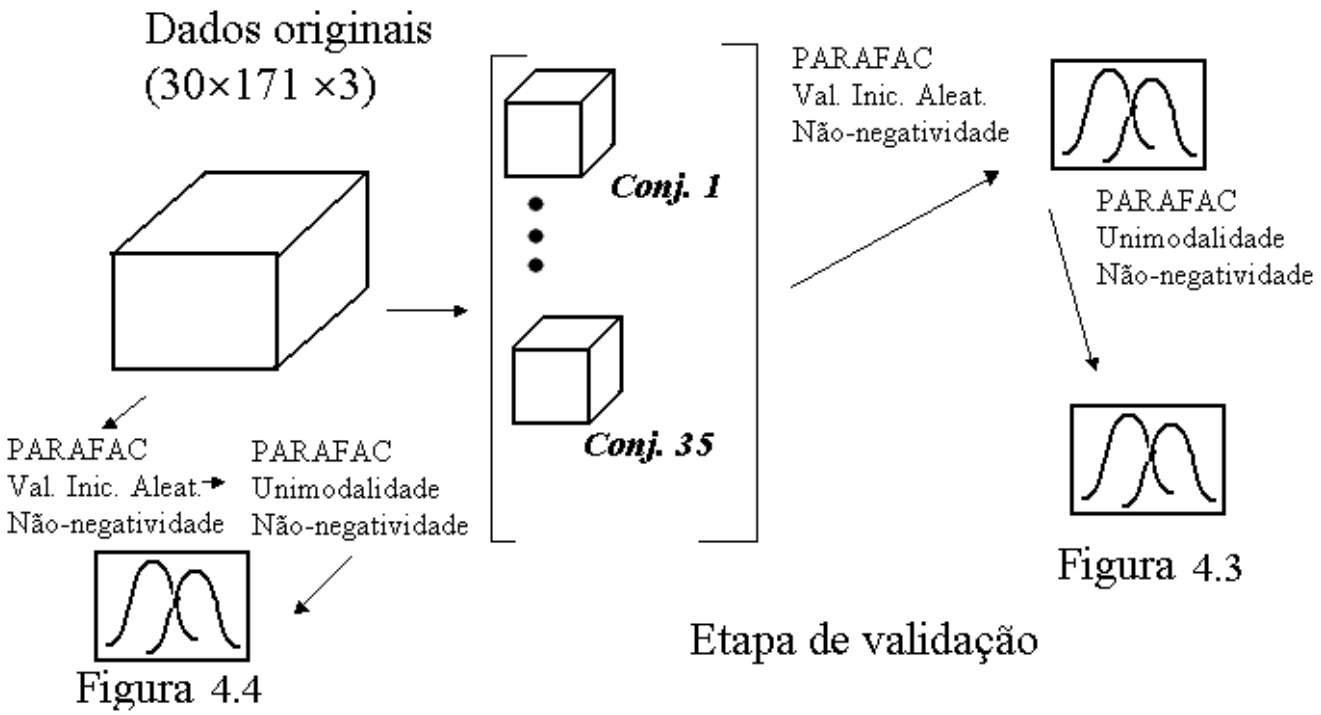
4.3.1.4 Número de fluoróforos

O número de fluoróforos foi escolhido através da comparação dos valores da função *P(F)*, como descrito na expressão 154, para seis modelos, onde o número de fatores (componentes), FACs (*i.e. F na* expressão 154), foi variado entre um e seis.

$$P(F) = \sum_{w=1}^{35} \left\| {}^{(w)}x_{ijk} - \sum_{1}^{F} {}^{(w)}a_{if}\, {}^{(w)}b_{jf}\, {}^{(w)}c_{kf} \right\|^2 \qquad 154$$

onde (*w*) – indica o conjunto de dados dentre os 35 diferentes arranjos em três modos descritos na etapa de validação.

Ao final, seis valores de *P(F)* foram calculados e sua variação analisada. Para confirmar o número de fluoróforos, três modelos foram testados: (1) com 3 fluoróforos; (2) com 4 fluoróforos; e (3) com 5 fluoróforos, isto feito para cada um dos 35 conjuntos de dados gerados na etapa de validação.
O Esquema 4.1 descreve a metodologia usada para a análise.



Esquema 4.1 – Resumo da análise de separação de curvas. Val. Inic. Aleat. – Valores Iniciais Aleatórios.

4.4 Resultados

O número de fluoróforos foi identificado, primeiro pela avaliação dos valores de *P(F)* (*ver* expressão **154**) de modelos decompostos com número de fatores variando entre 1 e 6 FACs, como mostrado na Tabela 4.1. A variação destes valores indica que o número de fatores poderia estar entre 3,4 ou 5, pois após 5 FACs esta variação se torna muito pequena. Em outras palavras, usando 6 ou mais fatores o valor de *P(F)* não decresce significantemente. A análise dos valores de *P(F)* não é suficiente para apontar o melhor número de fatores para a decomposição, mas ajuda a limitar o número de opções. Desta forma, três modelos PARAFAC foram obtidos para cada um dos 35 conjuntos de dados gerados na etapa de validação. Estes três modelos correspondem a três decomposições, onde foram considerados 3, 4 e 5 fatores. O melhor resultado foi obtido para o modelo com 4 fatores, onde os perfis decompostos concordam para os 35 conjuntos de dados (*ver* Esquema 4.1) como mostrado na Figura 4.3.

Tabela 4.1 – Valores da função *P(F)* para modelos PARAFAC usando números diferentes de fatores.

Número de Fatores	*P(F)*
1	2,0869
2	0,7962
3	0,1399
4	0,0222
5	0,0095
6	0,0045

P(F) - *ver* expressão **154**

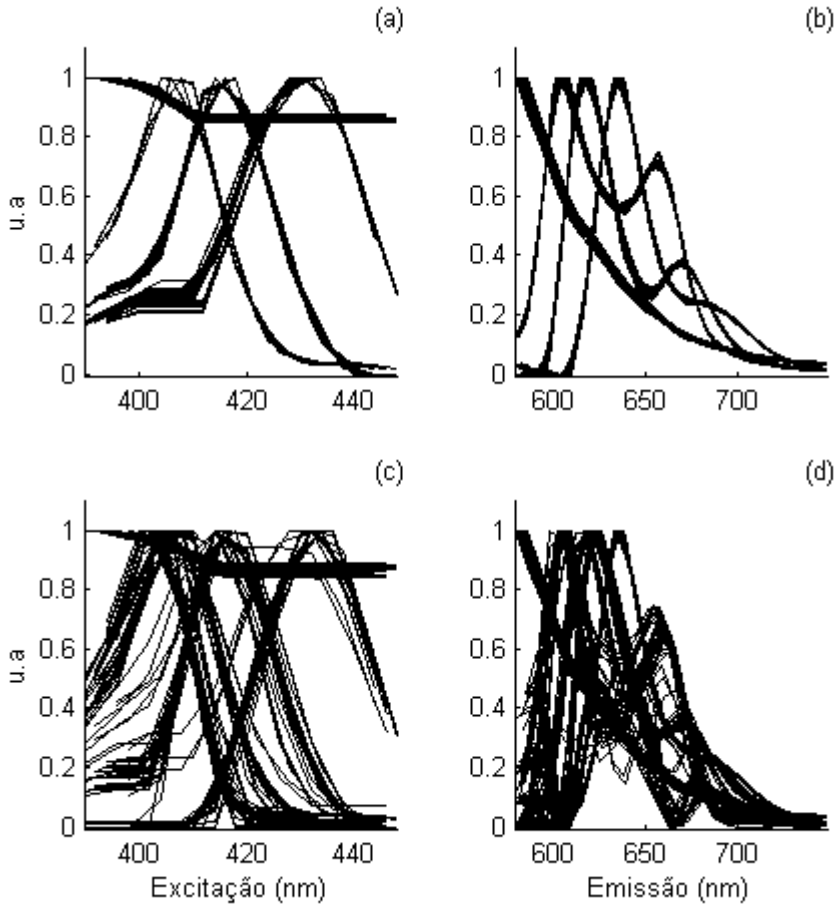


Figura 4.3 – Perfis resolvidos pelo método PARAFAC para: os modos de excitação, em (a), e emissão, em (b), empregando 4 fatores; para os modos de excitação, em (c), e emissão, em (d), para modelos com 5 fatores.

Os resultados para modelos considerando 3 fluoróforos apresentam perfis cuja forma sugere que mais perfis poderiam ser decompostos. Quanto ao modelo com 5 fluoróforos, os resultados para os 35 conjuntos de validação não concordam entre si, ou seja, um perfil de excitação que corresponde a um perfil de emissão obtido para um conjunto de dados dentre os 35, não corresponde ao mesmo perfil de emissão quando encontrado para outro conjunto de dados dentre os 35, como mostrado na Figura 4.3. Neste caso, dois ou mais perfis de excitação são considerados "iguais" se apresentarem o máximo da banda no "mesmo" comprimento de onda (*i.e.* a posição de dois máximos deve diferir de, pelo menos, 4 nm para se considerar os perfis diferentes, pois a resolução do modo de excitação é de 2 nm). Embora as medidas experimentais sejam diferentes para os 35 arranjos derivados da etapa de validação, a

posição do máximo dos perfis deve aparecer no "mesmo" comprimento de onda além de possuírem forma semelhante. Assim, o modelo com 4 FACs é o melhor.

O resultado final foi obtido com a decomposição do arranjo em três modos original (30×171×3) considerando 4 FACs, sendo mostrados na Figura 4.4. Um dos 4 perfis resultantes da decomposição, como mostrado na Figura 4.4, em (a) e (b) (identificado com o símbolo "×") não aparenta ser um perfil porfirinico, sendo considerado como interferência da banda presente na região compreendida entre 460 e 579nm (*ver* Figura 4.2) da região espectral de emissão. As concentrações apresentadas na Figura 4.5 são relativas, pois a unidade da concentração real não pode ser obtida, uma vez que as absortividades molares dos fluoróforos não são conhecidas e assim a devida escala para os perfis não pode ser obtida. Estas concentrações informam a composição relativa entre as três amostras.

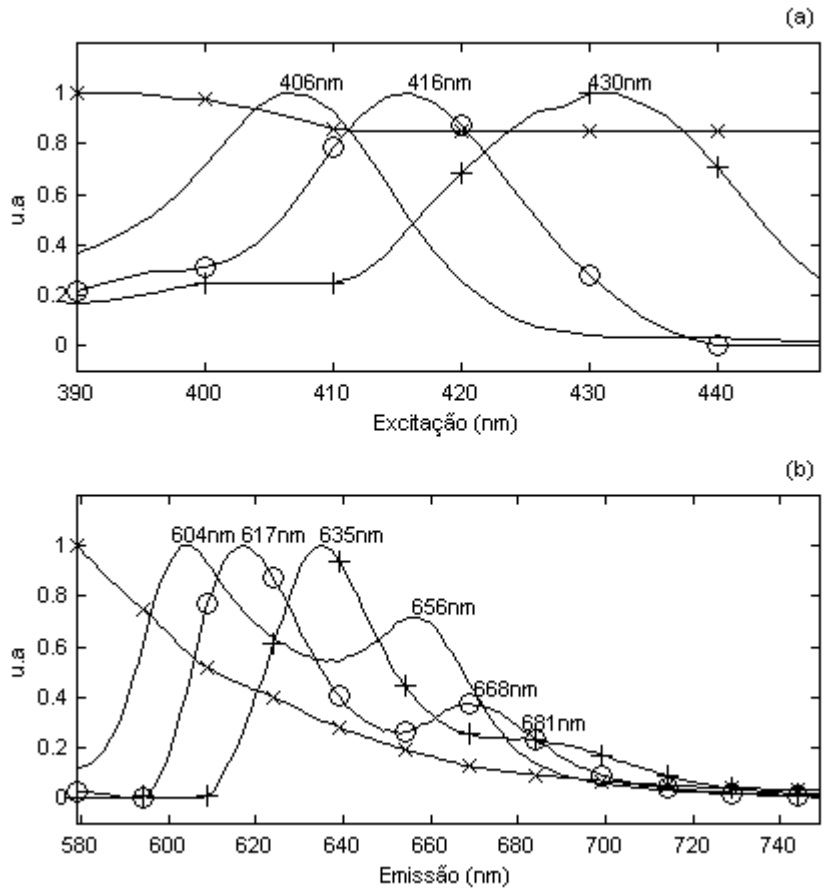


Figura 4.4 – Perfis resolvidos pelo método PARAFAC aplicado ao arranjo em três modos original, empregando 4 fatores. Em (a) para a excitação, (b) para emissão.

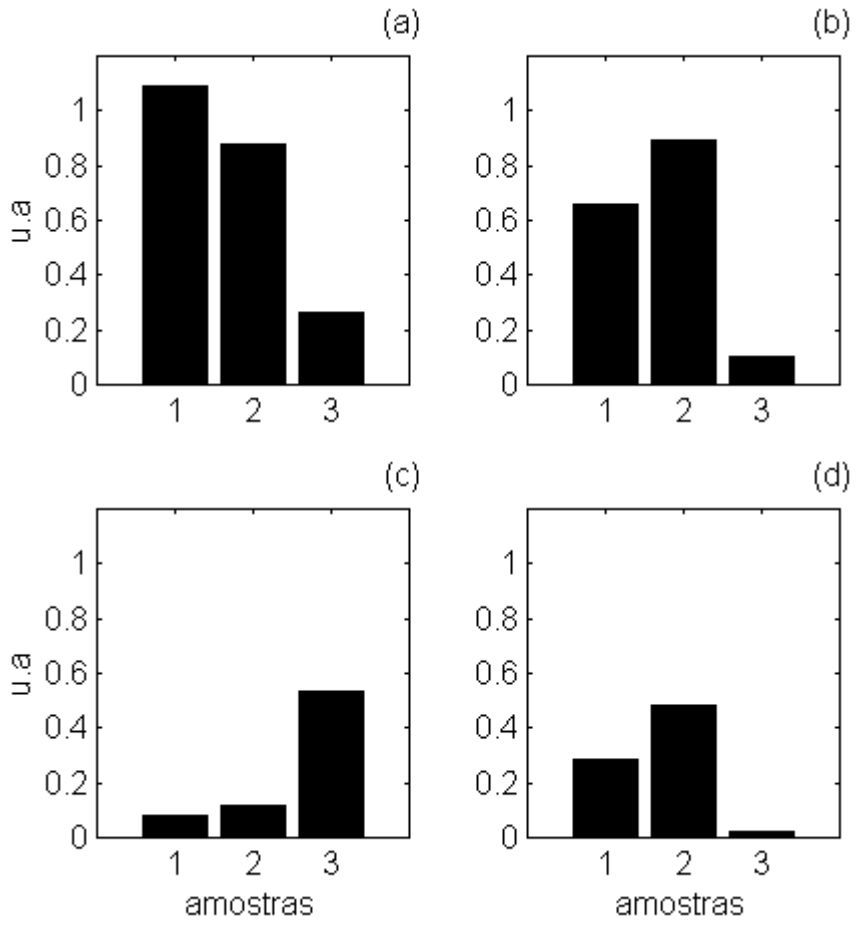


Figura 4.5 – Concentrações relativas encontradas pelo método PARAFAC aplicado ao arranjo em três modos original, empregando 4 fatores. A quantidade relativa das espécies são mostradas segundo sua discriminação na Figura 4.4, assim em: (a) para a espécie discriminada com linha contínua; (b) para espécie discriminada com o símbolo "O" ; (c) para a espécie discriminada com o símbolo "×" ; e (d) para espécie discriminada com o símbolo "+".

A Figura 4.6 apresenta o espectro de emissão de uma Hematoporfirina (*Hematoporphyrin Dihydrochloride-Sigma*) quando excitada com luz de comprimento de onda igual a 417 nm. Este espectro apresenta forma semelhante àquelas encontradas pelo PARAFAC e mostra que a banda na região de emissão 460-579nm deve, realmente, ser de um composto não porfirinico presente nas amostras de tártaro (*compare* Figuras 4.2 e 4.6).

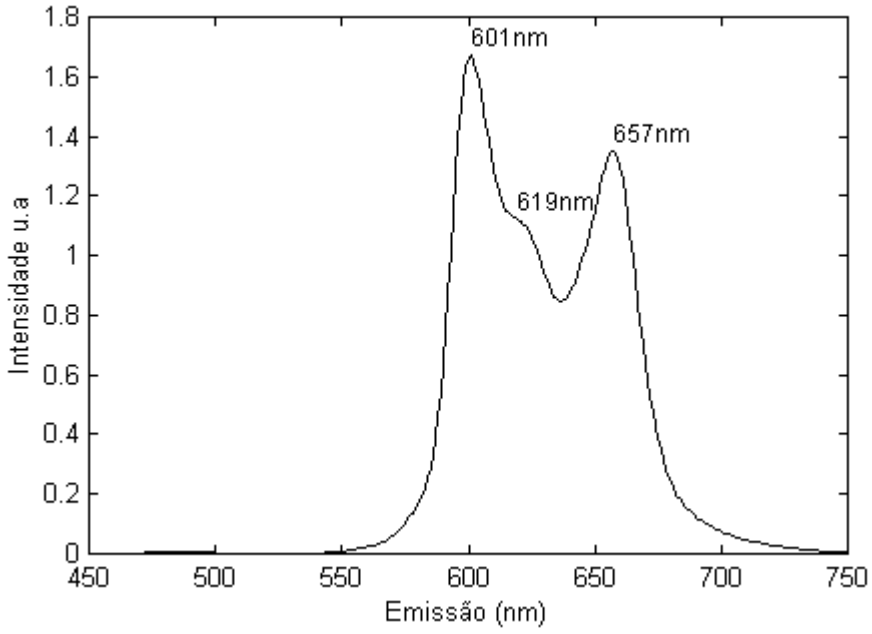


Figure 4.6 – Espectro de emissão de uma Haematoporfirina quando excitada com comprimento de onda de 417nm.

A Tabela 4.2 apresenta a posição dos máximos dos perfis de excitação obtidos neste trabalho e em trabalhos anteriores, sugerindo que espécies porfirinicas semelhante estão presentes nas amostras de tártaro humano, felino e canino, devido a proximidade dos máximos das bandas de excitação.

Tabela 4.2 – Banda Soret para espécies porfirinicas resolvidas neste trabalho e em trabalhos anteriores encontrados na literatura.

Referência	Comprimento de onda da excitação (nm)		
[Ferreira(a)]	410	417	436
[Reis(a)]	410	417	436
Este trabalho	407	416	431

4.5 Conclusões

A identificação dos compostos presentes nas amostras de tártaro poderia exigir a separação física destes compostos, pois seus espectros se apresentam sobrepostos. Alternativamente, o PARAFAC permitiu a identificação através de técnicas espectrocópicas. A identificação por completo das espécies porfirinicas exigiria uma etapa adicional, pois as amostras biológicas são com grande probabilidade

formadas de misturas de porfirinas muito semelhantes, o que dificulta a separação de curvas. Estas porfirinas podem possuir o mesmo núcleo, diferindo apenas nos grupos das cadeias laterais (por exemplo, um grupo acético em uma cadeia lateral que em outra porfirina aparece como um grupo propiônico) como descrito em Ferreira *et alli[Ferreira (a)]* Neste sentido, a maior contribuição deste trabalho é apresentar perfis de excitação e emissão obtidos através de uma separação de curvas através de uma solução validada para o PARAFAC.

5 *Análise Metodológica de Propriedades de Amidos extraídos de Fécula de Mandioca através de Modelos Tucker.*

5.1 Introdução

Os modelos Tucker, introduzidos por L. Tucker *[Tucker]* durante a década de sessenta para a interpretação de estudos psicológicos, têm sido aplicados na análise exploratória de dados ambientais ou dados químicos *[Kiers(a)]*, na identificação de compostos, calibração de ordem superior *[Smilde(c)]* e outros. Os métodos originais propostos por Tucker apresentam o problema de rotação livre, o que dificulta, em geral, a interpretação dos resultados *[Kiers(f)]*. Uma proposta para amenizar tal problema de rotação, é a rotação das matrizes componentes após a decomposição ter sido efetuada. Outra alternativa é restringir alguns parâmetros no modelo Tucker como sendo zero. Neste caso, o cálculo do modelo Tucker restrito para um conjunto de dados não significa determinar o modelo que descreve os dados, mas sim, determinar aqueles parâmetros que governam certos aspectos do conjunto de dados *[Kiers(f)]*.

A composição do amido e suas propriedades variam, principalmente, com a planta da qual é extraído *[Sarmento(a), Rickard]*. Este biopolímero é o principal componente da raiz da mandioca, sendo seu emprego industrial determinado, principalmente, por suas características físico-químicas. Neste trabalho é feita uma análise exploratória de propriedades físico-químicas de amidos extraídos de quatro cultivares de mandioca, colhidas em oito diferentes idades durante o período normalmente empregado para colheitas que se destinam ao uso industrial *[Sarmento(a)]*.

A análise exploratória destes dados se faz complexa devido à relação idade e fatores genéticos (diferentes cultivares) *versus* propriedades físico-químicas. Em trabalho anterior, esta complexidade foi verificada na análise multivariada de um conjunto similar de dados através da análise por componentes principais *[Sarmento(b)]*, o que confirma a importância e necessidade do emprego de técnicas sofisticadas de análise de dados, como modelos Tucker, para este tipo de estudo.

O aspecto metodológico deste trabalho está na avaliação da possibilidade da extração de informações, que permitam avaliar a relação idade e fatores genéticos *versus* propriedades físico-químicas, do conjunto de dados. Para tal, a metodologia proposta é baseada na decomposição do conjunto de dados em blocos de 3-modos através de modelos Tucker. O objetivo desta metodologia é "concentrar" em um bloco informações semelhantes a respeito das quatro cultivares, ou seja, amenizar

a influência de fatores genéticos que diferenciam as quatro cultivares. Três decomposições são testadas: A primeira, através de um modelo Tucker sem restrições; a segunda se diferencia da primeira por uma rotação; e a terceira e última, é um modelo Tucker restrito inspirado em modelos propostos por Kiers *[Kiers(d)]*, onde o arranjo do núcleo tem alguns de seus elementos fixados com o valor zero. Assim, o objetivo desta parte do trabalho de tese é fornecer informações sobre análise exploratória de dados através de métodos em três modos, sugerindo algumas ferramentas, como funções de inércia, para a avaliação dos modelos calculados.

