Universidade Federal de São Carlos
Centro de Ciências Exatas e Tecnológicas - CCET

Igor Ferreira Torquato

CORAL: Machine Learning e matrizes de rotação aplicados a resolução de sistemas multivariados

São Carlos - SP

4 de maio de 2022

Universidade Federal de São Carlos
Centro de Ciências Exatas e Tecnológicas - CCET

Igor Ferreira Torquato

CORAL: Machine Learning e matrizes de rotação aplicados a resolução de sistemas multivariados

Trabalho final de curso apresentado
como requisito para a obtenção do título de
Bacharel no curso de Engenharia Física pela
Universidade Federal de São Carlos.

Orientador: Prof. Dr. Fabio Aparecido Ferri
Co-Orientador: Dr. Santiago José Alejandro Figueroa

São Carlos - SP

4 de maio de 2022

Ferreira Torquato, Igor

CORAL: Machine Learning e matrizes de rotação aplicados
a resolução de sistemas multivariados / Igor Ferreira Torquato – 2022.
60f.

TFC (Graduação) - Universidade Federal de São Carlos, campus São Carlos, São Carlos

Orientador (a): Fábio Aparecido Ferri
Co-Orientador (a): Santiago José Alejandro Figueroa

Banca Examinadora: Vinicius Tribuzi Rodrigues Pinheiro Gomes, Javier Fernando Ramos Caro
Bibliografia

1. Física, Química, Computação. I. Ferreira Torquato, Igor. II. Título.

Ficha catalográfica desenvolvida pela Secretaria Geral de Informática (SIn)
DADOS FORNECIDOS PELO AUTOR
Bibliotecário responsável: Ronildo Santos Prado - CRB/8 7325

Resumo

O CORAL (Curve ResOlution for dAta anaLysis) é uma biblioteca de ferramentas quimiométricas baseada em Python para decomposição espectral multivariada de grandes conjuntos de dados, especialmente aqueles em que as técnicas envolvidas obedecem a lei de Beer ou uma forma de combinação linear de dados. Os pesquisadores podem usá-la no ambiente Jupyter para enfrentar os desafios relacionados ao grande número de espectros gerados pela rápida aquisição de dados durante estudos de reações catalíticas com resolução temporal. Além de permitir o estabelecimento de um caderno experimental único de controle das linhas de luz do acelerador de partículas de quarta geração, o Sirius, e análise de dados em Jupyter.

Apesar de versáteis, as técnicas multivariadas, como MCR-ALS (do inglês, *Multivariate Curve Resolution with Alternating Least Squares*) e PCA (do inglês, *Principal Component Analysis*), apresentam problemas intrínsecos à decomposição espectral, pois, em cada decomposição há múltiplas respostas que satisfazem o sistema de equações matriciais a variabilidade de respostas é conhecida como Ambiguidade Rotacional (ARs). Nesse aspecto, o estudo das ARs pode revelar informações importantes do sistema estudado, como o erro associado a cada composto puro encontrado pelos métodos multivariados.

O presente projeto pretende apresentar os desenvolvimentos do CORAL, complementar os estudos de ARs e avaliar como as restrições (constraints) afetam as respostas da decomposição espectral, aplicando o estudo de áreas de soluções plausíveis e matrizes de transformação para observar a evolução dos espaços de solução para um conjunto de dados com duas componentes. Dessa forma, o objetivo deste projeto é agregar uma nova ferramenta ao CORAL para que os usuários deste possam desfrutar de uma análise mais completa acerca dos conjuntos de dados, de modo a garantir a escolha das melhores soluções. Além disso, análises preliminares sugerem que o método pode ser utilizado como uma estimativa do erro associado à ambiguidade rotacional e os espectros e concentrações podem ser uma estimativa inicial do MCR-ALS.