5.2 Dados

O conjunto de dados é formado por propriedades físico químicas e funcionais de amidos extraídos de quatro cultivares (*i.e.* SRT 59-Branca de Santa Catarina, SRT 1287-Fibra, SRT 1105 Mico e IAC 12-829), cultivadas em Campinas, no estado de São Paulo e colhidas em oito diferentes idades: 10, 12, 14, 16, 18, 20, 22 e 24 meses, após terem sido plantadas (*i.e. primeira colheita em Julho, última colheita em Setembro do ano seguinte*). O período de colheita destas raízes é longo, sendo as condições climáticas bem definidas nesta região, em outras palavras, duas são as estações: A das chuvas, com altas temperaturas e, obviamente, chuvas; e condições de seca, com temperaturas baixas e pouca chuva. As variáveis estudadas foram: Conteúdo de amilose (Amilos[♣]); densidade absoluta do grânulo (Dens[♣]); poder de expansão (Pex[♣]) e porcentagem de solúveis (%sol[♣]) a 60, 75 e 90 graus Celsius; capacidade de ligação com água fria (Caplig[♣]); suscetibilidade enzimática (Pac[♣]) em termos de açúcares redutores produzidos após 3, 6 e 24 horas de ataque enzimático; área superficial específica do grânulo (Superf[♣]) e propriedade da pasta: Pico de viscosidade (PVisc[♣]) e tendência à retrogradação (Ret[♣]). Estes dados foram organizados em um arranjo em três modos tendo como modos: Idade×Propriedades×Cultivares. Estes dados são resultado da tese de doutoramento da Profa. Dra Silene B. S. Sarmento *[Sarmento(a)]* que gentilmente os cedeu e participou de forma efetiva na discussão da análise dos dados.

[♣] Legenda na figuras

5.3 Pré-processamento

As fatias do arranjo em multi modos (*i.e.* matrizes das cultivares (idade×propriedades)) foram justapostas verticalmente e a matriz resultante foi autoescalada (*i.e.* o valor médio e o desvio padrão de cada coluna foram calculados e cada elemento das colunas foram subtraídos de suas respectivas médias e divididos por seus respectivos desvios padrão).

5.4 Teoria

5.4.1 Modelo Tucker-MT

O modelo Tucker é o mais geral para dados em arranjos multi modos. Este modelo é descrito pela expressão **155**.

$$x_{ijk} = \left(\sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} g_{pqr} \right) + e_{ijk} \qquad 155$$

onde a_{ip}, b_{jq} e c_{kr} descrevem os elementos da matrizes componentes **A** (para o modo A), **B** (para o modo B) e **C** (para o modo C) de ordem $I \times P$, $J \times Q$ e $K \times R$ respectivamente. g_{pqr} é o elemento (p,q,r) do núcleo $\underline{\mathbf{G}}$ de dimensão $P \times Q \times R$. Finalmente e_{ijk} descreve o erro para o elemento x_{ijk}, quando decomposto pelo modelo Tucker, sendo um elemento do arranjo $\underline{\mathbf{E}}$ de dimensão $I \times J \times K$.

A formulação matricial do MT é mostrada na expressão **156** (*ver seção fundamentos*).

$$\underline{\mathbf{X}} = \mathbf{A}\underline{\mathbf{G}}\left(\mathbf{C}^T \otimes \mathbf{B}^T\right) + \mathbf{E} \qquad 156$$

Nesta parte do trabalho de tese, é considerada a decomposição do MT em R blocos como demonstrado na expressão **157**.

$$\underline{\mathbf{X}} = \mathbf{A}\mathbf{G}_1\left(\mathbf{c}_1^T \otimes \mathbf{B}^T\right) + \mathbf{A}\mathbf{G}_2\left(\mathbf{c}_2^T \otimes \mathbf{B}^T\right) + \cdots + \mathbf{A}\mathbf{G}_R\left(\mathbf{c}_R^T \otimes \mathbf{B}^T\right) + \mathbf{E} \qquad 157$$

5.4.2 Funções de inércia

Os modelos Tucker são calculados através de um algoritmo de Quadrados Mínimos Alternantes "QMA" *[Kiers(b)]* onde a função dada na expressão **159** é minimizada (*ver seção fundamentos*).

$$l(\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{G}) = \left\| \underline{\mathbf{X}} - \mathbf{A}\underline{\mathbf{G}}\left(\mathbf{C}^T \otimes \mathbf{B}^T\right) \right\|^2 \qquad 158$$

Os MTs são avaliados aqui através de dois tipos de "Funções de Inércia". A primeira "*FIT* ", expressão **159**, é baseada na função, minimizada pelo algoritmo QMA, usada no cálculo dos modelos Tucker.

$$f = \left(1 - \frac{\left\| \underline{\mathbf{X}} - \mathbf{A}\underline{\mathbf{G}}\left(\mathbf{C}^T \otimes \mathbf{B}^T\right) \right\|^2}{\left\| \underline{\mathbf{X}} \right\|^2} \right) \cdot 100\% \qquad 159$$

A outra função, "*FIP* ", expressão **160**, é usada para verificar como as fatias do arranjo núcleo descrevem as fatias do arranjo de dados.

IFP:

$$f\left(\mathbf{G}_k, c_{qk}, \mathbf{X}_q\right) = \left(1 - \frac{\left\| vec\mathbf{X}_q - \left[\left(\mathbf{A} \otimes \mathbf{B}\right) vec\mathbf{G}_k c_{qk} \right] \right\|^2}{vec\mathbf{X}_q^T vec\mathbf{X}_q} \right) \cdot 100\% \qquad 160$$

onde $vec\mathbf{X}_q$ é a fatia do arranjo de dados colocada em forma de vetor coluna pelo operador *vec*, $vec\mathbf{G}_k$, é a fatia do núcleo colocada na forma de um vetor coluna, c_{qk} é o elemento da matriz componente para o modo C.

5.4.3 Metodologia

O principal objetivo da análise em multi modos desta parte do trabalho de tese é efetuar uma decomposição, na qual informações a cerca das propriedades do amido, comum às quatro cultivares, sejam capturadas em um único bloco. Para tal, a primeira suposição é que, pelo menos, um vetor coluna da matriz componente para o modo das cultivares seja não-negativo. Com isto, é sugerido que este(s) vetor(es) represente(m) o mesmo tipo de informação em todas as cultivares mas em diferentes proporções. No caso deste trabalho, apenas um vetor coluna pode ser não-negativo devido à restrição de ortogonalidade aplicada aos três modos. Como resultado, a fatia do núcleo associada a este vetor não-negativo da matriz componente para o modo C (modo das cultivares) deve descrever informações, a respeito da relação entre propriedades e idade, semelhantes para as quatro cultivares. Desta forma, a

análise em três modos é conduzida considerando o vetor coluna não-negativo da matriz componente do modo C, isto feito através do Modelo Tucker "MT", Modelo Tucker Restrito "MTR" e Modelo Tucker com Rotação "MTRot".

5.4.4 Modelo Tucker Restrito-MTR

O modelo Tucker restrito é usado para extrair informações que permitam, ou ao menos ajudem, a compreender a relação entre a idade e propriedades do amido, isto feito em análise complementada por meio de gráficos. Para tal, supõe-se que o núcleo seja composto por duas partes: O núcleo restrito e o núcleo complementar. O núcleo restrito possui parte de seus elementos fixados com o valor zero e o oposto para o núcleo complementar. Assim, aqueles elementos iguais a zeros em um núcleo não serão fixados no outro e vice e versa. Desta forma, quando o modelo Tucker restrito é calculado com o núcleo restrito também existe um modelo complementar, ou seja, que descreve a parte não modelada pelo MTR. Os elementos fixados com zero no núcleo restrito são escolhidos, primeiro, considerando a ortogonalidade dos vetores coluna de $vec\underline{\mathbf{G}}_C$, conseguida com a imposição de que apenas um elemento de cada vetor linha de $vec\underline{\mathbf{G}}_C$ pode ser diferente de zero (*ver seção fundamentos para descrição da aplicação do operador vec em* $\underline{\mathbf{G}}_C$). Segundo, com a consideração do resultado do modelo Tucker com rotação (*a ser descrito a seguir*) e por último, os valores da função *FIP* que devem ser sempre não-negativos. Desta forma, a primeira fatia do núcleo restrito a ter os elementos escolhidos, para serem fixados a zero, é aquela associada ao vetor não negativo. Neste trabalho, esta escolha levou em consideração o resultado do modelo Tucker com rotação, ou seja, os elementos, com os menores valores em módulo, daquela fatia do núcleo do MTRot associada ao vetor coluna não negativo, foram fixados para zero. Na fatia seguinte, os menores valores em módulo foram fixados para zero, respeitando a imposição de que os elementos não fixados para zero em uma fatia devem ser fixados para zero nas outras fatias. Este procedimento foi seguido até a última fatia. O último parâmetro avaliado é o valor da função $f(\mathbf{G}_{G\text{-}qk})$, que deve ser não-negativo para todas as fatias do núcleo restrito. Desta forma, o MTR é calculado e os valores de $f(\mathbf{G}_{G\text{-}qk})$ são conferidos para a identificação de valores negativos: Se um ou mais valores negativos forem encontrados o núcleo restrito deve ser reestruturado. O resumo do modelo MTR é dado a seguir:

$$\underline{\mathbf{G}} = \underline{\mathbf{G}}_C + \underline{\mathbf{G}}_S \qquad 161$$

$$\underline{\mathbf{G}}_C = \left(\mathbf{G}_{C-1} \middle| \quad \mathbf{G}_{C-2} \middle| \quad \cdots \quad \middle| \quad \mathbf{G}_{C-R} \right) \tag{162}$$

$$\underline{\mathbf{G}}_S = \left(\mathbf{G}_{S-1} \middle| \quad \mathbf{G}_{S-2} \middle| \quad \cdots \quad \middle| \quad \mathbf{G}_{S-R} \right) \tag{163}$$

$$\underline{\mathbf{X}} = \mathbf{A}\left[\underline{\mathbf{G}}_C + \underline{\mathbf{G}}_S\right]\left(\mathbf{C}^T \otimes \mathbf{B}^T\right) \tag{164}$$

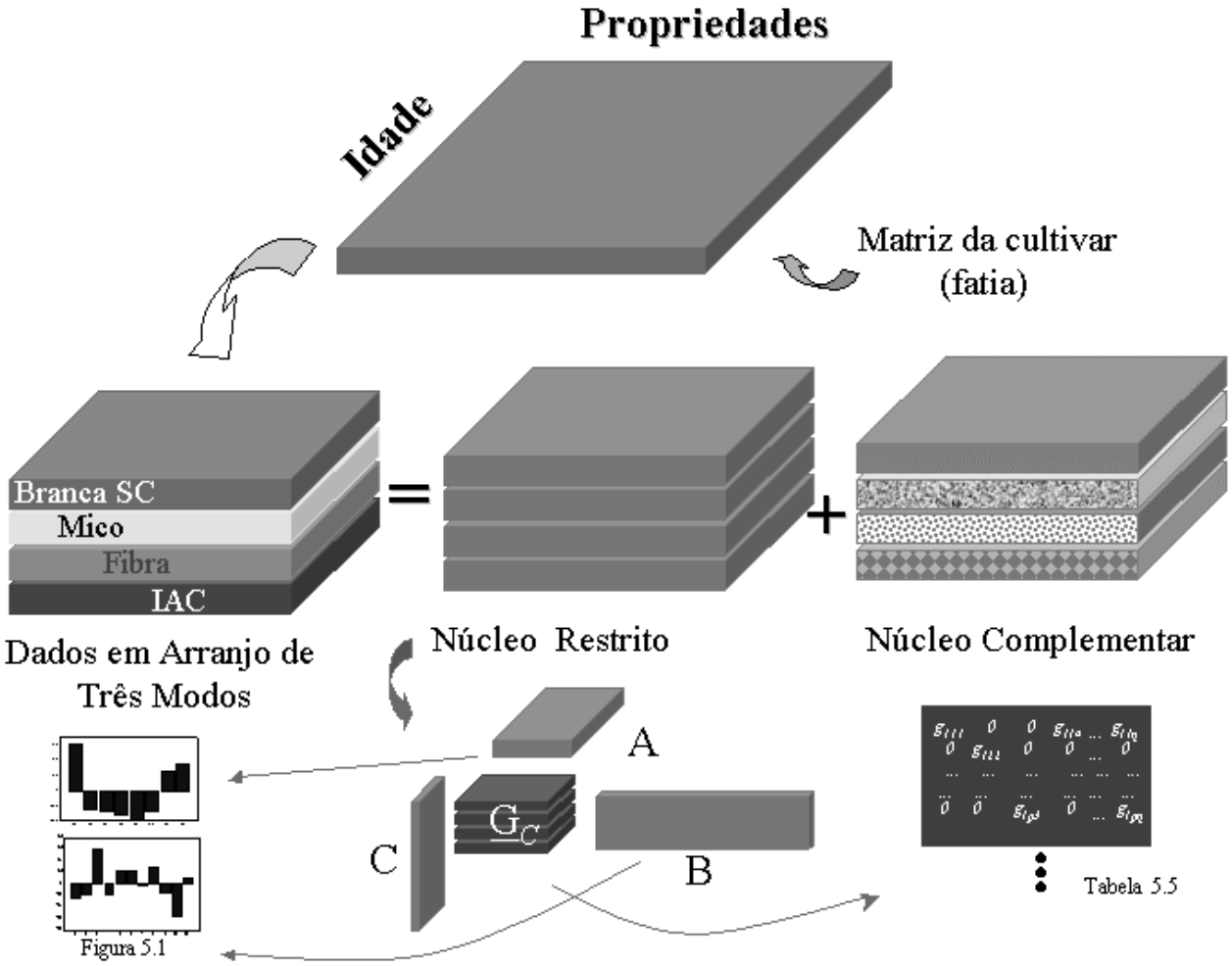
onde *c* e *s* indicam o núcleo restrito e complementar, respectivamente.

A decomposição proposta pelo MTR é baseada na decomposição do núcleo sem restrições do modelo Tucker em dois núcleos. O arranjo dado na expressão **164** pode ser reescrito na expressão **165**, e assim a decomposição do núcleo é dada como se segue:

$$vec\underline{\mathbf{X}} = \left(\mathbf{A} \otimes \mathbf{B}\right)\left(vec\mathbf{G}_{C-1}\mathbf{c}_1^T + \cdots + vec\mathbf{G}_{C-R}\mathbf{c}_R^T + vec\mathbf{G}_{S-1}\mathbf{c}_1^T + \cdots + vec\mathbf{G}_{S-R}\mathbf{c}_R^T\right) \tag{165}$$

$$vec\underline{\mathbf{X}} = \left(\mathbf{A} \otimes \mathbf{B}\right)\left(vec\mathbf{G}_{C-1}\mathbf{c}_1^T + \cdots + vec\mathbf{G}_{C-R}\mathbf{c}_R^T\right) + \left(\mathbf{A} \otimes \mathbf{B}\right)\left(vec\mathbf{G}_{S-1}\mathbf{c}_1^T + \cdots + vec\mathbf{G}_{S-R}\mathbf{c}_R^T\right) \tag{166}$$

Um importante aspecto sugerido por esta decomposição é que a informação capturada pelo arranjo $\underline{G}_C$, por exemplo em $vec\mathbf{G}_{C-1}\mathbf{c}_1^T$, é independente do arranjo $\underline{G}_S$, pois não é possível descrever os componentes de $vecG_C$ por meio dos componentes $vecG_S$. Por exemplo, $vec\mathbf{G}_{C-1}\mathbf{c}_1^T$ não pode ser descrito como uma combinação linear dos componentes de $vecG_S$ ($vec\mathbf{G}_{C-1}\mathbf{c}_1^T = vec\mathbf{G}_{S-1}w_1\mathbf{c}_1^T + \cdots + vec\mathbf{G}_{S-R}w_R\mathbf{c}_R^T$) devido à ortogonalidade dos vetores coluna de **C** e ao fato de que $vec\mathbf{G}_{S-1}^T vec\mathbf{G}_{C-1} = 0$, pois as posições dos elementos fixados em zero são opostas para os componentes de *S* e *C*. Esta análise também é resumida no Esquema 5.1.



Esquena 5.1 – Resumo da análise

5.4.5 Modelo Tucker com Rotação-MTRot

O modelo Tucker com rotação é gerado através da rotação do núcleo com uma rotação Orthomax *[Kiers(e)]* para os três modos, como se segue :

$$\underline{\mathbf{X}} = \mathbf{A}\mathbf{S}\mathbf{S}^T\,\underline{\mathbf{G}}\left(\mathbf{U}\otimes\mathbf{T}\right)\left(\mathbf{U}^T\otimes\mathbf{T}^T\right)\left(\mathbf{C}^T\otimes\mathbf{B}^T\right) \qquad 167$$

$$\mathbf{S}\mathbf{S}^T = \mathbf{S}^T\mathbf{S} = \mathbf{I} \therefore \mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I} \therefore \mathbf{T}^T\mathbf{T} = \mathbf{T}\mathbf{T}^T = \mathbf{I} \qquad 168$$

$$\underline{\mathbf{X}} = \tilde{\mathbf{A}}\underline{\tilde{\mathbf{G}}}\left(\tilde{\mathbf{C}}^{T} \otimes \tilde{\mathbf{B}}^{T}\right) \qquad 169$$

$$\tilde{\mathbf{A}} = \mathbf{AS} \therefore \underline{\tilde{\mathbf{G}}} = \mathbf{S}^{T}\underline{\mathbf{G}}\left(\mathbf{U} \otimes \mathbf{T}\right) \therefore \tilde{\mathbf{C}} = \mathbf{CU} \therefore \tilde{\mathbf{B}} = \mathbf{BT} \qquad 170$$

onde **S**, **T** e **U** são as matrizes de rotação para os modos **A**, **B** e **C**, respectivamente e $\tilde{\mathbf{A}}$, $\tilde{\mathbf{B}}$ e $\tilde{\mathbf{C}}$ correspondem às matrizes componentes após a rotação.

5.5 Resultados

A primeira análise foi efetuada com as propriedades: Conteúdo de amilose (Amilos); densidade absoluta do grânulo (Dens); poder de expansão (Pex) e porcentagem de solúveis (%sol) a 60 e 90 graus Celsius; capacidade de ligação com água fria (Caplig); suscetibilidade enzimática (Pac) em termos de açúcares redutores produzidos após 3, 6 e 24 horas de ataque enzimático; área superficial específica do grânulo (Superf). Estes dados foram organizados em arranjos em três modos tendo como dimensões (8×11×4).

Os valores singulares das matrizes (ou fatias) escaladas, em separado e justapostas, são apresentados na Tabela 5.1 e sugerem como valor para o posto, pelo menos, 8 para cada fatia. Assim, os modelos Tucker foram calculados tendo como dimensões para os modos A, B e C os valores (8×8×4) respectivamente. O MT descreve 95.80% dos dados, o que é medido pela função de inércia dada pela expressão **159**, e MTR descreve 85.81 %.

Tabela 5.1 – Valores singulares para matrizes (idade×propriedades) das 4 cultivares, justapostas horizontalmente e verticalmente.

	Valores singulares					
	BSC*	Mico*	Fibra*	IAC*	X1#	X2+
1	7,9349	4,6923	5,0551	5,9045	10,8104	10,0823
2	4,2190	3,9153	4,1699	3,9585	8,8409	8,3358
3	3,1204	3,4352	3,8742	3,5939	6,5597	6,5593
4	2,3468	2,8444	3,4254	2,8041	6,3111	6,0054
5	2,0159	2,2793	2,9542	2,4515	4,6355	5,1407
6	1,4671	1,8580	2,2703	1,6155	4,3996	4,8358
7	1,0417	1,5234	1,6500	1,2270	3,7283	4,1320
8	0,4981	0,4314	0,7402	0,8817	2,8932	3,0891

* Matrizes das cultivares, # Matrizes das cultivares justapostas verticalmente ,+ Matrizes das cultivares justapostas horizontalmente.

A Tabela 5.2 apresenta os vetores coluna não-negativos para o modo C (modo das variedades) para os três modelos. Nesta tabela, é possível verificar que os vetores para MTRot e MTR apresentam maior similaridade entre seus elementos, o que pode ser um bom aspecto, pois este vetor pode descrever as propriedades dos amidos que são semelhantes para as quatro cultivares. Embora o vetor não-negativo para o modelo MTRot apresente uma similaridade interessante entre os valores de seus elementos, a função *FIP* apresenta valores negativos como mostrado na Tabela 5.3. Mesmo sendo os valores negativos da *FIP* para o MTRot pouco significativos, eles sugerem que as fatias do núcleo, correspondentes a eles, capturaram mais informação do que poderiam. Em outras palavras, a distribuição dos valores no vetor coluna não-negativo não possui interpretação física mesmo tendo sido calculados de forma correta como mostra o Apêndice 5.1.

Tabela 5.2- Vetores coluna não-negativos das matrizes componentes para o modo das cultivares

	Vetores		
	MT	MTRot	MTR
BSC	0,2141	0,4738	0,4718
Mico	0,4636	0,4886	0,4981
Fibra	0,6163	0,4607	0,5107
IAC	0,5995	0,5697	0,5182

Tabela 5.3 – Valores das funções de inércia *FIP*.

Cultivares	Inércia %							
	MTRot				MTR			
	$^{*}G_1$	$^{*}G_2$	$^{*}G_3$	$^{*}G_4$	$^{*}G_1$	$^{*}G_2$	$^{*}G_3$	$^{*}G_4$
BSC	60,3392	22,7973	0,8490	1,8176	26,2985	62,3342	0,1074	0,4401
Mico	-0,4543	52,9427	-1,3967	42,6424	50,6217	0,0115	3,4313	26,7182
Fibra	35,8158	47,6670	26,5165	-0,3350	45,6891	25,7686	7,2344	7,9505
IAC	20,8183	57,8967	7,4220	9,3006	49,4570	16,2763	17,0759	0,7268

* Fatias do núcleo.

A Tabela 5.4 mostra os valores *FIP* associados ao vetor não-negativo para os três modelos, onde pode ser verificado que os valores para o MTR apresentam maior similaridade entre si.

Os valores de *FIP* são importantes para a determinação das melhores posições do núcleo do MTR a serem fixadas para zero, pois as melhores combinações destas posições devem gerar valores não-negativos de *FIP*. Os valores de *FIP*, quando usados como parâmetro de comparação, informam como o MTR descreve os dados. Valores negativos para *FIP* indicam que a parte não modelada pelo MTR possui grande importância como mostrado no Apêndice 5.1.

Tabela 5.4- Valores das funções de inércia *FIP* para fatias do núcleo correspondentes aos vetores coluna não-negativos das matrizes componentes para o modo das cultivares.