Palavras-chave: **CORAL, Combinação Linear, MCR-ALS, PCA, Jupyter, Ambiguidade rotacional**

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de siglas e abreviaturas

CORAL Curve Resolution for Data Analysis
CPU Central Processing Unit (Processador)
CuPy CUDA Python Numerical Processing Library
cuBLAS CUDA Basic Linear Algebra Subprograms
cuDNN Biblioteca CUDA para Deep Neural Network
cuRAND Biblioteca CUDA random number generation
cuSPARSE Biblioteca CUDA para matriz esparsa
cuSOLVER Biblioteca para aceleração com GPU para decomposição de sistemas lineares
EFA Evolving Factor Analysis (Análise de Fatores Evolucionários)
GPU Graphics processing unit (Placa de vídeo)
GUI Graphical User Interface (Interface gráfica de usuário)
ITTFA Interactive Target Transformation Factor Analysis
MCR-ALS Multivariate Curve Resolution with Alternating Least Squares
NCCL NVIDIA Collective Communication Library
Numpy Numerical Python
PCA Principal Component Analysis
QEXAFS Quick Scanning Extended X-ray Absorption Fine Structure
QUATI Linha de luz: Quick X-Ray Absorption Spectroscopy for Time and Space Resolved Experiments
RAM Random-access memory (Memória de acesso randômico)
SIMPLISMA SIMPLe-to-use Interactive Self-modeling Mixture Analysis
TR-XAS Time Resolved X-Ray Absorption Spectroscopy
UX User Experience (Experiência de usuário)
XAS X-Ray Absorption Spectroscopy
XAFS X-Ray Absorption Fine Structure
XRF X-Ray Fluorescense

1 Introdução

Os últimos 50 anos foram de grande importância para o desenvolvimento de técnicas quimiométricas, em especial àquelas destinadas a resolução de misturas. A quimiometria computacional pode ser entendida como a união da matemática e da estatística para interpretar, predizer e lidar com modelos químicos. Com os avanços computacionais e o maior poder de processamento, a área de Análise de Fatores (ou *Factor Analysis*, do inglês) se tornou uma ferramenta fundamental na análise de misturas. Tais misturas são exemplos comuns em sistemas multicomponentes presentes na química analítica, comumente estas são decorrentes de processos químicos ou reações. Nesse sentido, a crescente modernização de técnicas experimentais, além da instrumentação mais complexa possibilitou o surgimento de novas áreas, como a quimiometria computacional. [7]

A espectroscopia por absorção de raios-x com resolução temporal (TR-XAS), por exemplo, é um método de caracterização poderoso para estudos de catálise heterogênea e sistemas complexos que se aproveita da análise multivariada. Sendo uma técnica sensível aos elementos presentes na amostra, esta permite obter informações estruturais, químicas e eletrônicas de um dado elemento, possibilitando o estudo de sistemas complexos em condições *in situ* e/ou *operando*. Quando condições externas, tais como temperatura, atmosfera(gases) ou pressão, entre outros parâmetros, são alteradas é possível estudar a evolução dos processos fazendo uso das técnicas quimiométricas computacionais, permitindo acompanhar entre outras coisas a evolução temporal das cinéticas de reação.[8, 9]

De maneira geral, amostras analisadas com XAS são representadas por um ou mais espectros, os quais são resultantes de uma média ponderada de cada espectro do elemento absorvedor individualmente presente na amostra. Diferentes entornos Físico-Químicos dão lugar a distintas contribuições, sendo a superposição entre as contribuições de cada especie química individual uma mera combinação linear. Em situações em que amostras com diversas espécies são analisadas, há um problema de mistura com resolução temporal. Com isso, faz-se necessário o uso de técnicas computacionais capazes de separar as componentes puras em termos das concentrações relativas, isto é, resolver um sistema linear com múltiplas componentes. Para tanto, diferentes métodos matemáticos são aplicados à matriz de dados habitualmente tais soluções são obtidas por métodos de resolução multivariada ou através de rotações matriciais. [9, 10]

1.1 Motivações

Dentre as motivações para a implementação de métodos de avaliação de ambiguidade está a incerteza da comunidade de absorção de raios-X quanto ao uso de técnicas multivariadas. Pois, apesar de métodos como combinação linear serem amplamente aplicados em diversos estudos, ainda há uma certa desconfiança com relação aos métodos multivariados como PCA e MCR-ALS.