	Inércia %		
	MT	MTRot	MTR
BSC	6,7886	22,7973	26,2985
Mico	48,1736	52,9427	50,6217
Fibra	67,0748	47,6670	45,6891
IAC	67,0252	57,8967	49,4570

A Tabela 5.5 apresenta a fatia do núcleo do MTR correspondente ao vetor coluna não negativo da matriz componente do modo C. O elemento na coluna 6 e linha 7 na Tabela 5.5 é o terceiro mais importante, considerando seu valor em módulo, de todos os elementos do núcleo deste modelo. Este elemento corresponde ao vetor coluna 7 da matriz componente para o modo A, $\mathbf{a}_7$, e ao vetor coluna 6 da matriz componente para o modo B, $\mathbf{b}_6$. $\mathbf{a}_7$ informa sobre os efeitos sazonais nas características do amido. A Figura 5.1, apresenta o vetor coluna $\mathbf{a}_7$, onde aparecem dois grupos distintos de amostras: O Primeiro Grupo é formado por amostras correspondentes às idades de 10, 22 e 24 meses (estação seca e fria); e o Segundo Grupo para as idades de 12, 14, 16, 18 e 20 meses (estação quente e chuvosa).

Tabela 5.5 – Fatia do núcleo do modelo Tucker restrito correspondente ao vetor não-negativo da matriz componente do modo cultivares.

0,04	0	0	0	0	0	0	0
0	-7,10	0	0	0	0	0	0
0	3,93	0	0	0	0	0	1,38
0	0	0	-0,94	0	0	0	0
0	0	2,76	0	0	0	0	0
0	0	0	0	4,03	0	0	0
0	0	0	0	0	-6,04	0	0
0	0	0	0	0	0	-3,64	0

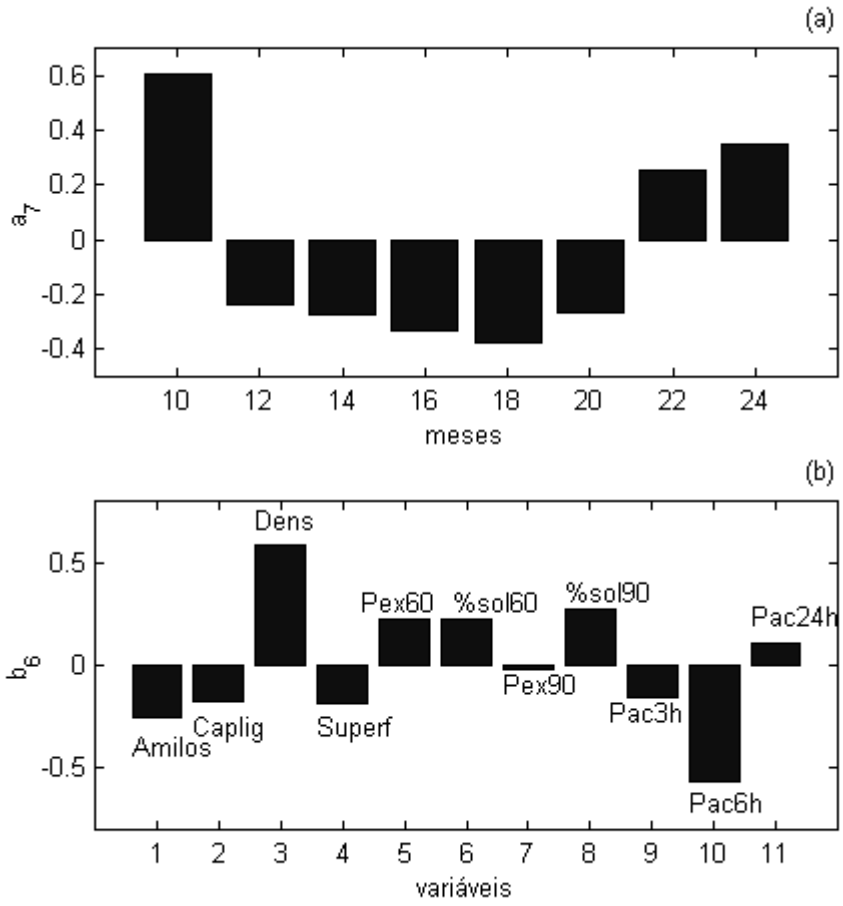
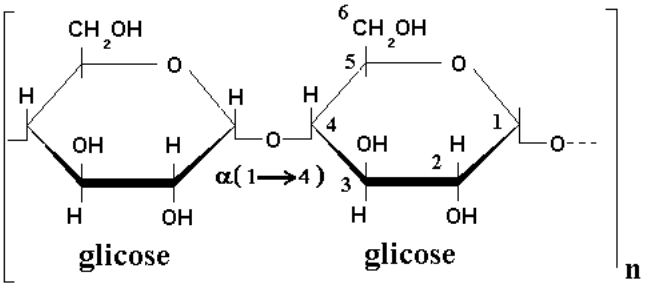


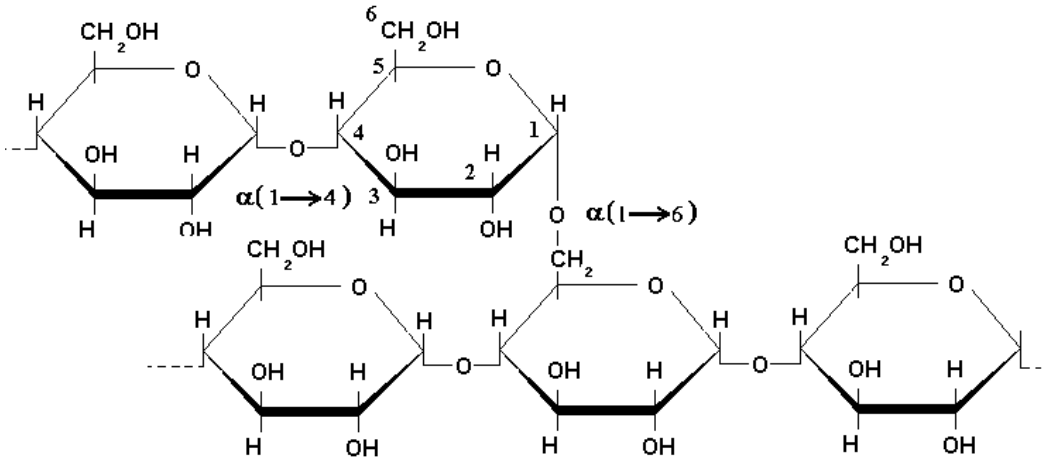
Figura 5.1 – Análise um. Valores dos vetores coluna não-negativo$\mathbf{a}_7$ em (a), para o modo da idade, e $\mathbf{b}_6$ em (b) para o modo das propriedades. (*Ver legenda no texto*).

O Primeiro Grupo corresponde ao período da seca quando a planta não produz ou produz apenas uma pequena quantidade de amido ou consome este componente para prover as necessidades fisiológicas da planta, pois ocorre a perda das folhas, seguida pela renovação da parte superior da planta. O Segundo Grupo representa a estação das chuvas, estágio onde ocorre intensa formação e acumulação de amido da raiz. É interessante notar a correlação entre os valores da idade e os valores do vetor coluna $\mathbf{a}_7$ no Segundo Grupo. A informação representada pelos vetores coluna $\mathbf{a}_7$ e $\mathbf{b}_6$ é única, pois apenas uma fatia do núcleo do MTR apresenta a coluna 6 e linha 7 com um único elemento diferente de zero.

Os grânulos de amidos são formados, principalmente, por amilose e amilopectina. A amilose é essencialmente um polímero linear de (1-4) α–D-glucose. A amilopectina é um polímero altamente ramificado, onde cadeias de (1-4) α–D-glucana são conectadas por ligações α (1-6), como ilustrado pela Figura 5.2. Estes dois biopolímeros podem se apresentar nos grânulos em diferentes conformações, comprimentos de cadeia e em diferentes proporções.



α-Amilose



Amilopectina

Figura 5.2 –Biopolímeros, α-Amilose e Amilopectina.

Ao aquecer uma suspensão de amido em água, os grânulos, aos poucos, começam a absorver água e a se expandir. Em uma faixa estreita de temperatura, todos os grânulos expandem irreversivelmente e são ditos gelatinizados. Como a temperatura da suspensão aquosa do amido foi aquecida acima da faixa de gelatinização, as ligações intramoleculares originárias de grupos hidroxilas continuam a se romper, moléculas de água passam a interagir com grupos hidroxilas, e os grânulos continuam a se expandir. O resultado direto da expansão do grânulo é o aumento da solubilidade*[Sarmento(a), Rickard]*.

O poder de expansão do amido, determinado com o aquecimento de uma amostra de amido previamente pesada a seco, em excesso de água, é definido como sendo o peso do sedimento expandido por grama de amido seco. A porcentagem de solúveis também pode ser determinada a partir da mesma solução através do sobrenadante *[Sarmento(a), Rickard]*. Segundo Leach *[Leach]*, o comprimento e o caráter micelar dentro dos grânulos é o fator mais importante na expansão do amido. A nível molecular, muitos fatores podem influenciar o grau e o tipo de associação entre os dois biopolímeros (*i.e.* amilose e amilopectina). Estes fatores incluem a razão da amilose em relação à amilopectina, a característica de cada fração em termos de peso/distribuição, grau/comprimento da ramificação e conformação destes polímeros. A presença de compostos diferentes dos carboidratos, como lipídeos, também é um fator importante. A formação de complexos amilose-lipídeo pode restringir a expansão e a solubilização. A diferença no ambiente de solubilidade nas raízes também pode estar atribuído à diferença na conformação dos componentes da amilose no grânulo nativo. Nas raízes, lipídeos, em geral, não ocorrem e a amilose ocorre em um estado amorfo sendo convertida pelo tratamento térmico a um estado helicoidal menos solúvel *[Sarmento(a), Rickard]*. Os valores dos elementos do vetor coluna $\mathbf{b}_6$ para poder de expansão e porcentagem de solúveis (*ver* Figura 5.1, *(b)*: Pex-60, %sol60, %sol90, poder de expansão a 60 graus Celsius, porcentagem de solúveis a 60 e 90 graus Celsius) apresentam correlação positiva com o Primeiro Grupo, ou seja, os valores dos elementos do vetor coluna $\mathbf{a}_7$ são positivos para o Primeiro Grupo (*i.e.* amostras na idade de 10, 22 e 24 meses) e também para o poder de expansão e porcentagem de solúveis nos elementos do vetor coluna $\mathbf{b}_6$. Em análise similar, é verificado que poder de expansão e porcentagem de solúveis apresentam correlação negativa com o Segundo Grupo (*i.e.*amostras com 12, 14, 16, 18 e 20 meses de idade). O sinal dos valores de poder de expansão e porcentagem de solúveis também são opostos ao do conteúdo de amilose no vetor coluna $\mathbf{b}_6$. Os valores dos elementos do vetor coluna $\mathbf{b}_6$ para porcentagem de solúveis sugerem que a diferença entre estes dois grupos de amostras, Primeiro e Segundo grupos, pode estar associada à conformação da amilose e não ao tamanho de sua porção, pois a porcentagem de solúveis e o conteúdo de amilose apresentam valores opostos, em resumo, a porcentagem de solúveis não deve aumentar com o aumento da porção de amilose no grânulo.

As moléculas de água, que interagem com as macromoléculas, são chamadas aqui de moléculas de água ligadas (seguindo a denominação encontrada na literatura deste tema) e refletem a habilidade da "superfície molecular" em formar ligações de hidrogênio com a água. A "superfície molecular" disponível para este tipo de ligação é reduzida em regiões onde ocorre um grande número de pontes de hidrogênio intramolecular. A quantidade de moléculas de água ligadas associadas com os grânulos de

amido influenciam as características de expansão dos grânulos. Assim, um alto valor de moléculas de água ligadas é atribuído à perda de associação entre os polímeros no grânulo do amido nativo. Os supostos sítios para ligação das moléculas de água são os grupos hidroxilas e os átomos de oxigênio das unidades de glicose. Durante a gelatinização, o número de sítios disponíveis para ligação das moléculas de água aumentam à medida que o calor rompe as pontes de hidrogênio intragranulares *[Sarmento(a), Rickard]*. Os valores dos elementos do vetor coluna $\mathbf{b}_6$ para a capacidade de ligação com água fria (*ver* Figura 5.1*(b)* Capilg) apresentam correlação positiva com o conteúdo de amilose e com o Segundo Grupo (*i.e.* amostras com idade de 12, 14, 16, 18 e 20 meses). Eles também apresentam correlação negativa com o Primeiro Grupo (*i.e.* amostras com idade de 10, 22 e 24 meses), poder de expansão e porcentagem de solúveis. Estas diferenças (correlação positiva do conteúdo de amilose com Segundo Grupo e capacidade de ligação com água fria e negativa com o Primeiro Grupo) entre os dois grupos de amostras pode ser devido ao tamanho da porção da amilose e à existência de diferentes números de sítios para ligação das moléculas de água, o que pode sugerir diferentes proporções de estados amorfos e cristalinos.

Os valores dos elementos do vetor coluna $\mathbf{b}_6$ correspondentes à área superficial específica do grânulo e a densidade absoluta do grânulo são opostos em sinal. A densidade absoluta do grânulo é correlacionada positivamente com o Segundo Grupo de amostras e área superficial específica do grânulo com o Primeiro Grupo. Área superficial específica do grânulo é dada por: 6/D[3,2], onde D[3,2] é um parâmetro proporcional ao diâmetro de uma esfera que possui a mesma área superficial da partícula analisada *[Sarmento(a)]*. Enfim, a área superficial específica do grânulo é inversamente proporcional ao tamanho do grânulo. Isto indica que, o Primeiro Grupo contém grânulos maiores e mais compactos se comparados ao Segundo Grupo.

A susceptibilidade dos grânulos de amido à digestão por glucoamilase é avaliada medindo a quantidade de unidades de glicose (*i.e.* açúcares redutores) produzida após um período de tempo sob o ataque enzimático. A glucoamilase quebra as ligações α (1-4) a partir das extremidades não redutoras dos polímeros, liberando moléculas de D-glucose na conformação β *[Sarmento(a)]*. As partes amorfas nos grânulos de amido são mais susceptíveis ao ataque enzimático, permitindo assim, o emprego da degradação enzimática para estudar a razão entre parte amorfa e cristalina dos grânulos de amido. Os valores dos elementos do vetor coluna $\mathbf{b}_6$ correspondentes à quantidade de açúcares redutores produzidos após 3 e 6 horas de ataque enzimático apresentam correlação positiva com o conteúdo de amilose, capacidade de ligação com água fria e área superficial específica do grânulo e

com o Segundo Grupo de amostras. A quantidade de açúcares produzida após 24 horas de ataque apresenta correlação positiva com poder de expansão (60) e porcentagem de solúveis (60,90) e com o Primeiro grupo de amostras.

Duas novas análises foram realizadas. Primeiro, as variáveis com informações sobre as propriedades da pasta foram incluídas. Para a segunda, as variáveis indicativas das propriedades da pasta e poder de expansão foram incluídas e a variável correspondente à quantidade de açúcares redutores após 24 horas foi excluída. Isto foi realizado para verificar se incluindo as variáveis, que também informam sobre propriedades semelhantes do amido, o modelo Tucker restrito ainda seria capaz de extrair o mesmo tipo de informação, isto nos dois sentidos: Incluindo e excluindo variáveis. Os resultados são mostrados nas Figuras 5.3 e 5.4.

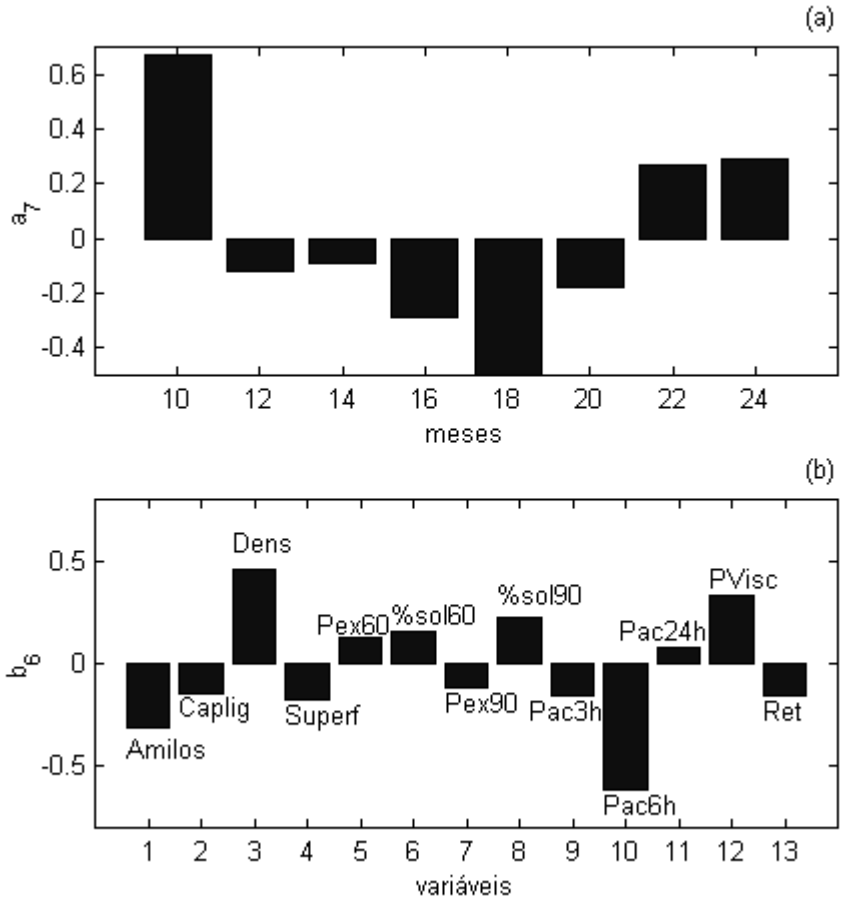


Figura 5.3 – Primeira análise após a inclusão de variáveis. Valores do vetor coluna não-negativo, $\mathbf{a}_7$ em (a) para o modo da idade, e $\mathbf{b}_6$ em (b) para o modo das propriedades. (*Ver legenda no texto*).

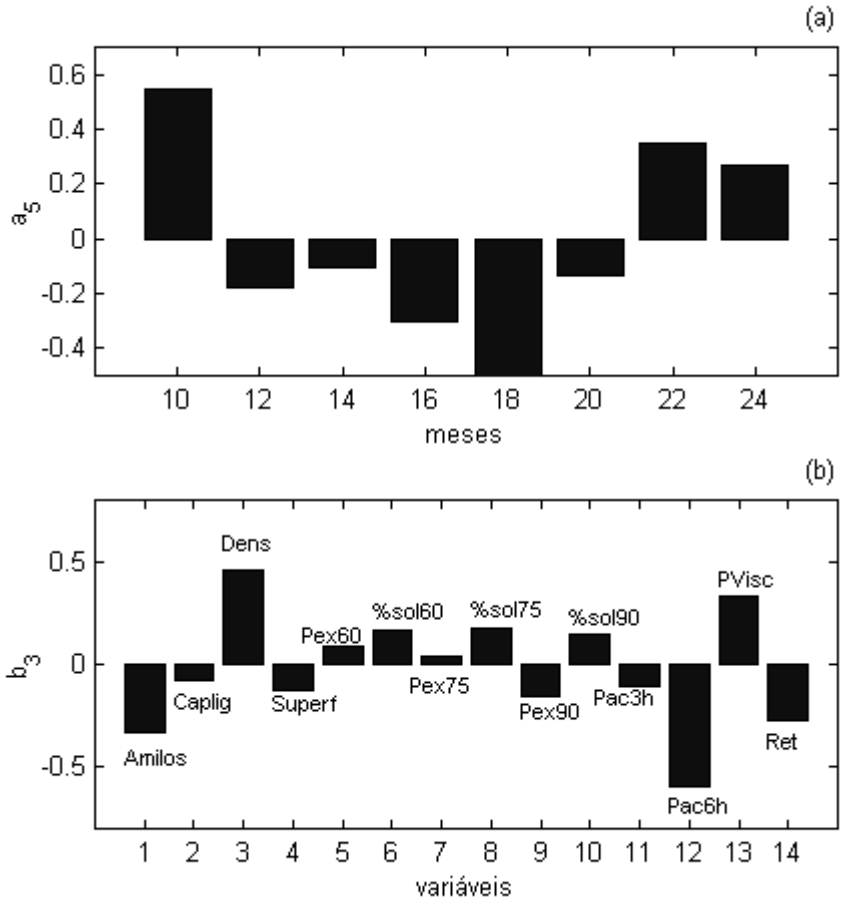


Figura 5.4 - Segunda análise após a inclusão de variáveis. Valores do vetor coluna não-negativo, $\mathbf{a}_7$ em (a) para o modo da idade, e $\mathbf{b}_6$ em (b) para o modo das propriedades. (*Ver legenda no texto*).

As propriedades da pasta mostram um alto valor para o elemento associado ao pico de viscosidade para amidos do Primeiro Grupo e correlação negativa com o conteúdo de amilose, capacidade de ligação com água fria, susceptibilidade enzimática e área superficial específica do grânulo. A tendência à retrogradação apresenta correlação negativa com o pico de viscosidade e positiva com o conteúdo de amilose, sendo o último reconhecido como o fator mais importante para esta propriedade.

Os valores dos elementos do vetor coluna $\mathbf{b}_6$ mostram uma correlação positiva entre capacidade de ligação com água fria (mais sítios disponíveis para a ligação de moléculas de água), produção de

açúcares redutores em um curto período (mais regiões susceptíveis ao ataque enzimático), área superficial específica do grânulo (menores grânulos) e conteúdo de amilose. Estas propriedades e composição indicam que os grânulos das amostras do Primeiro Grupo (para a estação da seca, *i.e.* 10, 22 e 24 meses de idade, *ver* Figura 5.1, 5.2 e 5.3 *(a)*) são mais compactos e com menos regiões amorfas. Por outro lado, as amostras do Segundo Grupo (estação das chuvas, *i.e.* amostras com 12, 14, 16, 18, e 20 meses de idade, *ver* Figura 5.1, 5.2 e 5.3 *(b)*) apresentam grânulos menos compactos com, provavelmente, mais áreas amorfas, que estão associadas à porção de amilose.

Deve ser notado também que, os valores do vetor coluna $\mathbf{a}_7$ apresentam um gradiente com relação à idade. Em resumo, o respectivo valor deste vetor para a idade de 10 meses é positivo, cai para a idade de 12 meses para um valor negativo e continua a decrescer até a idade de 18 meses quando volta a crescer se tornando positivo novamente para as idades de 22 e 24 meses. A precipitação pluviométrica média para o período *[Sarmento(a)]* e temperatura média mensal são apresentadas na Figura 5.5. Na Figura 5.6 são apresentadas as correlações entre precipitação pluviométrica e valores do vetor $\mathbf{a}_7$ e temperatura média mensal e $\mathbf{a}_7$. É possível verificar nesta figura, a existência de uma correlação significativa entre precipitação pluviométrica média e os valores do vetor coluna $\mathbf{a}_7$**,** mostrando que este vetor está relacionado aos efeitos sazonais.