Além disso, a linha de luz QUATI (Quick X-Ray Absorption Spectroscopy for Time and Space Resolved Experiments) será capaz de gerar uma quantidade expressiva de dados provenientes de medidas de alta qualidade. Isso ocorrerá devido a resolução temporal e espacial na escala de milissegundos. Portanto, a análise dos dados desta linha tem como base a combinação linear de espectros, nesse sentido, pressupõe um conhecimento a priori das especies que se combinam na reação de estudo, e com isso uma medida previa destas especies puras para poder realizar a tal combinação linear. Contudo, muitas vezes, não há como saber para cada processo em estudo todas as reações possíveis, ou melhor, não há como medir os intermediários pelas dificuldades de síntese destes. Com isso, se torna quase que humanamente impossível aplicar em processos químicos sobre estudo a combinação linear em forma direta. Dessa forma, faz-se necessário a utilização de técnicas que permitam obter informações sobre os processos de outras fontes, sendo assim, a análise multivariada vem no auxílio em conjunto com as técnicas computacionais otimizadas. [11]

Ainda assim, a implementação em Python permite a fácil modificação do código para utilização nas mais diversas técnicas experimentais, em especial àquelas que envolvem imagens hiperespectrais, como XRF e Raman.

2 Objetivos

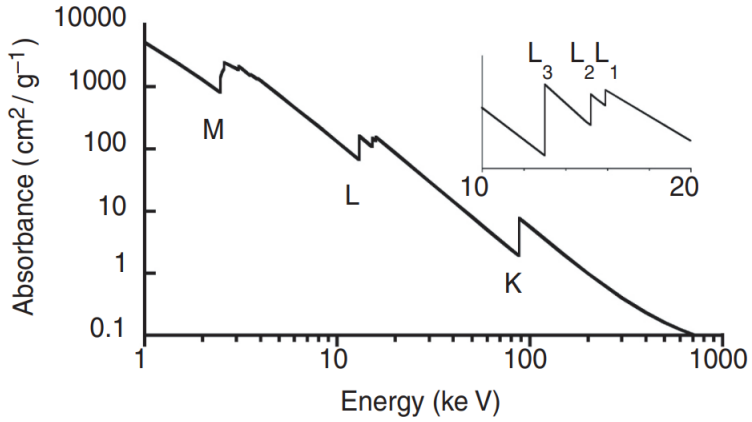
O objetivo do presente trabalho é apresentar os desenvolvimentos na área de técnicas de quimiometria, tendo principal foco nas implementações em Python do programa CORAL, cujo desenvolvimento e sofisticação foram parte deste trabalho. e a possibilidade de avaliar quanto a ambiguidade rotacional dos modelos multivariados, além de apresentar as novas funcionalidades implementadas e sua importância para a análise de dados provenientes de absorção de raios-x.

3 Introdução teórica

3.1 A Absorção de raios-X

Por definição, raios-x são considerados radiação ionizante capaz de excitar e/ou ejetar elétrons do núcleo atômico. Em termos energéticos, estão em uma faixa de 500 eV até 500 keV, isto é, comprimentos de onda na faixa de $25\mathring{A}$ a $0,25\mathring{A}$, respectivamente. Para cada camada energética de um átomo absorvedor desta radiação, existe um respectivo nível de energia de ligação, cuja radiação incidente é absorvida num processo único conhecido como efeito fotoelétrico. Ao escanear uma amostra variando as energias do raios-X é possível observar saltos abruptos em sua absorção, tal como na figura 3.1.1, os saltos são conhecidos como bordas de absorção.[1, 12, 13]

Figura 3.1.1: Espectro de absorção de múltiplas bordas de chumbo em função da energia de excitação. [1]



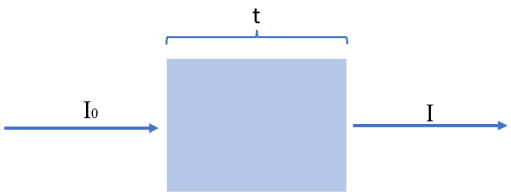
Cada uma das bordas representa um nível energético diferente, sendo as que ocorrem à maior energia (borda K, tal como na figura 3.1.1) são correspondentes a elétrons mais fortemente ligados (próximos ao núcleo atômico). Como os elétrons absorvem a radiação em energias próximas às de ligação, fótons com energias menores não perturbarão o estado quântico do elétron. Contudo, ao excitar elétrons do átomo com energias superiores, é possível que o elétron seja promovido de seu nível quântico ao contínuo, e a energia em excesso transforma-se em energia cinética do foto-elétron ejetado do átomo. [1, 13, 12]

De maneira geral, em experimentos de espectroscopia, o interesse é obter o valor da absorbância de um material em função da energia. Em amostras homogêneas a variação energética da absorbância é dependente apenas do coeficiente linear de absorção (μ). Tal coeficiente está diretamente relacionado com a probabilidade de absorção de raios-x pela amostra. A relação entre as grandezas envolvidas pode ser escrita em termos da Lei de Beer:

$$I = I_0 e^{\mu t} \tag{3.1.1}$$

onde I_0 é a intensidade da radiação incidente, I é a intensidade da radiação transmitida e t a espessura da amostra, assim como representado na figura 3.1.2.

Figura 3.1.2: Representação de um experimento de absorção.

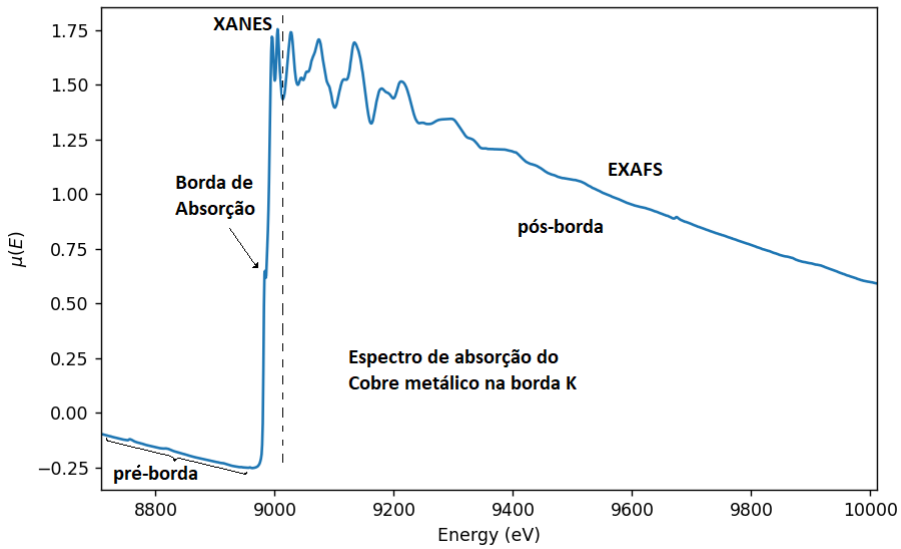


Em estudos de absorção, o intuito está em estudar a relação de μ com energias próximas às borda de absorção do elemento de interesse. Assim, pode-se entender um espectro de XAFS como uma medida de μ em energias próximas e maiores que as de ligação dos níveis energéticos. Como cada elemento possui elétrons com energias de ligação bem definidas, é possível medir diferentes bordas ajustando-se a varredura em energia do feixe incidente, sendo assim a técnica é sensível aos elementos químicos do material. [1]

Após a absorção, o átomo está em um estado excitado, pois um dos níveis eletrônicos foi deixado vazio, pois um fotoelétron foi emitido do átomo. Passado poucos femtossegundos após a interação, o estado excitado decai de duas maneiras diferentes: o primeiro mecanismo de decaimento é a fluorescência, em que elétrons mais externos preenchem a vacância do que foi ejetado, levando a emissão de um feixe de raios-x com energia bem definida é emitido; o segundo mecanismo consiste no efeito Auger, onde um elétron da camada externa preenche o nível atômico vazio e outro elétron é ejetado do átomo. Ambos processos podem ser utilizados para medir a absorção de raios-x de um material.