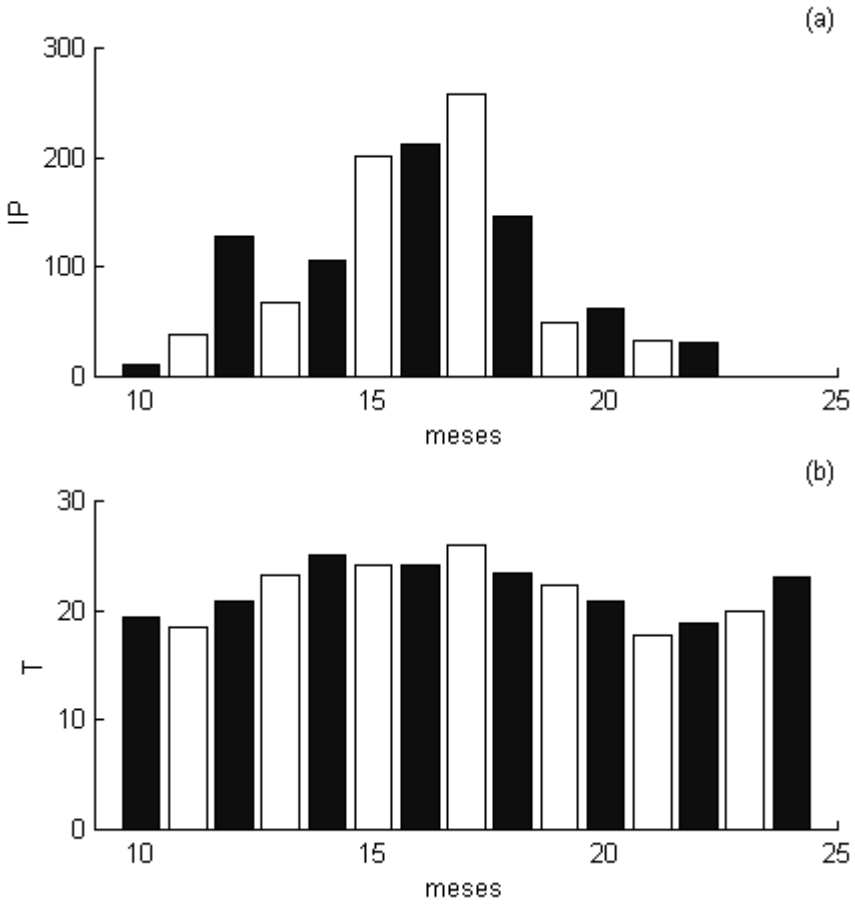


Figura 5.5 – PI-Precipitação pluviométrica, média mensal, para o período da colheita (i.e. Julho 93 - Setembro 94). T – temperatura média mensal (i.e. Julho 93 - Setembro 94). As barras em preto indicam os meses anteriores à colheita.

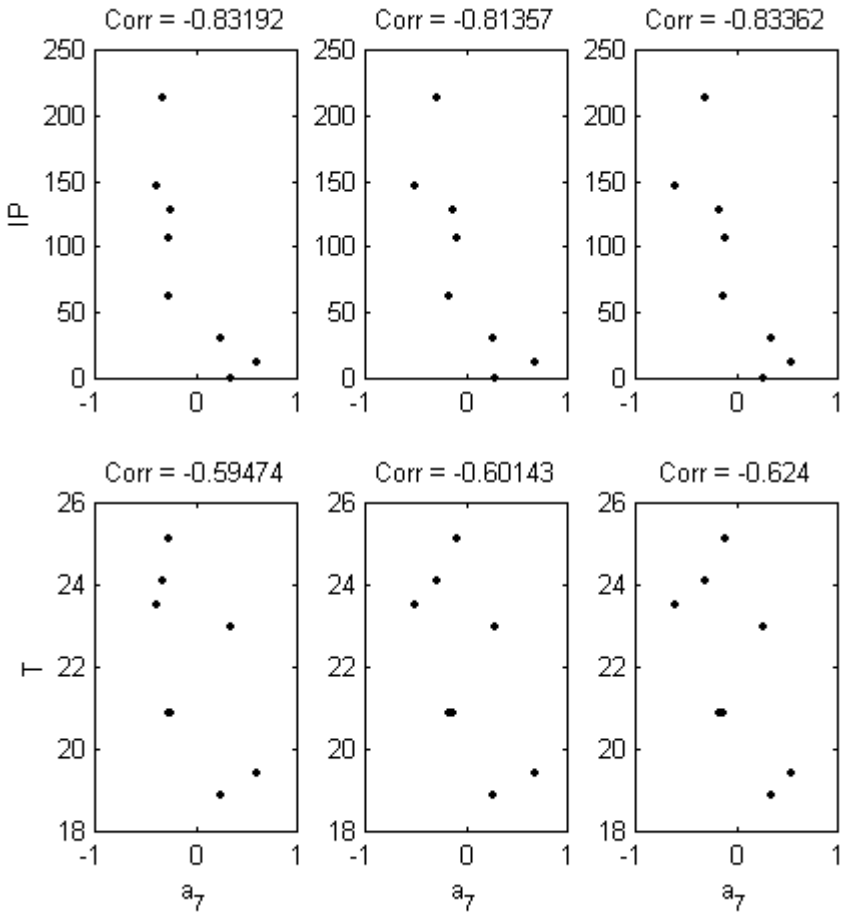


Figura 5.6 – Correlação entre os vetores $\mathbf{a}_7$ para as três análises (Figuras 5.1, 5.3 e 5.4) e precipitação pluviométrica média mensal, e temperatura média mensal, para o período da colheita (*i.e.* Julho 93 - Setembro 94). IP- precipitação pluviométrica média mensal. T- temperatura média mensal.

5.6 Conclusões

A metodologia proposta para análise dos dados de propriedades dos amidos das quatro cultivares mostrou que estas propriedades são susceptíveis às variações durante o longo período de colheita, estando estas variações mais propicias a ocorrerem devido às condições climáticas (*i.e.* temperatura e chuvas, principalmente) ou devido a estágios fisiológicos da planta do que à idade da planta. O

emprego desta metodologia ainda mostra que condições climáticas ou estado fisiológico agem sobre as propriedades do amido.

A correlação entre propriedades do amido, conformação dos polímeros, idade e fatores genéticos devido às diferentes cultivares é um complicado quebra cabeças. Por outro lado, este conjunto de dados (idade×propriedades×cultivares) pode ser arranjado com uma estrutura de arranjo em três modos e analisado como tal.

A análise de valores singulares destes dados mostra que não há redução do posto, podendo até ser questionado que a análise de fatores, mesmos em três modos, não seja interessante. Por outro lado, a análise em três modos, conduzida para a extração de um tipo de informação, neste caso, aquelas variações das propriedades dos amidos que são similares para as quatro cultivares (ou seja, eliminando as variações devido às diferenças genéticas), mostrou ser possível e produziu resultados muito importantes do ponto de vista agronômico. Nesta análise, supôs se que a informação similar entre os amidos das quatro cultivares foi capturada em uma estrutura em três modos chamada de bloco. Esta estrutura, bloco, foi aproximada através de um modelo Tucker restrito. Na formulação deste modelo foi considerada a independência entre blocos. Funções de inércia foram empregadas para avaliar os modelos, verificando como o conjunto de dados foi descrito por eles.

O ponto interessante desta metodologia é que os vetores colunas das matrizes componentes que descrevem a correlação entre idade e propriedades podem ser relacionados diretamente em um bloco, facilitando a análise dos dados.

Os resultados desta análise podem ser interessantes na escolha da época de colheita, pois relacionam as características do amido com a idade em que a planta foi colhida.

5.7 Apêndice 5.1

A proposta deste apêndice é mostrar como a função $f\left(G_k, c_{qk}, \mathbf{X}_q\right)$ pode resultar em valores negativos. Para tal, a função $f\left(\mathbf{G}_{G-qk}\right)$ é estudada, pois para os casos onde a inequação: $f\left(\mathbf{G}_{G-qk}\right) \leq vec\mathbf{X}_q^T vec\mathbf{X}_q$ é válida, a função $f\left(G_k, c_{qk}, \mathbf{X}_q\right)$ é não-negativa.

$$f\left(G_k, c_{qk}, \mathbf{X}_q\right) = \left(1 - \frac{\left\| vec\mathbf{X}_q - \left[(\mathbf{A}\otimes\mathbf{B})vec\mathbf{G}_k c_{qk}\right]\right\|^2}{vec\mathbf{X}_q^T vec\mathbf{X}_q}\right)\cdot 100\% = \left(1 - \frac{f\left(\mathbf{G}_{G-qk}\right)}{vec\mathbf{X}_q^T vec\mathbf{X}_q}\right)\cdot 100\% \quad 171$$

onde $vec\mathbf{G}_k$ é a coluna k da matriz $vec\underline{\mathbf{G}}$ e c_{qk} é o elemento k da linha q da matriz componente **C**.

Considere a função $f\left(\mathbf{G}_{G-qk}\right)$ usada para "medir" como a fatia q do arranjo de dados é descrita pela fatia k do núcleo.

$$f\left(\mathbf{G}_{G-qk}\right) = \left(\left\{vec\mathbf{X}_q - \left[(\mathbf{A}\otimes\mathbf{B})vec\mathbf{G}_k c_{qk}\right]\right\}^T \left\{vec\mathbf{X}_q - \left[(\mathbf{A}\otimes\mathbf{B})vec\mathbf{G}_k c_{qk}\right]\right\}\right), \quad 172$$

$$\mathbf{W_1} = vec\mathbf{X}_q^T vec\mathbf{X}_q - vec\mathbf{X}_q^T(\mathbf{A}\otimes\mathbf{B})vec\mathbf{G}_k c_{qk} - \left[(\mathbf{A}\otimes\mathbf{B})vec\mathbf{G}_k c_{qk}\right]^T vec\mathbf{X}_q \quad 173$$

$$\mathbf{W}_2 = \left[(\mathbf{A}\otimes\mathbf{B})vec\mathbf{G}_k c_{qk}\right]^T \left[(\mathbf{A}\otimes\mathbf{B})vec\mathbf{G}_k c_{qk}\right] \quad 174$$

$$f\left(\mathbf{G}_{G-qk}\right) = \mathbf{W}_1 + \mathbf{W}_2 \quad 175$$

Considerando que a fatia q corresponde a:

$$vec\mathbf{X}_q = (\mathbf{A}\otimes\mathbf{B})vec\underline{\mathbf{G}}\left({}^q\mathbf{c}\right)^T + \mathbf{e} \quad 176$$

e usando:

$$n = c_{qk} vec\mathbf{G}_k^T\left(\mathbf{A}^T\otimes\mathbf{B}^T\right)\mathbf{e} \quad 177$$

onde $({}^q\mathbf{c})$ indica a linha q da matriz componente **C** é a parte não modelada de $vec\mathbf{X}_q$.

A função $f(\mathbf{G}_{G-qk})$ pode ser reescrita como se segue:

$$\mathbf{W}_1 = \left(vec\mathbf{X}_q^T vec\mathbf{X}_q\right) - 2\left[\left({}^q\mathbf{c}\right)vec\underline{\mathbf{G}}^T vec\mathbf{G}_k c_{qk}\right] - 2n \quad 178$$

considerando a ortogonalidade imposta à matrizes componentes, **A**, **B** e **C**:

$$\left(\mathbf{A}^T\otimes\mathbf{B}^T\right)(\mathbf{A}\otimes\mathbf{B}) = \mathbf{I} \quad 179$$

e

$$vec\underline{\mathbf{G}}^{T} vec\underline{\mathbf{G}} = \mathbf{\Lambda} \tag{180}$$

onde $\mathbf{\Lambda}$ é uma matriz diagonal, pois:

$$vec\underline{\mathbf{X}} = (\mathbf{A} \otimes \mathbf{B}) vec\underline{\mathbf{G}}\mathbf{C}^{T} \tag{181}$$

e então

$$(\mathbf{A}^{T} \otimes \mathbf{B}^{T}) vec\underline{\mathbf{X}} = vec\underline{\mathbf{G}}\mathbf{C}^{T} \tag{182}$$

onde as colunas $vec\underline{\mathbf{G}}$ são ortogonais devido à ortogonalidade das colunas de **C** *(ver Magnus em " one mode component analysis")*. No caso do modelo Tucker restrito $\mathbf{\Lambda}$ é diagonal devido à restrição imposta ao núcleo.(*Para descrição da* expressão **181** *ver [Kiers(a), p.453] e seção fundamentos*).

Desta forma, tem-se:

$$vec\underline{\mathbf{G}}^{T} vec\mathbf{G}_k = \begin{pmatrix} 0 \\ \vdots \\ vec\mathbf{G}_k^{T} vec\mathbf{G}_k \\ \vdots \\ 0 \end{pmatrix} \tag{183}$$

$$\mathbf{W}_1 = (vec\mathbf{X}_q^{T} vec\mathbf{X}_q) - 2c_{qk}^{2} vec\mathbf{G}_k^{T} vec\mathbf{G}_k - 2n \tag{184}$$

$$\mathbf{W}_2 = c_{qk}^{2} vec\mathbf{G}_k^{T} vec\mathbf{G}_k \tag{185}$$

$$f(\mathbf{G}_{G-qk}) = (vec\mathbf{X}_q^{T} vec\mathbf{X}_q) - c_{qk}^{2} vec\mathbf{G}_k^{T} vec\mathbf{G}_k - 2n \tag{186}$$

ou seja, para $f(\mathbf{G}_{G-qk}) \leq vec\mathbf{X}_q^{T} vec\mathbf{X}_q$ ser válida o escalar 2*n*, se negativo, deve ser menor ou igual a ao termo positivo $c_{qk}^{2} vec\mathbf{G}_k^{T} vec\mathbf{G}_k$, indicando que a parte não modelada da fatia *q* tem pouca importância se comparada àquela descrita pela fatia *k* do núcleo. Naqueles casos onde a parte não modelada dos dados é importante $f(\mathbf{G}_{G-qk}) \leq vec\mathbf{X}_q^{T} vec\mathbf{X}_q$ não é válida resultando em valores negativos para $f(G_k, c_{qk}, \mathbf{X}_q)$. Este fato é importante na construção do modelo Tucker restrito.

Para o modelo Tucker com rotação $f\left(\mathbf{G}_{G-qk}\right)$ se torna $f\left(\widetilde{\mathbf{G}}_{\widetilde{G}\mathbf{U}-qk}\right)$ onde "~" se refere ao modelo com rotação. De forma semelhante àquela feita para $f\left(\mathbf{G}_{G-qk}\right)$, tem-se:

$$\widetilde{\mathbf{W}}_1 = \left(vec\mathbf{X}_q^T vec\mathbf{X}_q\right) - 2\left[\left({}^q\widetilde{\mathbf{c}}\right)vec\widetilde{\mathbf{G}}_q^T vec\widetilde{\mathbf{G}}_k\widetilde{c}_{qk}\right] - 2\widetilde{n} \tag{187}$$

$$\widetilde{\mathbf{W}}_1 = \left(vec\mathbf{X}_q^T vec\mathbf{X}_q\right) - 2\left[\left({}^q\mathbf{c}\right)\mathbf{U}\mathbf{U}^T vec\underline{\mathbf{G}}^T vec\underline{\mathbf{G}}\mathbf{u}_k\left({}^q\mathbf{c}\right)\mathbf{u}_k\right] - 2\widetilde{n} \tag{188}$$

ver expressão **170**

$$\widetilde{\mathbf{W}}_1 = \left(vec\mathbf{X}_q^T vec\mathbf{X}_q\right) - 2\left[\left({}^q\mathbf{c}\right)\mathbf{\Lambda}\mathbf{u}_k\left({}^q\mathbf{c}\right)\mathbf{u}_k\right] - 2\widetilde{n} \tag{189}$$

$$\widetilde{\mathbf{W}}_1 = \left(vec\mathbf{X}_q^T vec\mathbf{X}_q\right) - 2\left[\left({}^q\mathbf{c}\right)\mathbf{\Lambda}\mathbf{u}_k\mathbf{u}_k^T\left({}^q\mathbf{c}^T\right)\right] - 2\widetilde{n} \tag{190}$$

$$\widetilde{\mathbf{W}}_2 = \left({}^q\mathbf{c}\right)\mathbf{u}_k\mathbf{u}_k^T\mathbf{\Lambda}\mathbf{u}_k\mathbf{u}_k^T\left({}^q\mathbf{c}^T\right) \tag{191}$$

$$f\left(\widetilde{\mathbf{G}}_{\widetilde{G}\mathbf{U}-qk}\right) = \left(vec\mathbf{X}_q^T vec\mathbf{X}_q\right) - 2\left[\left({}^q\mathbf{c}\right)\mathbf{\Lambda}\mathbf{u}_k\mathbf{u}_k^T\left({}^q\mathbf{c}^T\right)\right] - 2\widetilde{n} + \left({}^q\mathbf{c}\right)\mathbf{u}_k\mathbf{u}_k^T\mathbf{\Lambda}\mathbf{u}_k\mathbf{u}_k^T\left({}^q\mathbf{c}^T\right) \tag{192}$$

$$f\left(\widetilde{\mathbf{G}}_{\widetilde{G}\mathbf{U}-qk}\right) = \left(vec\mathbf{X}_q^T vec\mathbf{X}_q\right) + \left[\left({}^q\mathbf{c}\right)\mathbf{\Omega}\mathbf{\Lambda}\mathbf{u}_k\mathbf{u}_k^T\left({}^q\mathbf{c}^T\right)\right] - 2\widetilde{n} + \left({}^q\mathbf{c}\right)\mathbf{u}_k\mathbf{u}_k^T\mathbf{\Lambda}\mathbf{u}_k\mathbf{u}_k^T\left({}^q\mathbf{c}^T\right) \tag{193}$$

onde

$$\mathbf{\Omega} = \begin{pmatrix} -2 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix}$$

$$f\left(\widetilde{\mathbf{G}}_{\widetilde{G}\mathbf{U}-qk}\right) = \left(vec\mathbf{X}_q^T vec\mathbf{X}_q\right) + \left({}^q\mathbf{c}\right)\left[\mathbf{\Omega} + \mathbf{u}_k\mathbf{u}_k^T\right]\mathbf{\Lambda}\mathbf{u}_k\mathbf{u}_k^T\left({}^q\mathbf{c}^T\right) - 2\widetilde{n} \tag{194}$$

assim, para $f\left(\tilde{\mathbf{G}}_{\tilde{G}\mathbf{U}-qk}\right) \leq vec\mathbf{X}_q^T vec\mathbf{X}_q$ ser válida, existe a dependência na forma de como o modelo é rodado e não apenas como o modelo descreve a fatia *q* do arranjo de dados. Um exemplo disto é apresentado a seguir, ou seja, o cálculo do elemento na coluna 3, linha 2 da Tabela 5.3 (***-1,3967***) correspondente à descrição da fatia 2 do arranjo de dados pela fatia 3 do núcleo, ou seja, *q=2* e *k=3*.

$$\mathbf{u}_3^T = \begin{pmatrix} 0{,}056887345 & 0{,}014471139 & 0{,}24262594 & -0{,}96834243 \end{pmatrix}$$

$$\left[\mathbf{\Omega} + \mathbf{u}_3\mathbf{u}_3^T\right]\mathbf{\Lambda}\mathbf{u}_3\mathbf{u}_3^T = \begin{pmatrix} -0{,}8984 & -0{,}2285 & -3{,}8315 & 15{,}2921 \\ -0{,}512 & -0{,}0385 & -0{,}6450 & 2{,}5743 \\ -0{,}6746 & -0{,}1716 & -2{,}8773 & 11{,}4836 \\ 1{,}4826 & 0{,}3772 & 6{,}3235 & -25{,}2377 \end{pmatrix}$$

$${}^q\mathbf{c} = \begin{pmatrix} 0{,}4636 & -0{,}2014 & 0{,}7529 & 0{,}4215 \end{pmatrix}$$

$$\left({}^q\mathbf{c}\right)\left[\mathbf{\Omega} + \mathbf{u}_k\mathbf{u}_k^T\right]\mathbf{\Lambda}\mathbf{u}_k\mathbf{u}_k^T\left({}^q\mathbf{c}^T\right) = 0{,}9552$$

$$vec\mathbf{X}_2^T vec\mathbf{X}_2 = 68{,}3922$$

$$f\left(G_3, c_{23}, \mathbf{X}_2\right) = \left(1 - \frac{68{,}3922 + 0{,}9552}{68{,}3922}\right)\cdot 100\% = -1{,}3967$$

6 PARAFAC com Splines: Um estudo de Caso

6.1 Introdução

Os métodos multi modos, que surgiram na psicometria *[Kiers(g), Tucker]*, tem recebido atenção nas últimas décadas no âmbito da quimiometria. Este tipo de método é apropriado para análise de dados estruturados, a cada dia mais comum na química e áreas afim, devido ao desenvolvimento de instrumentos hifenados (*e.g.* LC-UV, GC-MS, MS-MS). Em geral, os métodos quimiométricos se baseiam na decomposição da matriz de dados em variáveis latentes, como por exemplo, a Análise de Componentes Principais (*Principal Component Analysis* PCA) onde a matriz de dados **X** é decomposta nas matrizes de escores e *loadings* (*i.e.* $\mathbf{X}=\mathbf{T}\mathbf{P}^{T}$, onde o sobrescrito indica a operação de transposição sobre a matriz **P**). Neste sentido, métodos em multi modos podem ser considerados uma extensão de métodos como o PCA para arranjos multi modos de dados, pois o mesmo tipo de análise dos dados é usado.

Os métodos em multi modos desenvolvidos na psicologia, especialmente o PARAFAC e modelos Tucker, têm sido usados em aplicações na química como calibração de segunda ordem, separação de curvas e análise exploratória. Para tal, estes métodos tem recebido alguns *"refinamentos"*, que são necessários devido à presença de variações no conjunto de dados que não podem ser acomodadas pelos modelos empregados, como por exemplo, ruídos. Em separação de curvas, por exemplo, o objetivo de um método em multi modos é decompor um conjunto de perfis sobrepostos(*ver seção de separação de espectros*). No caso de espectros e perfis de tempo sobrepostos, por exemplo, a restrição de não-negatividade pode ser usada no processo de decomposição, pois é sabido que os espectros e perfis de tempo são não-negativos. Para o caso de perfis cromatográficos, podem ser usadas restrições de não-negatividade e unimodalidade, por exemplo. Suavidade (*dos inglês "Smoothness*"), pode ser necessária naqueles casos onde os dados apresentam uma grande variação local e a análise requer perfis de contorno suave. Este tipo de restrição também pode ser aplicado na decomposição promovida pelos métodos em multi modos. No contexto de análises em multi modos Bro *[Bro(d)]* aplicou PARAFAC com a restrição de suavidade, baseada em uma penalização, para a separação de curvas de dados de fluorescência.