Um espectro típico de absorção de raios-x, tal como, na figura 3.1.3 geralmente é dividido em duas porções: o XANES e o EXAFS.[1, 12]

Figura 3.1.3: Espectro de absorção típico de uma folha de cobre metálico. Se distinguem três regiões que compõe o espectro, a pré-borda, a borda e a região estendida ou pós-borda. Fenômenos físicos diferentes ocorrem nas regiões de XANES e o EXAFS.



Atualmente, não há uma convenção que defina exatamente onde cada parte do espectro começa ou termina, contudo alguns autores sugerem que o XANES se alonga por cerca de

30 a 70eV após a borda de absorção. Uma definição mais precisa para a interface pode ser a energia a qual o comprimento de onda do foto-elétron ejetado atinja a distância interatômica mais próxima do átomo absorvedor, o que em termos práticos se corresponde com o intervalo de energias antes mencionado para a grande maioria dos compostos. Porém, ressalta-se que a energia que separa a região XANES do EXAFS (de até cerca de 2keV após o XANES) não pode ser definida universalmente, pois a transição é gradual e em qualquer caso muda de acordo com as distâncias típicas do vizinho mais próximo no sistema sob pesquisa. As interações que geram as oscilações além do XANES, ou seja, a região de EXAFS, estão relacionadas a fenômenos de espalhamento (*scattering*) das ondas fotoeletrônicas. Tais ondas, ao saírem e voltarem ao átomo absorvedor, produzem uma interferência característica que depende de diferentes parâmetros: distâncias de primeiros vizinhos, questões geométricas, elementos na vizinhanças, livre caminho médio do fotoelétron e fenômenos de *scattering* simples e múltiplo. A explicação física do EXAFS pode ser consultada em maiores detalhes nas referências [12, 13].

3.2 Análise Fatorial

Em diversas técnicas experimentais, em especial as de combinação linear, é necessário a compreensão acerca de sistemas de muitas variáveis. Nesse contexto, a análise fatorial, cuja definição é amplamente discutida, é uma ferramenta fundamental de análise estatística, sendo possível correlacionar numericamente as variáveis envolvidas no processo estudado. Segundo Malinowski (1980, p. 19) a definição de Análise Fatorial pode variar conforme o tempo, sendo uma possível definição: "A análise fatorial é uma técnica multivariada para reduzir uma matriz de dados à menor dimensão possível, sendo para isso utilizado o espaço de fatores ortogonais e as transformações que resultam em predições ou fatores reconhecíveis."[1] Malinowski(1980, p. 19, tradução nossa). [7]

3.2.1 Generalizações e notações

É possível definir um conjunto de dados **D**, tal que a matriz formada é expressa como:

$$\mathbf{D}(M \times N) = \begin{bmatrix} d_{00} & d_{01} & \dots & d_{0N} \\ d_{10} & d_{11} & \dots & d_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{M0} & d_{M1} & \dots & d_{MN} \end{bmatrix} \tag{3.2.1}$$

Os sub escritos em d_{ij} representam, respectivamente, a linha e a coluna em que o dado está posicionado na matriz de dados original, ou seja, o dado d_{ij} é o dado da linha i da coluna j. O objetivo da análise fatorial é decompor a matriz **D** em outras duas matrizes, tal que:

$$\mathbf{D} = \mathbf{RC} \tag{3.2.2}$$

Tal decomposição tem sentido puramente matemático, de modo que, as soluções para o modelo são conhecidas como componentes abstratas. De maneira geral, a decomposição de n fatores é dada por:

$$d_{ik} = \sum_{j=0}^{n} r_{ij} c_{jk} \tag{3.2.3}$$

[1] Factor analysis is a multivariate technique for reducing matrices of data to their lowest dimensionality by the use of orthogonal factor space and transformations that yield predictions and or recognizable factors.

sendo os termos r_{ij} e c_{jk} os fatores resultantes da decomposição. Há de se notar que a equação 3.2.3 nada mais é que uma definição de produto matricial. As matrizes **R** e **C**, não necessariamente possuem significado físico, e por isso, são comumente conhecidas como matrizes abstratas e podem ser denotadas por: R_{abs} e C_{abs}.