Splines, historicamente, são originários da engenharia naval, onde eram usados para desenhar curvas entre pontos especificados, tornando-se mais tarde um termo matemático, consistindo na solução de um problema de otimização sob restrição. Os splines originais são interpolatórios por natureza. Embora este tipo de splines seja muito útil para dados sem ruídos, eles apresentam uso

limitado em dados experimentais. Por outro lado, existe um tipo de spline de suavização (*do inglês smoothing*) que pode descrever dados, sendo restritos a não interpolarem estes dados, e se tornando desta forma, valiosa ferramenta para dados experimentais *[Wegman]*. Ramsay e Silverman têm mostrado a importância de análise funcional aplicada a métodos similares ao PCA através de splines de suavização *[Ramsay, Besse, Silverman]*. Em quimiometria, splines tem sido usados desde o ajuste de curvas *[Wold]*, compressão de dados *[Alsberg]*, até linearização de problemas não lineares *[Ferreira(b)]*.

O conjunto de dados usados nesta parte do trabalho de tese corresponde à medidas da concentração de monóxido de carbono na cidade de São Paulo. Estas medidas foram efetuadas a cada hora para todos os dias ao longo de um ano. Estes dados foram arranjados em uma estrutura em três modos "Horas do Dia × Dias da Semana × Semanas do Ano (HD×DS×SA)". Esta estrutura foi sugerida para identificar o perfil da variação sistemática associada ao tráfego de veículos automotores, considerado a maior fonte deste gás, e os efeitos sazonais devido à mudanças climáticas. Neste caso, os efeitos sazonais a serem identificados são aqueles relacionados às diferentes estações ao longo do ano, ou seja, aquelas variações ocorridas em cada semana, aqui são consideradas como variação rápida local, não sendo de interesse deste trabalho descreve-las. Assim, o propósito desta parte do trabalho de tese é avaliar a utilização de splines de suavização na decomposição efetuada pelo PARAFAC para efetuar uma análise exploratória deste conjunto de dados.

Este tipo de dados também foi analisado por Paatero *[Paatero]* através de seu método para um modelo PARAFAC.

Em resumo, o objetivo desta parte do trabalho de tese é identificar o perfil de variação sistemática devido ao tráfego de veículos automotores e os efeitos sazonais ao longo do ano (aqueles que variam suavemente ao longo do ano), e verificar a utilidade de se combinar splines de suavização através da aproximação usada por Bro para as restrições de unimodalidade *[Bro(d)]*.

6.2 Dados

Os dados usados nesta parte do trabalho de tese são medidas da concentração de monóxido de carbono na cidade de São Paulo coletadas a cada hora, todos os dias ao longo de do ano, isto feito, em um único ponto de amostragem. Dados de dois anos são estudados neste trabalho, isto é, dados para os anos de 1997 e 1999. Os dados originais poderiam ser arranjados como um arranjo em três modos com dimensões de 24×7×52 (*i.e.* 24 horas de um dia×7 dias de uma semana×52 semanas ao longo do ano),

o que será representado aqui como (HD×DS×SA), entretanto, em trabalho anterior foi verificado que os finais de semana apresentam uma pequena contribuição daquela variação sistemática devido ao tráfego de veículos automotores encontrada para os dias da semana *[Barcellos]*. Assim, apenas os cinco dias da semana foram usados, de segunda a sexta, resultando em um arranjo em três modos de 24×5×52 (*i.e.* 24 horas de um dia×5 dias de uma semana, de segunda a sexta×52 semanas ao longo do ano).

6.3 Métodos

O PARAFAC é baseado na decomposição dos dados em arranjo em multi modos em uma combinação linear de componentes multilineares*[Harshman, Bro(b)]*. Esta decomposição também pode ser restrita (por exemplo, restrição de não-negatividade, restrição de unimodalidade, de suavidade, etc.). Os métodos avaliados neste trabalho são usados para impor a restrição de suavidade aos perfis, ou componentes, determinados na decomposição pelo PARAFAC e também para impor periodicidade quando necessário.

Antes de descrever os objetos funcionais a serem usados para impor suavidade e/ou periodicidade é preciso descrever o PARAFAC e como estas restrições são impostas na decomposição. A expressão **152** descreve a decomposição de uma arranjo em três modos pelo PARAFAC.

$$\underline{\mathbf{X}} = \mathbf{A}\underline{\mathbf{I}_{DS}}\left(\mathbf{C}^T \otimes \mathbf{B}^T\right) + \mathbf{E} \qquad 195$$

onde $\underline{\mathbf{I}_{DS}}$ ($F\times[F\cdot F]$) e $\underline{\mathbf{X}}$ ($M\times[N\cdot R]$) correspondem à forma matricial do arranjo diagonal superior e do arranjo do dados em três modos, respectivamente (*ver seção fundamentos*). A matriz $\underline{\mathbf{X}}$ é construída pela justaposição horizontal de *R* matrizes de dimensões (*M*×*N*), que são chamadas *fatias*, por exemplo, 52 matrizes de dimensões (24 horas ×5 dias). *F* representa o número de componentes trilineares decompostos pelo PARAFAC. O arranjo diagonal superior em três modos é construído de forma semelhante à $\underline{\mathbf{X}}$ onde cada fatia corresponde a uma matriz quadrada *F*, que possui apenas um elemento diferente de zero e igual a um, isto é, o elemento *f*,*f* da diagonal da matriz quadrada *F*, sendo *f* o número da fatia. **A**, **B**, e **C** são as matrizes componentes de dimensões (*M*×*F*), (*N*×*F*) e (*R*×*F*), respectivamente *[Bro(c)]*. **E** representa a parte do conjunto de dados que não pode ser descrita por um modelo trilinear.

A estimativa dos parâmetros do modelo PARAFAC, **A**, **B**, e **C**, através da expressão **152** pode ser encontrado através de um algoritmo por Quadrados Mínimos Alternantes "QMA" onde as matrizes componentes **A**, **B**, e **C** são determinadas em cada etapa do processo de otimização. Esta formulação é

denominada por Bro como Global *[Bro(d)]*(*ver seção fundamentos*). Outra aproximação possível sugere a determinação de **A**, **B** e **C** através de uma formulação no subespaço das coluna *[Bro(d)]*, onde a expressão **152** é primeiro reescrita em termos de **196** *[Bro(c),(d)]*.

$$\mathbf{X}_A^T = \mathbf{Z}_A\mathbf{A}^T + \mathbf{E} = \mathbf{z}_{A,1}\mathbf{a}_1^T + \mathbf{z}_{A,2}\mathbf{a}_2^T + \cdots + \mathbf{z}_{A,f}\mathbf{a}_f^T + \cdots + \mathbf{z}_{A,F}\mathbf{a}_F^T + \mathbf{E} \qquad 196$$

onde

$$\mathbf{Z}_A = \left[\underline{\mathbf{I}_{DS}}\left(\mathbf{C}^T \otimes \mathbf{B}^T\right)\right]^T = \left(\mathbf{z}_{A,1}\middle|\quad \mathbf{z}_{A,2}\middle|\quad \cdots\middle|\quad \mathbf{z}_{A,F}\middle|\right) \qquad 197$$

e $\mathbf{X}_A$, neste caso, é igual a matriz $\underline{\mathbf{X}}$, sendo o subscrito "*A*" usado para indicar que a matriz componente **A**, não participa diretamente do produto tensorial $(\cdots\otimes\cdots)$.

A decomposição global pelo PARAFAC é obtida em uma otimização por QMA sendo a função descrita pela expressão **198** minimizada.

$$l\left(\mathbf{z}_{A,1\cdots i}, \mathbf{a}_{1\cdots i}\right) = \left\|\mathbf{X}_A^T - \left(\mathbf{z}_{A,1}\mathbf{a}_1^T + \mathbf{z}_{A,2}\mathbf{a}_2^T + \cdots + \mathbf{z}_{A,f}\mathbf{a}_f^T + \cdots + \mathbf{z}_{A,F}\mathbf{a}_F^T\right)\right\|^2 \qquad 198$$

A formulação através do subespaço coluna é encontrada reescrevendo a expressão **198** com as expressões **199** e **200**, onde a determinação do componente multilinear f é obtida através da expressão **201**, neste caso, para o modo *A*.

$$l\left(\mathbf{a}_f\right) = \left\|\left(\mathbf{X}_A^T - \mathbf{z}_{A,1}\mathbf{a}_1^T - \mathbf{z}_{A,2}\mathbf{a}_2^T - \cdots - \mathbf{z}_{A,F}\mathbf{a}_F^T\right) - \mathbf{z}_{A,f}\mathbf{a}_f^T\right\|^2 \qquad 199$$

$${}^{(-f)}\mathbf{Y} = \mathbf{X}_A^T - \mathbf{z}_{A,1}\mathbf{a}_1^T - \mathbf{z}_{A,2}\mathbf{a}_2^T - \cdots - \mathbf{z}_{A,F}\mathbf{a}_F^T \qquad 200$$

$$l_f\left(\mathbf{a}_f\right) = \left\|{}^{(-f)}\mathbf{Y} - \mathbf{z}_{A,f}\mathbf{a}_f^T\right\|^2, \qquad 201$$

e assim para todo *f* a expressão **201** é resolvida.

Bro *[Bro(c),(d)]* mostra que a solução do problema de minimizar a função descrita em **201** quando $\mathbf{a}_f$ esta sob restrição é equivalente ao seguinte problema:

$$\underset{\mathbf{a}_f}{\text{minimize}} \left\|\boldsymbol{\beta}_f - \mathbf{a}_f\right\|^2 \qquad 202$$

sujeito a $\mathbf{a}_f$ restrito

onde $\boldsymbol{\beta}_f$ é a solução por quadrados mínimos sem restrição que minimiza a função dada na expressão **201**.

6.3.1 A restrição funcional

A solução por quadrados mínimos da expressão **201** sem restrições pode resultar em componentes que descrevem apenas parte do fenômeno estudado. Isto pode ocorrer devido à parte dos dados que não segue um modelo multilinear e que pode afetar o processo de otimização (por exemplo, dados com ruídos). Em alguns casos, variação rápida e local não é desejada e aos componentes decompostos pelo PARAFAC é requerida uma variação suave. Produzir uma boa descrição dos dados e evitar variações rápidas e locais pode ser obtida pela aplicação de regularização (suavização) aos componentes decompostos. Um método simples de aplicar tal regularização é representar cada componente através de uma combinação linear de um número limitado de funções de base. Outro método, é medir a variação rápida e local através de parâmetro de rugosidade *[Ramsay, Silverman(a)]*. Nesta parte do trabalho de tese, o parâmetro de rugosidade é o quadrado da segunda derivada integrado, que pode ser introduzido na expressão **202**, como mostrado em **203**, para suavizar o vetor coluna f da matriz componente.

$$l_{f\lambda} = \sum_t \left(\beta_{f,t} - \alpha_{f,t}\right)^2 + \lambda \int \left(g_f^{''}\right)^2 dt \qquad 203$$

onde

$$\boldsymbol{\beta}_f^T = \left(\mathbf{z}_f^T \mathbf{z}_f\right)^{-1} \mathbf{z}_f^{T\,(-f)} \mathbf{Y}, \qquad 204$$

$$\boldsymbol{\alpha}_f = \left(g_f(t_1)\, g_f(t_2) \cdots g_f(t_M)\right)^T, \qquad 205$$

e $\boldsymbol{\alpha}_f$ indica o vetor coluna f da matriz componente correspondendo aos valores da função $g_f(t)$ calculados em t pontos para um dado intervalo. O termo $\int \left(g_f^{''}\right)^2 dt$ da expressão **203** é responsável pela curvatura da função $g_f(t)$ (ou taxa de troca entre o erro residual e a variação local). Neste sentido, a alteração dos valores de λ altera o valor de $\int \left(g_f^{''}\right)^2 dt$ e conseqüentemente a curvatura da função $g_f(t)$ é ajustada *[Silverman(a)]*, ou melhor, a função a ser encontrada deve possuir uma curvatura que se ajuste ao valor de λ para compensar o processo de minimização. O sobrescrito ‘ $''$ ’ indica a segunda derivada de $g_f(t)$.

A solução da expressão **203** é discutida no Apêndice 6.1 e também pode ser encontrada na literatura *[Ramsay,Hastie, Green]*.

A função $g_f(t)$ empregada para representar os vetores coluna das matrizes componentes são expansões em funções de base, ou seja, uma combinação linear de funções de base, neste caso, B-splines e séries de Fourier. A função $g_f(t)$ representada por funções B-spline é dada pela expressão **206** (*ver as referências [Wold, Wegman, Hastie] para introdução e de Boor[de Boor] para detalhes técnicos*).

$$g_f(t) = \sum_{j=1}^{n_{bases}} \tau_j Q_j(t) = \mathbf{Q}\boldsymbol{\tau} \quad , \qquad 206$$

onde $Q_j(t)$ é a função de base j disposta na coluna j da matriz $\mathbf{Q}$, $\boldsymbol{\tau}$ é o vetor de coeficientes da combinação linear das funções de base, n_{bases} é o número de funções de base. A seqüência de nós é denotada por $\xi_\sigma ; \sigma = 1,2,\cdots,n_{nos}$ mais os nós de extremos: ξ_0 e $\xi_{\sigma+1}$, onde n_{nos} é o número de nós.

A descrição da função $g_f(t)$ em termos da série de Fourier é dada pela expressão **207**

$$g_f(t) = c_0 + c_1 sin(\varpi t) + c_2 \cos(\varpi t) + c_3 sin(2\varpi t) + c_4 \cos(2\varpi t) + \cdots \quad , \qquad 207$$

que pode ser descrita como

$$g_f(t) = \boldsymbol{\varphi}\boldsymbol{\gamma} \qquad 208$$

onde as colunas da matriz de funções de base $\boldsymbol{\varphi}$ são as funções:

$\phi_0(t) = 1$, $\phi_{2r-1}(t) = sin(r\varpi t)$, $\phi_{2r}(t) = \cos(r\varpi t)$ e o parâmetro ϖ determina o período $2\pi/\varpi$, que é igual ao comprimento do intervalo T onde $g_f(t)$ é periódica *[Ramsay, Silverman(b)]*. $\boldsymbol{\gamma}$ é o vetor de coeficientes.

A Figura 6.1 apresenta exemplos de funções de base usadas nesta parte do trabalho de tese.

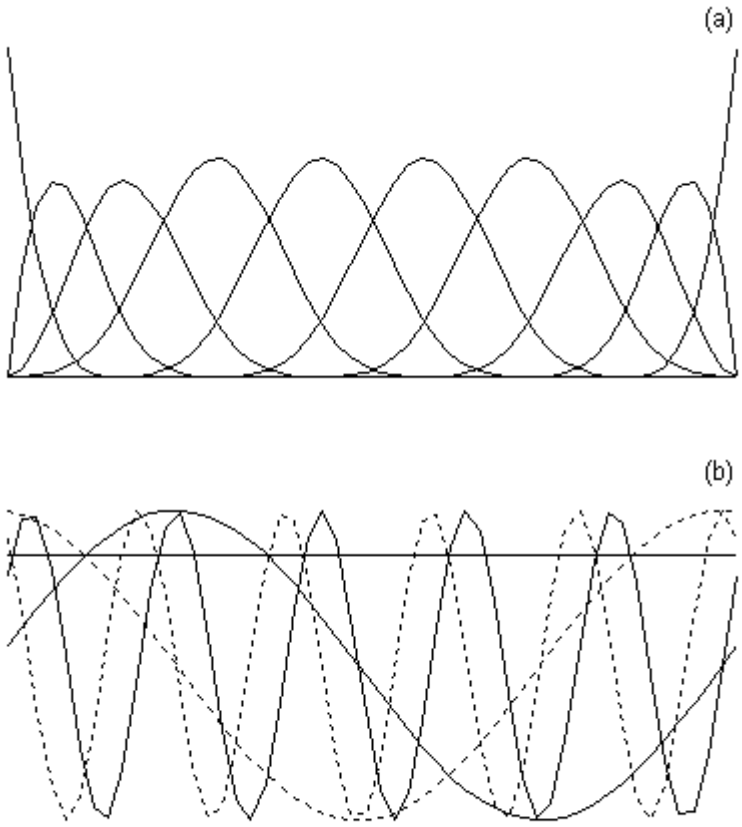


Figura 6.1- Exemplo de funções de base usadas neste trabalho. Em (a) funções B-splines, em (b) séries de Fourier.

6.3.1.1 Determinação do parâmetro de penalização ou número de bases

A suavização dos componentes é obtida neste trabalho através de dois métodos, como já mencionado. O primeiro método corresponde ao uso de um número restrito de funções de base para controlar a quantidade de regularização e o segundo, emprega um parâmetro de penalização. Desta forma, duas metodologias são usadas: A primeira emprega apenas a aproximação por quadrados mínimos penalizados para aplicar a suavização aos perfis decompostos (neste caso, o número de funções de base é igual ao número de pontos do perfil a ser suavizado e quando necessário é adicionado uma função de base na série de Fourier). Na segunda metodologia, a aproximação por quadrados mínimos penalizados (o número de funções de base é igual ao número de pontos do perfil a ser

suavizado,) é aplicada a um dos modos e para outro modo é usado o número de funções de base para controlar a suavização.

Os dados de CO, provavelmente, apresentam uma variação sistemática com um período de 24 horas e um efeito sazonal ao longo do ano devido às mudanças climáticas segundo as diferentes estações ao longo do ano. Assim, o PARAFAC foi testado com duas combinações diferentes de splines: A série de Fourier, como função de base, com a aproximação por quadrados mínimos penalizados foi usada para suavizar o perfis dos modos horas do dia (HA) e semanas do ano (SA), as funções B-spline foram usadas apenas para representar o modo SA sendo a suavização controlada pelo número de funções de base. A contribuição dos dias da semana foi considerada a mesma para os cinco dias tendo sido mantida constante. A Tabela 6.1 apresenta um sumário das metodologias e as duas combinações de splines foram chamadas de Método-A e Método-B.

Tabela 6.1 – Resumo da metodologia

Modo	**Método para suavização**		**Nome do Método**	
	Funções de base	Controle da suavização	Método-A	Método-B
Horas do Dia (HD)	Séries de Fourier	-Parâmetro de penalização (λ) (Quadrados mínimos penalizado)	X	X
Semanas do Ano (SA)	Séries de Fourier	- Parâmetro de penalização (λ) (Quadrados mínimos penalizado)	X	
	B-splines	-Número de funções de base		X
Dias da Semana (DS)	Mantido constante	_____	X	X

A suavização, naqueles casos onde o PARAFAC é restrito pelo Método-A, depende de um parâmetro: O parâmetro de penalização. Para o Método B, que emprega B-splines para o modo SA, vários parâmetros devem ser encontrados (*i.e.* número de nós, posições dos nós, número de funções de base). Neste caso, seria proibitiva a decomposição através do PARAFAC com este grande número de parâmetros a serem encontrados. Assim, o número de parâmetros requeridos pelos B-splines foi reduzido através da utilização de nós igualmente espaçados, onde o número de nós é dado por : n_{nos} =

n_{bases}–(ord_{pol}+1) + 2, sendo ord_{pol} a ordem do polinômio, pois (ord_{pol}+1) é o número de nós necessários para descrever o espaço definido pelos splines e "2" corresponde aos nós de extremo.

O parâmetro de penalização λ na expressão **203** ou o número de funções B-spline foram determinados através de um método de validação cruzada ordinária *[Silverman(a), Wahba]*. Silverman considera este tipo de método como sendo um dos mais atrativos para determinar parâmetros de splines *[Silverman(a)]*. O método de validação cruzada neste trabalho tem como princípio a eliminação de pontos do vetor $\boldsymbol{\beta}_f$ (*ver* expressão **204**), um a cada vez. O valor do parâmetro desejado é determinado quando os valores daqueles pontos eliminados forem preditos da melhor forma possível através de funções ajustadas com os pontos remanescentes do vetor $\boldsymbol{\beta}_f$. Assim, o melhor valor para o parâmetro desejado é aquele que minimiza a função de validação cruzada dada pela expressão **209**.

$$VCO(\lambda) = \sum_{i=1}^{I} \left(\beta_{f,i} - {}^{\lambda,i}\alpha_{f,i}\right)^2 \qquad 209$$

onde ${}^{\lambda,i}\alpha(t)$ é a função ajustada através da aproximação dada na expressão **203** quando o ponto *i* do vetor $\boldsymbol{\beta}_f$ é eliminado e ${}^{\lambda,i}\alpha_{f,i}$ é o valor do ponto *i* calculado pela função ${}^{\lambda,i}\alpha(t)$ para o parâmetro λ.

Para o Método-B, a variável da função *VCO* é o número de funções de base e o termo $\int \left(g_f''\right)^2 dt$ na expressão **203** desaparece, pois neste caso, não é aplicado nenhuma penalização.

O método de *VCO*, como descrito na expressão **209**, consome um grande tempo computacional, mas existem formas eficientes de efetuar a validação cruzada *[Green]*. No Apêndice 6.1 é ilustrada uma destas formas.

6.3.2 Modelo PARAFAC para os dados de CO

O tráfego de veículos automotores é considerado a principal fonte emissora de monóxido de carbono na cidade de São Paulo *[CETESB]*. Desta forma, a concentração medida de monóxido de carbono deve ser proporcional ao número de veículos automotores e ao efeito sazonal devido à mudanças climáticas (por exemplo, a quantidade de CO medida em um dia com ventos deve ser menor que aquela obtida em um dia sem ventos). Em resumo, se as condições climáticas não são favoráveis à dispersão de CO, então a maior parte do gás que poderia atingir o instrumento o faz, no entanto, se as

condições climáticas são favoráveis à dispersão, apenas parte do gás que poderia atingir o instrumento é medida. As condições climáticas não são as mesmas ao longo do ano, dependendo, principalmente, da estação*[CETESB]* e desta forma agindo de formas diferentes sobre a dispersão de CO.

Com as suposições: De que o tráfego de veículos automotores é o mesmo para os 5 dias da semana possuindo uma variação sistemática com período de 24 horas; e que não há mudanças significativas das condições climáticas durante cada semana, e aquelas que por ventura ocorram são consideradas com variação aleatória (ou variação rápida e local), um modelo multiplicativo, ou trilinear, é sugerido na expressão **210** para efetuar a análise exploratória destes dados.

$$x_{mnr} = \sum_{f=1}^{F} a_{mf} b_{nf} c_{rf} \qquad 210$$

onde x_{mnr} é a concentração de CO medida na hora *m* no dia *n* da semana *r*. a_{mf} é a contribuição do número de carros na hora *m* para o fator *f*, b_{nf} é a contribuição do dia *n* da semana para o fator *f* e c_{rf} é a contribuição do efeito sazonal para a semana *r* no fator *f*. *F* é o número de fatores usados na decomposição.

Assim, o PARAFAC foi usado para decompor o conjunto de dados em componentes trilineares cujos perfis são: **a-** o perfil sistemático para o modo HD devido à contribuição do tráfego de veículos automotores, **b**-é mantido fixo, pois assumiu-se que todos o cinco dias da semana apresentam a mesma contribuição, e **c** é o perfil para o modo SA que descreve o efeito sazonal ao longo do ano. Os três perfis são descritos com letra minúscula pois apenas um fator foi usado na decomposição (*i.e. F*=1). Adicionalmente, o arranjo em três modos foi rearranjado para a dimensão 25×5×52, onde 25 para o modo HD corresponde a 24 horas mais a primeira hora para completar um ciclo completo (*i.e.* um período de zero a 2π).

6.4 Resultados

A principal variação sistemática presente nestes dados é devido ao tráfego de veículos automotores, como já mencionado. Considerando que o principal objetivo nesta parte do trabalho é identificar tal variação no CO emitido, o PARAFAC foi usado para decompor o conjunto de dados em um componente trilinear (*i.e. F=1*).

Duas curvas de validação cruzada são mostradas na Figura 6.2 para o ano de 1999 empregando o Método-A e os valores finais para número de funções de base e do parâmetro de penalização são mostrados na Tabela 6.2 para os anos de 1997 e 1999.

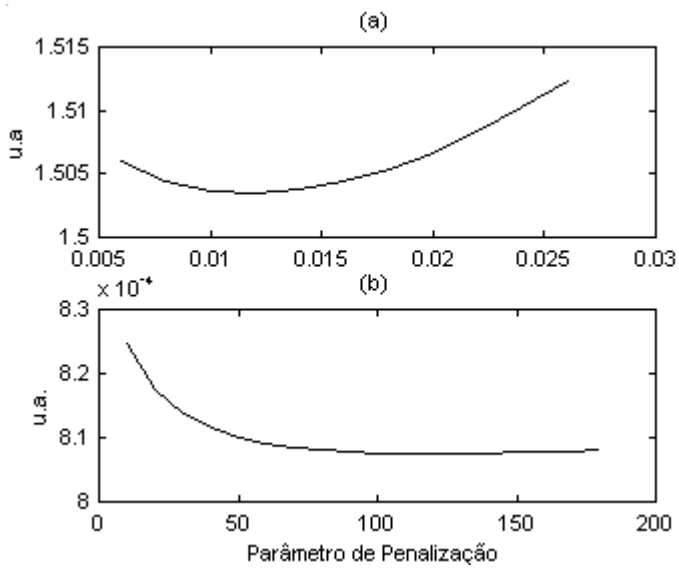


Figure 6.2 – Funções da validação cruzada ordinária (parâmetro de penalizção como variável) para os modos horas do dia (HD) em (a) e para semanas do ano (SA) em (b), para o ano de 1999. u.a.- unidade arbitrária.

Tabela 6.2 – Resultados da validação cruzada na determinação dos parâmetros dos splines

	Controle da suavização	**Ano**	
		1997	1999
Método-A	Parâmetro de penalização λ (HD)	0,149	0,012
	Parâmetro de penalização λ(SA)	180	120
Método-B	Parâmetro de penalização λ(HD)	0,151	0,012
	Número de bases (SA)	5	5

A primeira informação que se pode obter a partir da Tabela 6.2, através da comparação entre os valores dos parâmetros de penalização, é que há muito mais variação rápida e local para o modo SA comparado ao modo HD, (*i.e.* quanto maior o valor do parâmetro de penalização maior é a presença de

variação rápida e local). Isto é confirmado pelas Figuras 6.3, 6.4, 6.5 e 6.6, onde os perfis decompostos pelo PARAFC, para os anos de 1999 e 1997, com e sem restrição de suavidade são apresentados. O perfil para o modo HD, Figuras 6.3 e 6.4, apresenta alta correlação com o fluxo de veículos, principalmente para a posição de picos e vales, segundo um relatório do departamento de tráfego de São Paulo para o mês de abril destes dois anos *[http(d)]*.

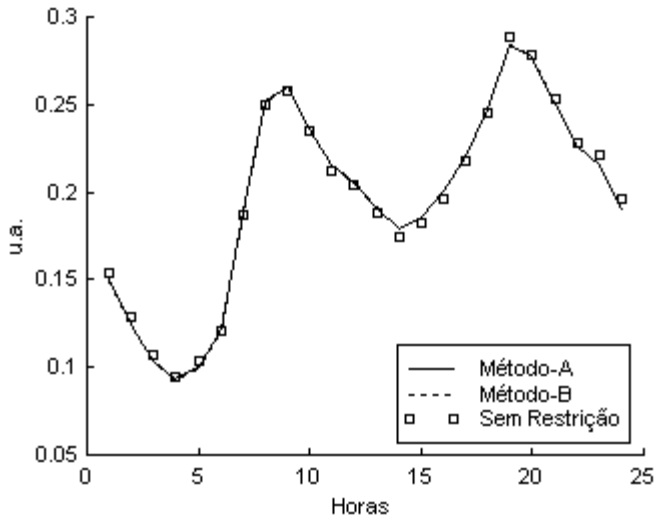


Figure 6.3- Perfis decompostos pelo PARAFAC para o modo Horas do Dia (HD) para o ano de 1999. u.a.- unidade arbitrária.

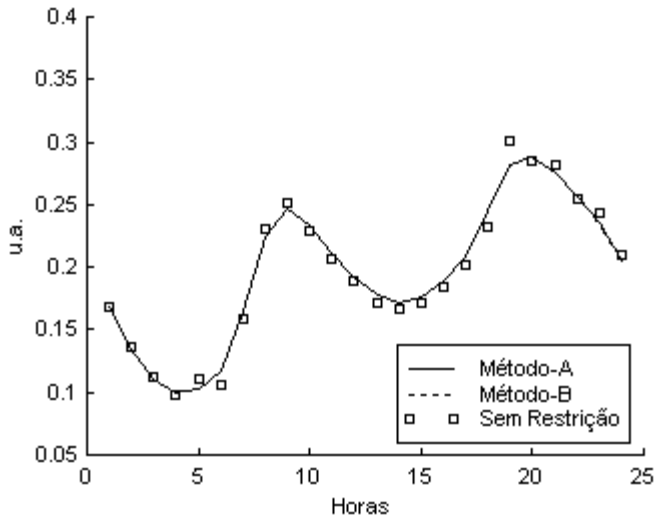


Figure 6.4- Perfis decompostos pelo PARAFAC para o modo Horas do Dia (HD) para o ano de 1997. u.a.- unidade arbitrária.

Os perfis decompostos pelo PARAFAC para modo SA são mostrados nas Figuras 6.5 e 6.6 para os anos de 1999 e 1997, tanto para solução restrita quanto para a sem restrição. A Figura 6.7 mostra os perfis para o modo SA para os anos de 1997 e 1999, para solução restrita, onde é possível verificar que o efeito sazonal para estes dois anos é diferente, principalmente, para estação da seca.

Sazonalmente, as concentrações de CO são maiores durante a estação da seca, onde é mais provável a estagnação da atmosfera. Dentre os fatores responsáveis por este tipo de estagnação, a freqüência de inversões de temperatura e ventos de baixa velocidade são fatores importantes para gerar condições pouco favoráveis à dispersão de CO *[Colucci]*. A Figura 6.8 mostra as freqüências de inversões de temperatura*[CETESB]* durante a estação seca para os dois anos envolvidos neste trabalho, onde é possível verificar que o ano de 1997 apresenta uma freqüência de inversões de temperatura maior que o ano de 1999. A inversão de temperatura limita a capacidade de ventilação vertical da

atmosfera *[Colucci]*, gerando condições de estagnação, assim, uma alta freqüência de inversões de temperatura pode ser associada à maior acumulação de CO. Desta forma, o fato de 1997 ter apresentado maiores freqüências de inversão de temperatura durante a seca sugere que este ano apresentou mais condições favoráveis à acumulação de CO se comparado ao ano de 1999. A Tabela 6.3 apresenta o número de dias em cada mês com condições não favoráveis à dispersão de CO durante a seca (*i.e.* semanas ~19-39, meses ~5 a 8). A Figura 6.8 e a Tabela 6.3 trazem informações a respeito das condições climáticas destes dois anos, sugerindo que no ano de 1997 houve mais condições favoráveis à acumulação de CO se comparado ao ando de 1999 durante a seca.

Os perfis para o modo SA obtidos pelo PARAFAC com a restrição de suavidade, mostrados na Figura 6.7, sugerem que foi mais "fácil" medir a concentração de CO durante a seca, por exemplo, em 1997 se comparado ao ano de 1999. Considerando que é mais fácil medir o CO em condições de estagnação, ou seja, com condições não favoráveis à dispersão do mesmo, a comparação das informações climáticas (Figura 6.8 e Tabela 6.3) mostram que o PARAFAC com a restrição de suavidade é capaz de diferenciar e identificar os efeitos sazonais destes dois anos.

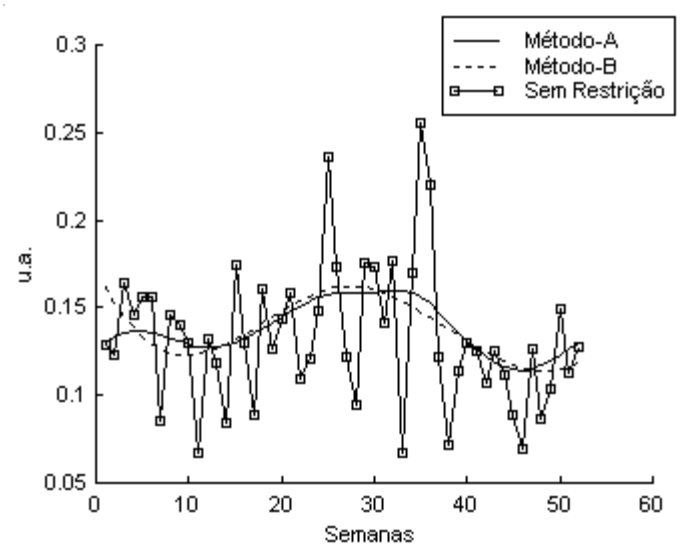


Figure 6.5- Perfis decompostos pelo PARAFAC para o modo Semanas do Ano (SA) para o ano de 1999. u.a.- unidade arbitrária.

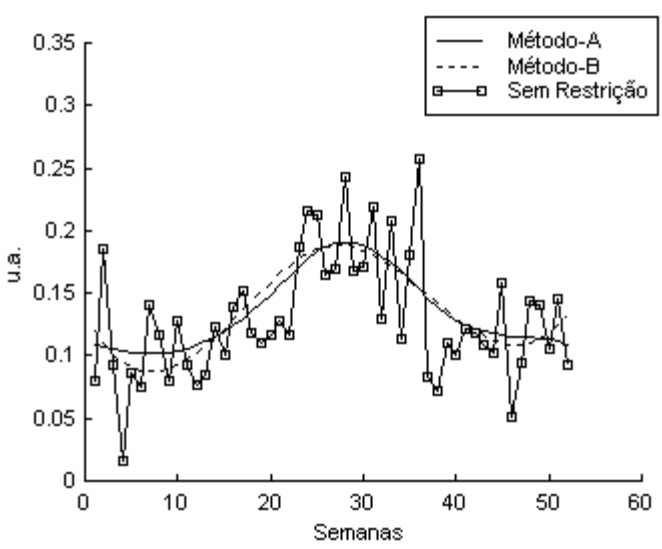


Figure 6.6- Perfis decompostos pelo PARAFAC para o modo Semanas do Ano (SA) para o ano de 1997. u.a.- unidade arbitrária.

Tabela 6.3- Número de dias com condições não propícias para a dispersão do monóxido de carbono.

Ano	**Número de dias**						
	Mês						
	Janeiro	Maio	Junho	Julho	Agosto	Setembro	Novembro
	Semanas						
	1-4	19-21	23-25	27-30	32-34	36-39	45-47
1997	0	4	3	10	14	5	1
1999	1	0	1	4	10	8	0

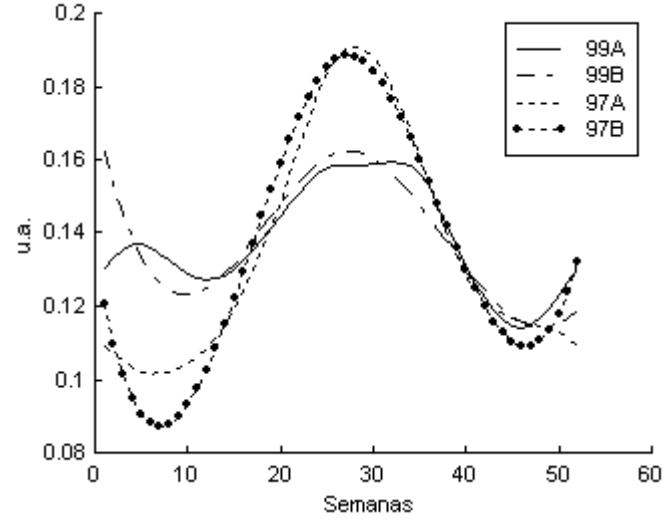


Figure 6.7- Perfis decompostos pelo PARAFAC para o modo Semanas do Ano (SA) para os anos de 1997 e 1999. u.a.- unidade arbitrária.

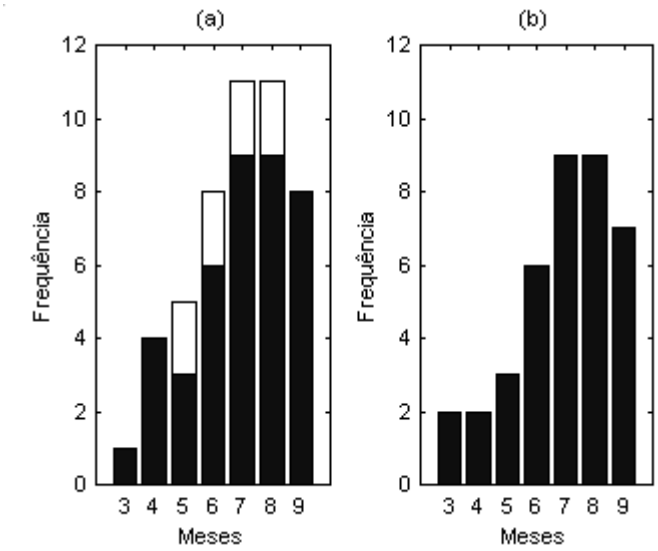


Figure 6.8 – Freqüências de inversões de temperatura, até 200m da superfície da terra, para 1997 em (a) e 1999 em (b). A parte branca das barras em indica o quanto maior foi a freqüência em 1997 se comparada com aquela para o mesmo período em 1999. u.a.- unidade arbitrária.

6.5 Conclusões

A decomposição do conjunto de dados de CO através do PARAFAC, com restrição de suavidade, resultou em perfis que permitiram diferenciar e identificar efeitos sazonais de dois anos diferentes, o que não foi possível através da solução sem o emprego desta restrição. Com isto, a metodologia estudada, modelo trilinear com restrição de suavidade, mostra ser uma boa ferramenta de análise exploratória para casos onde é possível usar um modelo multiplicativo com a necessidade de suavidade.

A importância desta metodologia é considerar os dados com funções amostradas com erro e provendo, ao final, uma descrição funcional da variação de CO.

6.6 Apêndice 6.1

Considere o problema de Quadrados Mínimos Penalizados como mostrado na expressão **211**, onde a função $l_{f\lambda}$ deve ser minimizada.

$$l_{f\lambda} = \sum_t \left(\beta_{ft} - \alpha_{ft}\right)^2 + \lambda \int \left(g_f^{"}\right)^2 dt \qquad 211$$

Reescrevendo a expressão **211** em termos dos vetores dos dados e da função discretizada no pontos correspondentes aos do vetor de dados como dado nas expressões **212** e **213**, tem-se a expressão **215**.

$$\mathbf{y} = \begin{pmatrix} \beta_{f1} & \beta_{f2} & \cdots & \beta_{fn} \end{pmatrix}^T \qquad 212$$

$$\mathbf{g} = \begin{pmatrix} \alpha_{f1} & \alpha_{f2} & \cdots & \alpha_{fn} \end{pmatrix}^T \qquad 213$$

onde

$$\alpha_{f,j} = g_f\left(t_j\right) \qquad 214$$

$$l_{f\lambda} = \left\|\mathbf{y} - \mathbf{g}\right\|^2 + \lambda \int \left(g_f^{"}\right)^2 dt \qquad 215$$

Antes de ilustrar a solução do problema descrito na expressão **215** (onde $l_{f\lambda}$ deve ser minimizada), considere as notações descritas nas expressões **216** e **217**, usadas para indicar o operador diferencial para a derivada segunda.

$$g_f^{"} = \frac{d^2}{dt^2}\left(g_f(t)\right) \qquad 216$$

$$\frac{d^2}{dt^2} = D^2 \qquad 217$$

A função g_f corresponde a uma expansão em funções de base como mostrado na expressão **218**

$$g_f = \boldsymbol{\psi}\boldsymbol{\gamma} \qquad 218$$

onde

$$\boldsymbol{\psi} = \begin{pmatrix} \psi_1(t) & \psi_2(t) & \cdots & \psi_k(t) \end{pmatrix} \qquad 219$$

sendo $\psi_j(t)$ as funções de base e $\boldsymbol{\gamma}$ o vetor de coeficientes.

O termo da derivada segunda integrada da expressão **215** pode ser reescrito em termos da expressão 229, como mostra a formulação descrita nas expressões de **220** a 229.

$$\int_T \left(g''_f\right)^2 dt = \int_T \left(D^2[\boldsymbol{\psi\gamma}]\right)^2 dt \qquad 220$$

$$\boldsymbol{\psi\gamma} = \gamma_1\psi_1(t) + \gamma_2\psi_2(t) + \cdots + \gamma_k\psi_k(t) \qquad 221$$

onde γ_1,..., γ_j ,..., γ_k são os elementos do vetor de coeficientes $\boldsymbol{\gamma}$.

$$D^2[\boldsymbol{\psi\gamma}] = D^2[\gamma_1\psi_1(t)] + D^2[\gamma_2\psi_2(t)] + \cdots + D^2[\gamma_k\psi_k(t)] \qquad 222$$

$$D^2[\boldsymbol{\psi\gamma}] = \gamma_1 D^2\psi_1(t) + \gamma_2 D^2\psi_2(t) + \cdots + \gamma_k D^2\psi_k(t) \qquad 223$$

$$\left\{D^2[\boldsymbol{\psi\gamma}]\right\}^2 = \left[\gamma_1 D^2\psi_1(t) + \gamma_2 D^2\psi_2(t) + \cdots + \gamma_k D^2\psi_k(t)\right] \cdot \left[\gamma_1 D^2\psi_1(t) + \gamma_2 D^2\psi_2(t) + \cdots + \gamma_k D^2\psi_k(t)\right] \qquad 224$$

$$\left\{D^2[\boldsymbol{\psi\gamma}]\right\}^2 = \sum_j \sum_i \left\{\gamma_i\gamma_j \left[D^2\psi_i(t)\right]\left[D^2\psi_j(t)\right]\right\} \qquad 225$$

$$\int_T \left(D^2[\boldsymbol{\psi\gamma}]\right)^2 dt = \int_T \sum_j \sum_i \left\{\gamma_i\gamma_j \left[D^2\psi_i(t)\right]\left[D^2\psi_j(t)\right]\right\} dt \qquad 226$$

$$\int_T \left(D^2[\boldsymbol{\psi\gamma}]\right)^2 dt = \sum_j \sum_i \gamma_i\gamma_j \int_T \left[D^2\psi_i(t)\right]\left[D^2\psi_j(t)\right] dt \qquad 227$$

$$\int_{\mathrm{T}} \left[D^2\psi_i(t)\right]\left[D^2\psi_j(t)\right]dt = \Omega_{ij} \tag{228}$$

$$\int_{\mathrm{T}} \left(D^2[\boldsymbol{\psi}\boldsymbol{\gamma}]\right)^2 dt = \boldsymbol{\gamma}^T\boldsymbol{\Omega}\boldsymbol{\gamma} \tag{229}$$

Com isto, a expressão **215** pode ser escrita em termos da expressão **230** ou da expressão **231**.

$$l_{f\lambda} = \|\mathbf{y}-\mathbf{g}\|^2 + \lambda\boldsymbol{\gamma}^T\boldsymbol{\Omega}\boldsymbol{\gamma} \tag{230}$$

$$l_{f\lambda} = (\mathbf{y}-\mathbf{g})^T(\mathbf{y}-\mathbf{g}) + \lambda\boldsymbol{\gamma}^T\boldsymbol{\Omega}\boldsymbol{\gamma} \tag{231}$$

Para ilustração da solução da expressão **231** (onde $l_{f\lambda}$ deve ser minimizada) considere, primeiro, a forma discretizada das funções de base nos pontos correspondentes aos dados, neste caso, a função de base discretizada é indicada com o chapéu "~", como mostrado na expressão 232.

$$l_{f\lambda} = (\mathbf{y}-\tilde{\boldsymbol{\psi}}\boldsymbol{\gamma})^T(\mathbf{y}-\tilde{\boldsymbol{\psi}}\boldsymbol{\gamma}) + \lambda\boldsymbol{\gamma}^T\boldsymbol{\Omega}\boldsymbol{\gamma} \tag{232}$$

O problema descrito na expressão **232** (onde $l_{f\lambda}$ deve ser minimizada) tem como variável o vetor de coeficientes **γ**. Assim, através da diferenciação descrita nas expressões de **233** a 236 e fazendo a derivada dada na expressão 236 igual a zero, obtém-se a solução, ou seja, **γ** como mostrado na expressão 242 (*ao leitor é sugerido uma consulta a obras de cálculo diferencial, uma boa sugestão é [Magnus], para este trabalho em particular, a descrição do problema de quadrados mínimos na página 232 deste livro é de grande interesse, ver também o tópico interpretação diferencial em quadrados mínimos no apêndice da seção de fundamentos*).