Apesar dos resultados apresentados até aqui não possuírem significado físico, o objetivo da análise fatorial é obter um modelo físico que permita realizar predições com o conjunto de dados. Para tanto, algumas transformações matemáticas são necessárias, frequentemente são conhecidas como rotações (apesar da possibilidade de ser uma transformação linear qualquer) e podem ser expressas em termos de **T** reversíveis ou que possuam pseudo-inversas.

Quando **T** é aplicada em matrizes abstratas é possível obter suas transformadas que, quando impostas às restrições corretas podem reduzir o espaço solução do problema. Dessa forma, a decomposição ficaria: [7]

$$\mathbf{D} = (\mathbf{R_{abs}T})(\mathbf{T^{-1}C_{abs}}) = \mathbf{R_{real}C_{real}} \tag{3.2.4}$$

Com as transformações é possível encontrar um conjunto de soluções fisicamente plausíveis, cujo o objetivo da análise fatorial é sumarizado na equação 3.2.4. Dentre as técnicas mais utilizadas para a transformação das matrizes abstratas está a *target transformation* e a matriz de transformação.

3.2.2 Quimiometria

A análise fatorial é a base da quimiometria computacional, e encontra-se uso de especial interesse em técnicas em que a Lei de Beer é válida. De maneira geral, a lei pode ser expressa como:

$$\frac{I}{I_0} = e^{-\alpha tc} \tag{3.2.5}$$

A equação relaciona a intensidade da luz incidente (I_0) e a intensidade que atravessa um meio (I). Além disso, a equação possui outros parâmetros como: a absortividade molar (α), a espessura da amostra (t) e a concentração do meio absorvente (c). Quando multiplicados (α, t, c) dão origem à absorbância (A). Ademais, há outros modos de expressar a mesma lei, tal como na equação 3.2.6 [7, 9]:

$$A = \alpha tc = -log(\frac{I}{I_0}) \tag{3.2.6}$$

Dessa maneira, sistemas que possam acompanhar a evolução da absorbância com a mudança de uma variável interna (comprimento de onda/energia) podem ser descritos com uma matriz **A** com M comprimento de ondas e N misturas, tal que [7, 9]:

$$\mathbf{A}(M \times N) = \begin{bmatrix} a_{00} & a_{01} & \dots & a_{0N} \\ a_{10} & a_{11} & \dots & a_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M0} & a_{M1} & \dots & a_{MN} \end{bmatrix} \tag{3.2.7}$$

Essas informações podem advir de diversas técnicas cujo princípio físico esteja relacionado com a Lei de Beer, tal como a absorção de raios-x. Nesse caso, a decomposição fatorial é dada pela equação 3.2.3, que pode ser adaptada para o caso de XAS [7, 9]:

$$A_{ik} = \sum_{j=0}^{n} e_{ij} m_{jk} \tag{3.2.8}$$

Nesse caso e_{ij} e m_{jk} são elementos das matrizes abstratas relacionadas com a i-ésima energia e a k-ésima mistura. Ainda assim, tais matrizes são componentes abstratas passíveis de transformações. Caso o conjunto de dados obedeça a Lei de Beer, cada fator (n) pode ser interpretado em termos químicos como as n componentes absorvedoras (espécies ativas ou puras). Para um modelo cujas transformações resultam em componentes reais, tem-se [7, 9]:

$$A_{ik} = \sum_{j=0}^{n} \alpha_{ij} c_{jk} \tag{3.2.9}$$

Desse modo, os elementos α e $\mathbf{C}$ são, respectivamente, a de absorvidade molar por unidade de energia e de concentração molar [7, 9].