$$dl_{f\lambda} = 2(\mathbf{y}-\tilde{\boldsymbol{\psi}}\boldsymbol{\gamma})^T d(\mathbf{y}-\tilde{\boldsymbol{\psi}}\boldsymbol{\gamma}) + 2\lambda\boldsymbol{\gamma}^T\boldsymbol{\Omega}d\boldsymbol{\gamma} \tag{233}$$

$$dl_{f\lambda} = 2(\mathbf{y}-\tilde{\boldsymbol{\psi}}\boldsymbol{\gamma})^T(-\tilde{\boldsymbol{\psi}}d\boldsymbol{\gamma}) + 2\lambda\boldsymbol{\gamma}^T\boldsymbol{\Omega}d\boldsymbol{\gamma} \tag{234}$$

$$dl_{f\lambda} = -2\mathbf{y}^T\tilde{\boldsymbol{\psi}}d\boldsymbol{\gamma} + 2\boldsymbol{\gamma}^T\tilde{\boldsymbol{\psi}}^T\tilde{\boldsymbol{\psi}}d\boldsymbol{\gamma} + 2\lambda\boldsymbol{\gamma}^T\boldsymbol{\Omega}d\boldsymbol{\gamma} \tag{235}$$

$$\frac{dl_{f\lambda}}{d\boldsymbol{\gamma}} = -2\mathbf{y}^T\tilde{\boldsymbol{\psi}} + 2\boldsymbol{\gamma}^T\tilde{\boldsymbol{\psi}}^T\tilde{\boldsymbol{\psi}} + 2\lambda\boldsymbol{\gamma}^T\boldsymbol{\Omega} \tag{236}$$

$$\frac{dl_{f\lambda}}{d\boldsymbol{\gamma}} = 0 \tag{237}$$

$$-2\mathbf{y}^T\tilde{\boldsymbol{\psi}} + 2\boldsymbol{\gamma}^T\tilde{\boldsymbol{\psi}}^T\tilde{\boldsymbol{\psi}} + 2\lambda\boldsymbol{\gamma}^T\boldsymbol{\Omega} = 0 \tag{238}$$

$$\boldsymbol{\gamma}^T\left(\tilde{\boldsymbol{\psi}}^T\tilde{\boldsymbol{\psi}} + \lambda\boldsymbol{\Omega}\right) = \mathbf{y}^T\tilde{\boldsymbol{\psi}} \tag{239}$$

$$\left(\tilde{\boldsymbol{\psi}}^T\tilde{\boldsymbol{\psi}} + \lambda\boldsymbol{\Omega}\right)^T\boldsymbol{\gamma} = \tilde{\boldsymbol{\psi}}^T\mathbf{y} \tag{240}$$

$$\left(\tilde{\boldsymbol{\psi}}^T\tilde{\boldsymbol{\psi}} + \lambda\boldsymbol{\Omega}\right)\boldsymbol{\gamma} = \tilde{\boldsymbol{\psi}}^T\mathbf{y} \tag{241}$$

$$\boldsymbol{\gamma} = \left(\tilde{\boldsymbol{\psi}}^T\tilde{\boldsymbol{\psi}} + \lambda\boldsymbol{\Omega}\right)^{-1}\tilde{\boldsymbol{\psi}}^T\mathbf{y} \tag{242}$$

O algoritmo empregado neste trabalho, proveniente do pacote computacional descrito por Ramsay e Silverman *[Ramsay]*, não emprega a forma discretizada das funções de base, mas sim a interpolação do vetor dos dados. Desta forma, o vetor dos dados passar a ser descrito segundo a expressão **243**.

$$\mathbf{y} = \begin{pmatrix} y(1) & y(2) & \cdots & y(n) \end{pmatrix}^T \tag{243}$$

onde

$$y(t) = \boldsymbol{\psi}\boldsymbol{\gamma}_y \tag{244}$$

Ou seja, a expressão **242** pode ser reescrita em termos da expressão **245**.

$$\boldsymbol{\gamma} = \left(\Theta + \lambda\boldsymbol{\Omega}\right)^{-1}\Theta\boldsymbol{\gamma}_y \tag{245}$$

onde

$$\int_{T} \psi_i(t)\psi_j(t)dt = \Theta_{ij} \qquad 246$$

6.6.1 Validação Cruzada Ordinária

A determinação do parâmetro de penalização, λ, através da validação cruzada ordinária pode ser obtido segundo a seguinte formulação (uma excelente abordagem deste tema é encontrado na seguinte literatura: *[Ramsay,Hastie, Green]*). Primeiro, considere a multiplicação de ambos os lados da expressão **242** pela matriz das funções de base discretizadas, como mostrado na expressão **247**.

$$\tilde{\psi}\boldsymbol{\gamma} = \tilde{\psi}\left(\tilde{\psi}^T\tilde{\psi} + \lambda\boldsymbol{\Omega}\right)^{-1}\tilde{\psi}^T\mathbf{y} \qquad 247$$

A expressão **247** mostra que o vetor de dados suavizado, dado pela expressão **248** , pode ser obtido através do mapeamento do vetor de dados por uma matriz que depende apenas do parâmetro de penalização λ. Assim, a expressão **247** pode ser escrita em termos da expressão **248**.

$$\mathbf{g} = \mathbf{W}_\lambda\mathbf{y} \qquad 248$$

onde

$$\mathbf{W}_\lambda = \tilde{\psi}\left(\tilde{\psi}^T\tilde{\psi} + \lambda\boldsymbol{\Omega}\right)^{-1}\tilde{\psi}^T \qquad 249$$

Desta forma, a função de validação cruzada pode ser dada pela expressão **250** *[Ramsay, Hastie, Green]*.

$$VCO(\lambda) = n^{-1}\sum_{1}^{n}\left(\frac{y_i - g_i}{1 - W_{\lambda,ii}}\right)^2 \qquad 250$$

onde $W_{\lambda,ii}$ é o elemento *i,i* da matriz $\mathbf{W}_\lambda$ e *n* é o número de pontos do vetor de dados.

A expressão **250** mostra uma grande redução no tempo computacional gasto para o cálculo de λ, pois para cada λ apenas uma matriz $\mathbf{W}_\lambda$ é calculada, ao contrário daquela forma descrita no texto principal, onde *n* matrizes similares a $\mathbf{W}_\lambda$ são determinadas para cada λ.

7 Conclusões gerais

A conclusão geral deste trabalho de tese é: *"o emprego de métodos de ordem superior deve considerar os desvios das suposições, feitas por estes métodos acerca dos dados, tanto para elaboração dos métodos quanto para análise dos dados"*. Esta conclusão pode parecer óbvia para muitos, mas para este autor não o é. Um exemplo disto é o desempenho do método NBRA usado para calibração de segunda ordem. Este método apresentou grande eficiência para os casos onde os desvios das suposições, feitas por ele acerca dos dados, eram pequenos, o que poderia lhe conferir destaque frente a outros métodos de mesmo propósito, por possuir o NBRA grande desempenho computacional. No entanto, para os casos onde os desvios das suposições, os quais não podem ser eliminados do experimento, foram significativos, o NBRA errou, em muito, na predição da concentração de amostras desconhecidas. A análise exploratória, dos dados daquelas amostras onde o NBRA falhou, mostrou que as variações experimentais (deslocamentos de perfis e colinearidades entre eles) eram suficientes para alterar a estrutura dos dados e causar a falha do NBRA. O interessante neste problema é que combinações binárias de três isômeros foram estudadas e apenas as combinações entre dois isômeros específicos se mostraram problemáticas, fato relacionado à maior similaridade entre estes dois isômeros. Por outro lado, se os efeitos da colinearidade (relacionada aos desvios experimentais devido à similaridade entre os compostos), por exemplo, são considerados na elaboração do modelo de calibração, ótimos resultados são obtidos *[kiers(f)]* para o mesmo caso onde NBRA falha. Ainda sobre a análise dos resultados do NBRA, uma operação que pode parecer óbvia, para muitos, isto é, determinar os perfis de tempo a partir dos espectros conhecidos das espécies ácida e básica da mistura, na verdade não o é. Isto pois, se as características do experimento, neste caso, as propriedades semelhantes de difusão para os isômeros, não forem consideradas, os perfis de tempo encontrados podem não possuir significado físico.

Neste mesmo sentido, a separação de curvas, onde o maior problema encontrado foi a determinação do número de curvas a serem resolvidas, mostra que tanto a solução para 4 curvas quanto para 5 curvas (resultado não apresentado) seriam interessantes. No entanto, a etapa de validação mostrou que para a resolução espectral empregada, a melhor solução obtida é aquela para 4 curvas. Enfim, aquela parte dos dados que não pode ser acomodada no modelo trilinear usado na separação de curvas afeta a solução da otimização envolvida no método, mas a validação permitiu identificar a solução com significado físico. Em resumo, se as características experimentais (ruídos, interferentes, compostos com grande similaridade, que se tornam "iguais" dependendo da resolução espectral) que

podem gerar mínimos locais no processo de otimização não são consideradas, a resolução de curvas pode gerar um resultado que não é, por completo, verdadeiro. Por outro lado, se estas características são consideradas e uma validação é sugerida, pode-se obter um parâmetro que auxilia na escolha do melhor resultado.

A análise das propriedades físico-químicas de amidos extraídos de féculas de mandioca foi efetuada para identificar a influência da idade (estágio fisiológico) e efeitos sazonais nas propriedades dos amidos. Se considerado que o amido é constituído de uma população de grânulos, os quais, em geral, determinam tais propriedades, deve se esperar que, em média, os grânulos sejam afetados tanto pelo efeito sazonal quanto pelo estágio fisiológico. No entanto, as respostas frente a estes efeitos não devem ocorrer de forma semelhante para toda a população, mas todos os grânulos apresentam respostas aos experimentos. Adicionalmente, há as características genéticas das quatro variedades estudadas. Assim, as propriedades físico-químicas apresentam variações segundo a idade, efeito sazonal e aquelas adversas. Baseado nisto, a análise exploratória foi efetuada para separar estes dois tipos de variação, aplicando para isto, métodos que acomodassem pelo menos aquela variação de interesse. Esta análise resultou em perfis relacionados a efeitos sazonais que caracterizam as variações daquelas propriedades segundo estes efeitos.

No estudo da concentração horária de monóxido de carbono ao longo de um ano, a análise foi efetuada para verificar como aquele perfil diário, associado à determinada fonte emissora, seria afetado pelos efeitos sazonais, ou seja, se aquele perfil diário seria mais ou menos influenciado pelas diferentes condições climáticas ao longo do ano. Neste caso, era de interesse apenas aquela variação gradativa, segundo as estações do ano. Assim, aquelas variações "repentinas" ocorridas no período de uma semana deveriam ser separadas daquela gradativa ao longo do ano. Desta forma, a análise exploratória dos dados foi efetuada sob a restrição de suavidade tendo sido obtidos resultados que descrevem os efeitos sazonais ao longo do ano como confirmado por fatores ambientais.

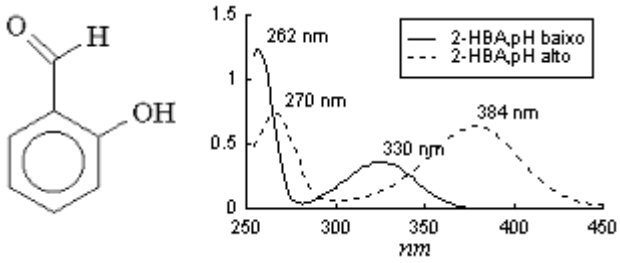
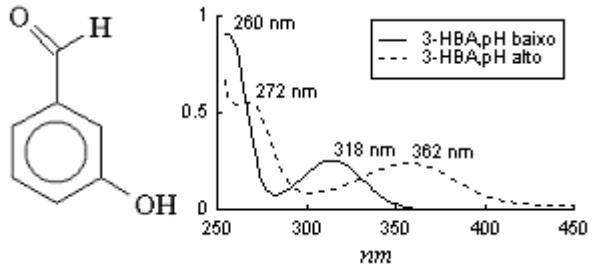
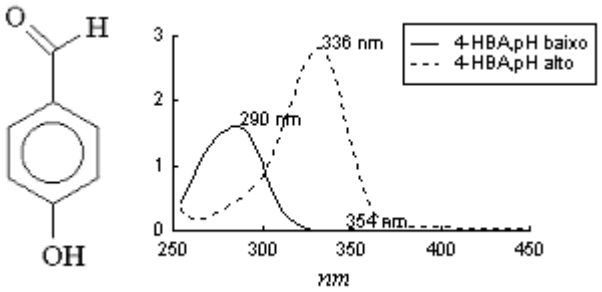
Concluindo, um dos quatro tópicos estudados, a calibração de ordem superior, evidencia a influência das variações não acomodadas pelos métodos de ordem superior nos resultados destes métodos. Nos outros três tópicos, em geral, são sugeridas formas para amenizar a influência destas variações e até mesmo sua identificação.

8 Notação

$\left\| \cdot \right\|^2$	Soma dos quadrados dos elementos de "·"
$(\cdot)^T$, ex.: $\begin{pmatrix} \alpha & \delta & \varphi \\ \beta & \varepsilon & \gamma \\ \chi & \phi & \eta \end{pmatrix}^T = \begin{pmatrix} \alpha & \beta & \chi \\ \delta & \varepsilon & \phi \\ \varphi & \gamma & \eta \end{pmatrix}$, em termos formais para uma matriz $\mathbf{A}=a_{ij}$ a operação de tranposição é dada por $\mathbf{A}^T=a_{ji}$	Operação de transposição de Matrizes ou vetores.
$\begin{pmatrix} \alpha & \delta & \varphi \\ \beta & \varepsilon & \gamma \\ \chi & \phi & \eta \end{pmatrix} \times \begin{pmatrix} a & d \\ b & e \\ c & f \end{pmatrix} = \begin{pmatrix} a\alpha + b\delta + c\varphi & d\alpha + e\delta + f\varphi \\ a\beta + b\varepsilon + c\gamma & d\beta + e\varepsilon + f\gamma \\ a\chi + b\phi + c\eta & d\chi + f\phi + fn \end{pmatrix}$, em termos formais, o produto entre duas matrizes $\mathbf{A}=a_{ij}$ e $\mathbf{B}=b_{ij}$ é definido como $\mathbf{AB} = \mathbf{C}$, sendo $c_{ik} = \sum_j a_{ij} b_{jk}$	Produto matricial
$\left(\cdot_1 \middle| \cdot_2 \middle| \cdots \middle| \cdot_w \right)$	Matriz de w colunas se "." for um vetor. Caso "." seja uma matriz o resultado é uma matriz de (w·z) colunas, onde z é o número de colunas desta matriz. "|" é usado para separar colunas,
$\mathbf{X}_1, \mathbf{G}_1$	"Fatias" (matrizes) dos arranjos em 3 modos (N×M) e (P×Q)
$\underline{\mathbf{X}} = \left(\mathbf{X}_1 \middle\| \mathbf{X}_2 \middle\| \cdots \middle\| \mathbf{X}_O \right), \underline{\mathbf{G}} = \left(\mathbf{G}_1 \middle\| \mathbf{G}_2 \middle\| \cdots \middle\| \mathbf{G}_R \right)$	Arranjos em 3 modos (N×M×O) e (P×Q×R)
$\mathbf{x}_1 = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{N1} \end{pmatrix}^T, \mathbf{a}_1 = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{N1} \end{pmatrix}^T$, $\mathbf{b}_1 = \begin{pmatrix} b_{11} & b_{21} & \cdots & b_{M1} \end{pmatrix}^T, \mathbf{c}_1 = \begin{pmatrix} c_{11} & c_{21} & \cdots & c_{O1} \end{pmatrix}^T$	Vetores coluna

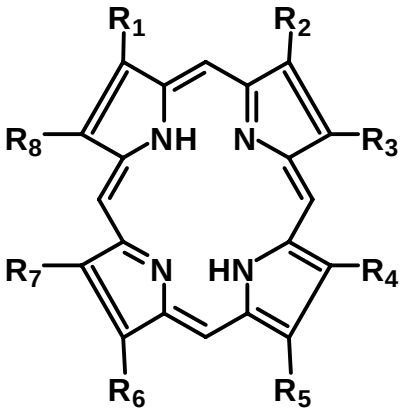
$\mathbf{A}=\begin{pmatrix}\mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_P\end{pmatrix}, \mathbf{B}=\begin{pmatrix}\mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_Q\end{pmatrix},$ $\mathbf{C}=\begin{pmatrix}\mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_R\end{pmatrix}$	Matrizes componentes
${}^1\mathbf{c}=\begin{pmatrix}c_{11} & c_{12} & \cdots & c_{1R}\end{pmatrix}$	Vetor linha
$vec\mathbf{X}_1=\begin{pmatrix}\mathbf{x}_1\\ \mathbf{x}_2\\ \vdots\\ \mathbf{x}_M\end{pmatrix}$, onde $\mathbf{X}_1=\begin{pmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_M\end{pmatrix}$	Forma vetorizada da fatia 1, sendo $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_M$, vetores coluna
$\left(vec\mathbf{X}_1 \middle\vert \; vec\mathbf{X}_2 \middle\vert \; \cdots \middle\vert \; vec\mathbf{X}_O\right)=vec\underline{\mathbf{X}},$ $\left(vec\mathbf{G}_1 \middle\vert \; vec\mathbf{G}_2 \middle\vert \; \cdots \middle\vert \; vec\mathbf{G}_R\right)=vec\underline{\mathbf{G}}$	Aplicação do operador de "vetorização" aos arranjos em 3 modos.
$(\mathbf{C}\otimes\mathbf{B})=\begin{pmatrix}c_{11}\mathbf{B} & c_{12}\mathbf{B} & \cdots & c_{1O}\mathbf{B}\\ \vdots & \vdots & \ddots & \vdots\\ c_{R1}\mathbf{B} & c_{R2}\mathbf{B} & \cdots & c_{RO}\mathbf{B}\end{pmatrix}$	Produto tensorial de Kronecker
$\mathbf{A}\mathbf{G}_1\left(\mathbf{c}_1^T\otimes\mathbf{B}^T\right),\ (\mathbf{A}\otimes\mathbf{B})vec\mathbf{G}_{C-1}\mathbf{c}_1^T$	"Blocos" (P×Q×R), (N×M×O) e (N×M×O) respectivamente

9 Glossário

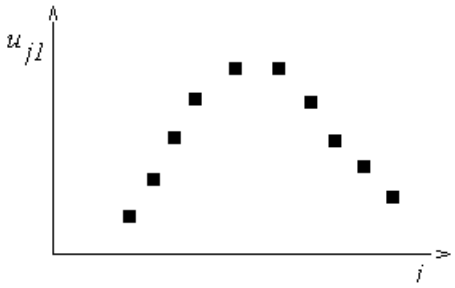
Termo	Significado
i.e.	*do Latin* "id est" significa "em outras palavras".
u.a.	Unidade arbitrária
2-HBA	2-hidroxibenzaldeído 
3-HBA	3-hidroxibenzaldeído 
4-HBA	4-hidroxibenzaldeído 
GC-MS	*do Inglês* "Gas Chromatography-Mass Spectrometry"- indica instrumento de cromatografia a gás com um espectrômetro de massas acoplado.
LC-UV	*do Inglês* "Liquid Chromatography-UltraViolet Spectroscopy"- indica instrumento de cromatografia líquida com um espectrofotômetro de absorção na região do ultravioleta.

MS-MS	*do Inglês* "Mass-Mass Spectrometry"- indica espectrômetro de massa-massa.
FIA	*do Inglês s "Flow Injection Analysis"- análise por injeção em fluxo*
GRAM	*do Inglês "Generalized Rank Annihilation Method", [Sanchez]*
NBRA	*do Inglês "Non-Bilinear Rank Annihilation", [Wilson, Wang(a)]*
RAFA	*do Inglês "Rank Annihilation Factor Analysis", [Ho]*
RBL	*do Inglês "Residual Bilinearization", [Öman]*
PARAFAC	*do Inglês "PARAllel FACtor analysis", [Bro (b), Harshman]*
PCA	*do Inglês* "Principal Component Analysis " Análise de Componentes Principais – método usado para efetuar a decomposição da matriz de dados, **X**, em "escores",**T** , e "loadings", **P**, ou seja, uma projeção da matriz de dados em uma base ortonormal: **X**=**TP**^{T}, onde **P**^{T}**P**=**I** e **T**^{T}**T**=**Λ**, sendo **I** a matriz identidade e **Λ** uma matriz diagonal.
SVD	*do Inglês* "Singular Value Decomposition" Decomposição em Valores Singulares – Método usado para decompor a matriz de dados, **X**, em duas matriz ortogonais, **U** e **V**, e uma matriz diagonal não-negativa, com os valores singulares, **S**: **X=USV**^{T}, onde **V**^{T}**V**=**VV**^{T}=**I** e **U**^{T}**U**=**UU**^{T}=**I**, sendo **I** a matriz identidade e **S** uma matriz diagonal.
Aditividade linear	Neste trabalho é usado para indicar que o posto (*ver* posto) de uma matriz resultante da soma de duas outras é igual à soma do posto destas matrizes: posto(**X**)=posto(**A**)+posto(**B**), onde **X**=**A**+**B**.
Autovalores-autovetores	Indica a solução de um problema de autovalores-autovetores, definido como: $\mathbf{W}\psi=\psi\mathbf{\Pi}$, sendo **W** uma matriz quadrada, **Π**

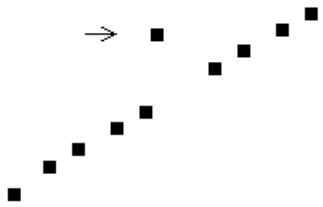
	uma matriz diagonal de autovalores e $\boldsymbol{\Psi}$ a matriz de autovetores.
Base	Uma "base" corresponde a um conjunto de vetores linearmente independentes usados para descrever um espaço vetorial.
Base ortonormal	É uma "base" (ver base) onde os vetores usados para descrever um vetor ou matriz são ortogonais ($\mathbf{u}_i^T\mathbf{u}_j=0$, $i\neq j$) e com norma igual a um ($\mathbf{u}_i^T\mathbf{u}_j=1$, $i=j$) .
Bilinear (forma bilinear)	Por definição, uma **forma bilinear** em V é uma transformação f:V×V→K, onde V é um espaço vetorial de dimensção finita e K um dado "corpo", $\mathbb{Q}, \mathbb{R}, \mathbb{C}$, que satisfaz: (i) $f(au_1+bu_2\ , v) = a f(u_1\ ,v) + b f(u_2,\ , v)$ *(ii)* $f(u\ , av_1+bv_2) = a f(u\ ,v_1) + b f(u\ , v_2)$ para quaisquer $a, b \in$ K e quaisquer u_i e $v_j \in$ V. A condição (i) diz que f é linear na primeira variável e a condição (ii) diz que f é linear na segunda variável. Considere o exemplo da seção Fundamentos para mistura de dois compostos em concentração unitária, para a qual é medida a absorbância no comprimento de onda j e tempo i. Neste caso, a absorbância é da por $f_{i,j}(t_i\ ,\ s_j) = (t_{i1}\ s_{j1}) + (t_{i2}\ s_{j2})$, onde $t_i=(t_{i1}\ \ t_{i2})$, $s_j = (\ s_{j1}\ s_{j2})$, sendo s_{j1} e s_{j2} as absortividades molares no comprimento de onda j para os dois compostos, respectivamente; t_{i1} e t_{i2} as constantes proporcionais ao tempo i para os dois compostos, respectivamente.
Decomposição em Valores Singulares	*Ver SVD*
Escores	*Ver PCA*
Funções de base	Por simplicidade são definidas aqui, com sendo funções usadas para descrever um espaço funcional, ou seja, formar uma base para este espaço.