No caso especial da absorção de raios-x, é possível simplificar a interpretação, uma vez que os dados obedecem a relação:

$$\frac{I}{I_0} = e^{-\mu t} \tag{3.2.10}$$

onde I é a intensidade do feixe incidente, I_0 a intensidade do feixe transmitido, t a espessura da amostra e μ é o coeficiente de absorção (ao variar a energia as medidas de μ resultam na matriz S^T). Isto é, dado uma matriz de dados $\mathbf{D}$ proveniente de XAS, é possível obter uma decomposição tal que [7, 9]:

$$\mathbf{D} = \mathbf{S}\mathbf{C}^T + \mathbf{R} \tag{3.2.11}$$

cujas linhas de $\mathbf{C}$ são os perfis de concentração das n espécies ativas, $\mathbf{S^T}$ são os n espectros puros e $\mathbf{R}$ os resíduos. Tal relação advém da equivalência entre as equações 3.2.9 e 3.2.10. Desse modo, pode-se entender que, quando uma amostra é composta por uma mistura de átomos absorvedores de diferentes espécies químicas, o espectro de XAS será uma média ponderada dos espectros de cada uma delas. Os espectros que se combinam para formar o conjunto de dados da amostra são denominados *espectros puros*. Assim, em uma medida de XAS $\mu(E)$ pode ser expresso como a soma das absorbâncias das espécies puras balanceados pela sua quantidade (concentração) na amostra, logo [7, 9]:

$$\mu(E) = \sum_{j=1}^{n} c_j s_j(E) + r \tag{3.2.12}$$

A equação 3.2.12 é uma forma mais generalizada da equação 3.2.11. Nessa representação, os elementos r da matriz de resíduos são o ruído experimental e o erro associado a decomposição do modelo. Além disso, devido a seletividade de elementos das medidas de XAS e a normalização espectro a espectro, é possível impor a condição de balanço de massas, isto é [9]:

$$\sum_{j=1}^{n} c_j = 1 \tag{3.2.13}$$

3.3 Técnicas de aprendizado de máquina não supervisionados

O aprendizado de máquina não supervisionado (do inglês, *non-supervised machine learning*) executa processos com dados não rotulados ou não classificados. Os modelos aprendem baseados na semelhança de dados, desse modo sua resposta está condicionada à ordenação destes conjuntos. Dentre estas técnicas, destaca-se para o PCA (do inglês, *Principal Component Analysis*) e o MCR-ALS (do inglês, *Multivariate Curve Resolution With Alternating Least Squares*).

3.4 Análise de componentes principais

Dentre as principais técnicas de aprendizado não supervisionado, o PCA é uma das estratégias utilizadas para reduzir a dimensão dos conjuntos de dados. Em estudos com catalisadores em reações *in situ*, em especial quando diferentes espécies co-existem na amostra, o PCA é aplicado para decompor a mistura em um número menor de componentes puras. Nesse contexto, seria uma tarefa complicada tentar classificar cada espectro do conjunto de dados, pois não há modelos que descrevam com boa precisão e sem ambiguidade o que está sendo observado. Este método obedece a decomposição descrita pela equação 3.2.12 [9, 14].

O objetivo geral da redução de dimensionalidade é minimizar a descrição do sistema, simplificando o conjunto de dados. Apesar de ser um método não supervisionado, a interpretação do resultado necessita de interferência humana para classificar espécies puras e, consequentemente, o que ocorre durante o processo químico. Em outras palavras, interpretar as mudanças de energia de absorção e de pré-picos, comuns na pré-borda do espectro de absorção, que podem caracterizar diferentes espécies químicas. [9, 14].

No caso ideal em que há um conjunto de dados com n espécies puras e sem ruído experimental presente, é possível que cada coluna da matriz de dados $\mathbf{D}$ (os espectros) seja representada como uma combinação linear das n espécies, assim, o número de linhas linearmente independentes de $\mathbf{D}$ é igual a n. Contudo, devido ao ruído na realidade experimental, cada coluna do conjunto de dados (ou a maioria) se torna linearmente independente , isto é, o posto matricial se torna igual ao número de colunas, sem ligação direta com o número de espácies ativas.