Hematoporfirina	 R1-$CHOHCH_3$, R2-CH_3, R3- $CHOHCH_3$, R4-CH_3, R5-CH_2CH_2COOH, R6-CH_2CH_2COOH, R7-CH_3 , R8-CH_3 *[Falk, pag. 5]*
Hifenados	Instrumentos acoplados, como por exemplo “GC-MS”, “LC-UV”.
Justaposição (Justapor)	Indica a ação de formar uma nova matriz através da união de duas ou mais matrizes, por exemplo, **X**= (**A** | **B** | **C**), onde “ | ” indica que as matrizes foram unidas lateralmente, isto é, as colunas de **B** foram colocadas ao lado das de **A**, o mesmo válido para **C**.
“Loadings”	*Ver PCA*
Matrizes componentes	Matrizes resultantes da decomposição por meio dos métodos de ordem superior, por exemplo **A**, **B**, e **C** em: $\underline{\mathbf{X}} = \mathbf{A}\underline{\mathbf{I}_{DS}}\left(\mathbf{C}^T \otimes \mathbf{B}^T\right)$ (PARAFAC) $\underline{\mathbf{X}} = \mathbf{A}\underline{\mathbf{G}}\left(\mathbf{C}^T \otimes \mathbf{B}^T\right)$ (TUCKER)
Matriz diagonal	Matriz quadrada com elementos $x_{ii} \neq 0$ e $x_{ij} = 0$ para $i \neq j$ $\begin{pmatrix} x_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & x_{nn} \end{pmatrix}$

Matriz identidade	Matriz quadrada com elementos $x_{ii}=1$ e $x_{ij}=0$ para $i \neq j$ $\begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}$
Matriz ortogonal	É uma matriz quadrada, digamos **U**, onde a seguinte propriedade é válida: $\mathbf{U}^T\mathbf{U}=\mathbf{U}\mathbf{U}^T=\mathbf{I}$, sendo **I** a matriz identidade.
Matriz semi-ortogonal	É uma matriz, digamos **T**, onde uma das seguintes igualdades é válida: $\mathbf{T}^T\mathbf{T} = \mathbf{I}$ ou $\mathbf{T}\mathbf{T}^T=\mathbf{I}$, sendo **I** a matriz identidade.
Métodos em Multi (Três) Modos	São métodos usados para tratar dados em multi modos.
Não-negatividade	*Ver restrição de não-negatividade*
Parâmetro de rugosidade ou penalização	É o parâmetro que controla a quantidade de suavização a ser aplicada em um dado vetor de dados. Por exemplo o λ na seguinte Equação : $l_{f\lambda} = \sum_t \left(\beta_{f\,t} - \alpha_{f\,t}\right)^2 + \lambda \int \left(g_f^{''}\right)^2 dt$ (*ver definição da Equação no texto principal seção: PARAFAC com splines*)
Posto	O posto de uma matriz é o maior número de colunas ou linhas desta matriz linearmente independentes. O posto de uma matriz é um número inteiro positivo e pode ser identificado através de uma Decomposição em Valores Singulares (*ver* SVD), ou seja, o posto corresponde ao número de valores singulares diferentes de zero.
Posto completo	Uma matriz é dita posto completo quando seu posto é igual ao número de colunas ou linhas, ou seja, o menor entre os dois.
Posto-um	Matrizes que possuem **posto-um** neste contexto são aquelas que a aditividade linear é válida, **X** é dita **posto-um** se for resultado da soma de duas ou mais matrizes e o valor de seu posto for igual a soma do posto de cada matriz. Por exemplo: *posto*(**X**)=*posto*(**A**)+*posto*(**B**), onde **X**=**A**+**B**.

Posto-maior-que-um	Matrizes que possuem **posto-maior-que-um** neste contexto são aquelas para as quais a aditividade linear **não** é válida, **X** é dita **posto-maior-que-um** se for resultado da soma de duas ou mais matrizes e o valor de seu posto for diferente da soma do posto de cada matriz. Por exemplo: $posto(\mathbf{X}) \neq posto(\mathbf{A}) + posto(\mathbf{B})$, onde $\mathbf{X}=\mathbf{A}+\mathbf{B}$.
Pseudo-posto	A matriz de dados experimentais, **A**, pode apresentar-se como sendo a soma de dois tipos de matrizes, aquela com as informações úteis, $\hat{\mathbf{A}}$, e a de ruídos, **E**, ou seja, $\mathbf{A}=\hat{\mathbf{A}}+\mathbf{E}$. Se $\hat{\mathbf{A}}$ não for posto completo, então é dito que o **pseudo-posto** de $\hat{\mathbf{A}}$ é igual ao posto de $\hat{\mathbf{A}}$.
Restrição de não-negatividade	Significa que os elementos de um vetor ou matriz devem ser maiores ou iguais a zeros.
Restrição de ortogonalidade	Significa que os vetores de uma matriz, digamos **U**, devem ser ortogonais ($\mathbf{u}_i^T\mathbf{u}_j=0$, $i \neq j$).
Restrição de unimodalidade	Significa que a variação dos elementos de um vetor deve formar uma única banda na ordem em que aparecem no vetor. Por exemplo para um vetor $\mathbf{u}$ $(= \mathbf{u}_{11}\ \mathbf{u}_{21} \ldots \mathbf{u}_{10,1})^T$: 
Rotação livre	É a rotação da matriz resultante da projeção de uma dada matriz de dados, digamos **X**. Por exemplo, o PCA ($\mathbf{X}=\mathbf{TP}^T$) é dito como sendo de **rotação livre**, pois com um emprego de uma matriz **W**, pode-se "girar" a matriz **T**: $\mathbf{X}=\mathbf{TWW}^T\mathbf{P}^T$, pois $\mathbf{WW}^T=\mathbf{W}^T\mathbf{W}=\mathbf{I}$, onde **I** é a matriz identidade. Assim, **X** também pode ser representado por $\mathbf{X}=\mathbf{ZQ}^T$, onde $\mathbf{Z}=\mathbf{TW}$ e $\mathbf{Q}=\mathbf{PW}$.

Sobreposição de posto	Ocorre quando a aditividade linear não é válida: $posto(\mathbf{X}) \neq posto(\mathbf{A}) + posto(\mathbf{B})$, onde $\mathbf{X} = \mathbf{A} + \mathbf{B}$.
Sobrescrito	É usado para indicar a posição de um símbolo em dada equação ou expressão, quando na parte superior da linha, por exemplo o "*T*" em: $\mathbf{P}^T$.
Suavização	Significa eliminar a variação rápida e local.
Subscrito	É usado para indicar a posição de um símbolo em dada equação ou expressão, quando na parte inferior da linha, por exemplo o "*j*" em: $\mathbf{u}_j$.
Trilinear (forma trilinear)	A **forma trilinear** é uma extensão natural da **forma bilinear** em V, ou seja, uma transformação $f: V \times V \times V \rightarrow K$,
Truncada	Uma das formas de se efetuar o PCA é através de uma SVD, ou seja, $\mathbf{X} = \mathbf{USV}^T \Rightarrow \mathbf{X} = \mathbf{TP}^T$ onde $\mathbf{T} = \mathbf{US}$ e $\mathbf{P} = \mathbf{V}$. Uma matriz, por exemplo **T**, é dita truncada quando apenas parte de **U**, e conseqüentemente de **S**, é usada para descrever **T**, ou seja, apenas um número reduzido de colunas de **U** e **V**.
Unimodalidade	*Ver Restrição de unimodalidade*
Valores singulares	*Ver SVD*
Variação rápida	É uma variação "repentina" ocorrida ao longo de uma variação gradativa, alterando o curso desta. Por exemplo: 

Variação rápida e local	É uma variação “repentina” ocorrida ao longo de uma variação gradativa sem alterar o curso desta. Por exemplo: 

10 Notas Computacionais

Os cálculos deste trabalho de tese foram efetuados em um Pentium-Intel 300 MHz, em ambiente Windows®. O programa empregado foi o MATLAB tendo sido usados os “*toolboxes*” disponíveis na internet nos sites *[http (b) e http(c)]*. Também foram elaborados códigos próprios (*para* NBRA, RBL, PARFAC para incluir os splines).

Um programa computacional de domínio público alternativo é a linguagem-R disponível em www.r-language.org, onde há diversos pacotes, inclusive para dados de ordem superior (PTAk), splines e várias ferramentas empregadas em quimiometria.

11 Referências Bibliográficas

Alsberg, B.K.; Kvalheim, O. M.; *"Compression of nth-Order Data Arrays by B-Splines. Part 1:Teory." Journal of Chemometrics,* **1993, *7,*** 61-73.

Andre, J.C.; Bouchy, M., Viriot, M.L.; "*Synchronous Excitation Methods for Increasing Sensitivity in Fluorimetry , the Limitations Caused by Raman and Rayleigh Sscatter, Analytica Chimica Acta,* **1979,** *105,* 297-310.

Atkins, P. K.; *"Physical Chemistry"*, Oxford University Press: 5ª ed. 1994.

Barcellos, S. B.; Reis, M. M.; Ferreira, M. M. C.; *"Quimiometria II dados de ordem superior: Métodos, Modelos, Aplicações". Em preparação.*

Besse, P.; Ramsay, J. O.; *"Principal Components Analysis of Sampled Functions*", *Psychometrika*, **1986**; *51*, (2), 285-311.

Booksh, K. S; Kowalski, B. R.; *"Theory of Analytical Chemistry", Analytical Chemistry* **1994,** *66*, (15), A782-A791.

(a) Bro, R.; Heimdal, H.; "*Enzymatic browning of vegetables. Calibration and analysis of variance by multi-way methods*", *Chemometrics and Intelligent Laboratory Systems*, **1996,** *34*, 85-102

(b) Bro, R.; "*PARAFAC: tutorial & applications*", *Chemometrics and Intelligent Laboratory Systems*, **1997,** *38,* 149-171

(c) Bro, R.; Sidiropoulos, N. D.; *"Least Squares Algorithms Under Unimodality and Non-Negativity Constraints", Journal of Chemometrics* **1998,** 12, 223-247.

(d) Bro, R.; *"Multi-Way Analysis in the Food Industry, Models Algorithms and Applications"*, Ph.D Thesis (1998), University of Amsterdam.

Buguera, J. L.; *"Flow Injection Atomic Spectroscopy",* Marcel Dekker, inc:**1989**, pag.19

CETESB-*Relatório de qualidade do ar de São Paulo para* **1999**.

Colucci, J. M.; Begeman, C. R.; "*Carbon monoxide in Detroid, New York and Los Angeles air*", *Enviromental Science and Technology* **1969**; *4*, (1), 3-39.

de Boor C.; *"Pratical Guide to Splines"*, Springer-Verlag: New York,**1987**.

Efron, B.; Tibshirani, R. J.; *"An introduction to the bootstrap",* CRC Press: London, **1994**.

Faber, N. M.; Buydens, L. M. C.; Kateman, G.; *"Aspects Of Pseudorank Estimation Methods Based On The Eigenvalues Of Principal Component Analysis Of Random Matrices", Chemometrics and Intelligent Laboratory Systems* **1994,** *25*, (2), 203-226.

Falk, J E.; "*Porphyrins and Metalloporphyrins"*, vol. 2, Elsevier, Amsterdam, **1964**.

(a) Ferreira, M. M. C.; *et al.;* "*Chemometric study of the fluorescence of dental calculus by trilinear decomposition*", *Applied Spectroscopy*, **1995,** *49*, (9) 1317-1325.

(b) Ferreira, M. M. C.; Ferreira, W. C. Jr; Kowalski, B. R.; *"Rank Determination and Analysis of non-linear Processes by Global Linearizing Transformation", Journal of Chemometrics,* **1996,** *10,*: 11-30.

Green, P. J.; Silverman, B. W.; *"Nonparametric regression and Generalized Linear Models"*, Chapman Hall: London UK, **1994**.

Harshman, R. A.; Lundy, M. E.; "*PARAFAC: parallel factor analysis*", *Computational Statistics & Data Analysis*", **1994,** *18*, 39-72.

Hastie, T. J.; Tibshirani, R. J.; "*Generalized Additive Models",* Chapman and Hall: London, **1997**, 24-26.

Henderson, H. V.; "*The Vec-Permutation Matrix, The Vec Operator and Kronecker Products: A Review", Linear Algebra and Multilinear Algebra,* **1981**, *9*, 271-281

Hirschfeld, T.; "*The Hy-phen-ated Methods", Analytical Chemistry* **1980,** *52*, (2), A297-A312.

Ho, C.-N.; Christian, G. D.; Davidson, E. R.; *"Application of Method of Rank Annihilation to Quantitative-Analyses of Multicomponent Fluorescence Data From Video Fluorometer*", *Analytical Chemistry* **1978,** *50*, 1108-13.

(a) http://www.models.kvl.dk/users/rasmus/, [versão 1.03, 1998].

(b) http://www.models.kvl.dk/source/ [*site* atual, Agosto 2001].

(c) http://www.psych.mcgill.ca/faculty/ramsay/software.html [verificado em 30 de Julho 2001].

(d) http://200.19.93.5/internew/informativo/balanco/relad.html [verificado em 31 de Julho 2000].

(e) http://web.mit.edu/18.06/www/ [verificado em 8 de março 2002].

(a) Kiers, H. A. L.; "*Hierarchical Relations Among Three-Way Methods ", Psychometrika*, **1991**, *56*, (3), 449-470.

(b) Kiers, H. A. L.; Kroonenberg, P. M.; ten Berge, J. M. F.; *"An Efficient Algorithm for TUCKALS3 on Data With Large Numbers of Observation Units", Psychometrika*, **1992**, *57*, (3), 415-422.

(c) Kiers, H. A. L.; Smilde, A. K. "*Some Theoretical Results On 2nd-Order Calibration Methods For Data With And Without Rank Overlap*", *Journal of Chemometrics* **1995,** *9*, (3), 179-95.

(d) Kiers, H. A. L.; ten Berge, J. M. F.; Rocci, R.; "*Uniquess of Three-Mode Factor Models with Sparse Cores: The 3×3×3 Case*", *Psychometrika*, **1997**, *62*, (3), 349-374.

(e) Kiers, H. A. L.; "*Three-Mode Orthomax Rotation*", *Psychometrika*, **1997**, *62*, (4), 579-598.

(f) Kiers, H. A. L.; Smilde, A. K.; "*Constrained Three-Mode Factor Analysis As A Tool For Parameter Estimation With Second-Order Instrumental Data*", *Journal of Chemometrics,* **1998,** *12*, 125-47.

(g) Kiers, H. A. L.; "*Towards a Standardized Notation and Terminology in Multiway Analysis*", *Journal of Chemometrics,* **2000,** *14*, 105-122.

Leach, H. W.; "*Gelatinization of starch. em: Starch chemistry and technology Vol. I, R.L. Whistler, E.F. Paschall*" , New York: Academic Press, **1965** p.289-307.

Magnus, J. R.; Neudecker, H.; "*Matrix differential calculus with applications in statistics and econometrics"*, John Wiley & Sons, **1988**.

Nørgaard, L.; Ridder, C.; "*Rank Annihilation Factor-Analysis Applied To Flow-Injection Analysis With Photodiode-Array Detection*", *Chemometrics and Intelligent Laboratory Systems* **1994,** *23*, 107-114.

Öhman, J.; Geladi, P.; Wold, S.; "*Residual Bilinearization.Part 1:Theory and Algorithms*" *Journal of Chemometrics* **1990,** *4*, 79-90.

Paatero, P.; Juntto, S.; "*Determination of underlying components of cyclical time series by means of two-way and three-way factor analytic techniques"*, *Journal of Chemometrics* **2000**; *14*, 241-259.

Perrin, D. D.; Demsey, B.; "*Buffers for pH and metal ion control",* Chapman & Hall: London, **1974**.

Ramsay, J. O.; Silverman, B. W.; "*Functional Data Analysis",* Springer-Verlag: New York, **1997**.

(a) Reis, M. M.; Ferreira,M. M. C.; "Separação de espectros simulados e de luminescência total através do método generalizado de anulação do posto (GRAM)", Química Nova, **1999,** 22, 11-17.

(b) Reis, M. M.; Gurden S.P; Smilde A.P.; Ferreira, M. M. C.; "*Calibration And Detailed Analysis Of Second-Order Flow Injection Analysis Data With Rank Overlap*", *Analytica Chimica Acta* **2000,** *422*, (1), 21-36.

(c) Reis, M. M.; Biloti, D. N.; Ferreira, M. M. C.; Pessine F. B. T.; "*Curve Resolution Of Total Luminescence Of Human Dental Tartar By Parafac", Applied Spectroscopy***, 2001**, *55*,(7) 847-851.

(d) Reis, M. M.; Ferreira,M. M. C.; Sarmento, S. B. S.; "*A Methodological Multi-Way Analysis Of Starch Properties*" **Artigo submetido.**

Rickard, J. E.; Asaoka, M.; Blanshard, J. M. V.; *"The Physico-Chemical properties of cassava starch"*, *Tropical Science*, **1991,** *31*, 189-207.

Ross ,R. T.; Leurgans, S.; "*Component resolution using multilinear models, Methods in Enzymology*", **1995,** *246*, 679-700.

Sanchez, E.; Kowalski, B. R.; "*Generalized Rank Annihilation Factor-Analysis*", *Analytical Chemistry*, **1986,** *58*, (2), 496-99.

(a) Sarmento, S. B. S; "*Caracterização da Fécula de Mandioca (Manihot esculenta C.) no Período de Colheita de Cultivares de uso Industrial*" *Tese de Doutorado*, Universidade de São Paulo - Faculdade de Ciências Farmacêuticas – Departamento de Alimentos e Nutrição – **1997**.

(b) Sarmento, S. B. S; Reis, M. M.; Ferreira, M. M. C.; Cereda, M. P.; Penteado, M.V.C.; Anjos, C. B.; *"Análise Quimiométrica de Propriedades Físicas, Físico-Químicas e Funcionais de Féculas de Mandioca"*, *Braz. J. Food Technology*, **1999**, *2*, (1-2), 131-137.

Saurina, J.; Hernández-Cassou, S.; Tauler, R.; Izquierdo-Ridorsa, A.; *"Multivariate resolution of rank-deficient spectrophotometric data from first-order kinetic decomposition reactions"*, *Journal of Chemometrics,* **1998,** *12*, (3) 183-03.

Serjeant, E. P.; Demsey, B.; *"Ionisation constants of organic acids in aqueous solution (IUPAC chemical data series; 23)"*, Pergamon Press: Oxford, **1979**.

(a) Silverman, B. W.; *"Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting"*, *Jornal of Royal Statistcal Society*, **1985**, *47*, (1), 1-52.

(b) Silverman, B. W.; *"Smoothed Functional Principal Components Analysis by Choice of Norm"*, *The Annals of Statistics*, **1996**, *24*, (1), 1-24.

(a) Smilde, A. K.; Wang, Y.; Kowalski, B. R.; "*Theory Of Medium-Rank 2nd-Order Calibration With Restricted-Tucker Models*", *Journal of Chemometrics* **1994,** *8*, (1), 21-36.

(b) Smilde, A. K.; Tauler, R.; Henshaw, J. M.; Burgess, I. W.; Kowalski, B. R.; "*Multicomponent Determination Of Chlorinated Hydrocarbons Using A Reaction-Based Chemical Sensor .3. Medium-Rank 2nd-Order Calibration With Restricted Tucker Models"*, *Analytical Chemistry* **1994,** *66*, (20), 3345-51.

(c) Smilde, A. K.; Tauler, R.; Saurina, J.; Bro, R.; "*Calibration Methods for Complex Second-Order Data*", *Analytica Chimica Acta* **1999,** *398*, 237-51.

Strang, G.; "*Linear Algebra and Its Application*", Harcourt Brace Jovanovich, Publisher**, 1976.**

Wahba, G.; Wold, S.; *"A completely automatic french curve: Fitting spline fucntions by cross validation"*, *Communication in Statistics*, **1975**, *4*, **(**1), 1-17.

(a) Wang, Y.; Borgen, O. S.; Kowalski, B. R.; Gu, M.; Turecek, F.; "*Advances in Second-Order Calibration*", *Journal of Chemometrics* **1993,** *7*, 117-30.

(b) Wang, Y.; Borgen, O. S.; Kowalski, B. R.; "*Comments on the Residual Bilinearization Method*", *Journal of Chemometrics* **1993,** *7*, 439-45.

Wilson, B. E.; Lindberg, W.; Kowalski, B. R.; "*Multicomponent Quantitative Analysis Using Second-Order Nonbilinear Data: Theory and Simulations*",*Journal of the American Chemical Society* **1989,** *111*, 3797-804.

Wegman, E. J.;, Wright, I. W.; "*Spline in Statistics*", *Journal of the American Statistical Association*, **1983,** *78*, (382), 351-365.

Wold, S.; "*Spline Functions in Data Analysis"*, *Technometrics*, **1974**, *16*, (1), 1-11.

Tucker, L. R.; "*Some Mathematical Notes on Three-Mode Factor Analysis*", *Psychometrika*, **1966**, *31*, (3), 279-311.

Vandeginste B.G.M.; Sielhorst C.; Gerritsen M.; "*The Nipals Algorithm For The Calculation Of The Principal Components of a Matrix"*, *TRAC-Trends In Analytical Chemistry*, **1988**, *7*, (8), 286-287.

Índice Remissivo