Apesar dessas restrições, o PCA quando combinado com o ITTFA (do inglês, *Iteractive Target Transformation Factor Analysis*), é capaz de encontrar e reduzir a dimensionalidade do sistema baseado em um número n de componentes. Dentre os principais algorítimos utilizados para realizar o PCA, a decomposição em valores singulares (SVD, do inglês *Singular Value Decomposition*) é a mais comum devido a precisão dos resultados [9].

3.4.1 Decomposição em Valores Singulares

A decomposição em valores singulares é uma fatoração de uma matriz retangular de dados. Seja $\mathbf{D}$ uma matriz retangular com M linhas e N colunas, o teorema do SVD diz que:

$$\mathbf{D}_{MxN} = \mathbf{U}_{MxM}\mathbf{S}_{MxN}\mathbf{V}^{T}_{NxN} \tag{3.4.1}$$

onde $\mathbf{U}$ e $\mathbf{V}$ são matrizes ortogonais, em termos matemáticos, $\mathbf{U}^{-1} = \mathbf{U}^{T}$ e $\mathbf{V}^{-1} = \mathbf{V}^{\mathbf{T}}$. As colunas de $\mathbf{U}$ são os autovetores à esquerda; S é uma matriz diagonal, cujos elementos são autovalores; e as linhas de $\mathbf{V}$ são os autovetores à direita. Por ser uma decomposição geral (funciona para qualquer posto matricial) é comum que os autovetores sejam denominados vetores singulares e os autovalores como valores singulares [9].

Seja $\mathbf{C}$ uma matriz de covariância N x N simétrica é possível diagonalizar a matriz, de modo que:

$$\mathbf{C} = \mathbf{V}\mathbf{L}\mathbf{V}^{\mathbf{T}} \tag{3.4.2}$$

sendo as colunas de **V** seus autovetores, e a diagonal de **L** seus autovalores em ordem decrescente. Quando os dados são projetados nos eixos principais (autovetores), a projeção é conhecida como componentes principais (*scores*), ou melhor, as colunas de **DV**. Fazendo uma analogia com a decomposição em valores singulares:

$$\mathbf{D} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathbf{T}} \tag{3.4.3}$$

então, a matriz de covariância pode ser reescrita como:

$$\mathbf{C} = \frac{\mathbf{V}\mathbf{S}\mathbf{U}^{T}\mathbf{U}\mathbf{S}\mathbf{V}^{\mathbf{T}}}{M-1} = \frac{\mathbf{V}\mathbf{S}^{\mathbf{2}}\mathbf{V}^{T}}{M-1} \tag{3.4.4}$$

assim, a matriz **V** de autovetores à direita representa as direções principais, os valores singulares são dados por $\lambda_i = \frac{s_i^2}{M-1}$ e as componentes principais por $\mathbf{DV} = \mathbf{U}\mathbf{S}\mathbf{V}^{T}\mathbf{V} = \mathbf{U}\mathbf{S}$. Como **U** e $\mathbf{V}^{T}$ são matrizes unitárias, suas colunas formam uma base para o espaço do conjunto de dados. Desse modo, o PCA tem por objetivo mudar a base do conjunto maximizando a variância explicada e reduzindo os resíduos.

3.4.2 Varimax

As técnicas que envolvem o PCA também oferecem métodos como o Varimax, mais conhecido como Varimax Rotation. Essa rotação em conjunto com o PCA tem o intuito de reduzir o número de componentes de um sub-espaço particular em termos das componentes majoritárias. O algorítimo tende a maximizar a variância do quadrado dos "*loadings*", cujos valores são dados pelo produto matricial entre **V** e **S**, para qual, cada coluna representa um *loading*. Essa maximização preserva a ortogonalidade da base obtida com o PCA, isto é, preserva o resultado do produto interno, deixando o sub-espaço invariante à rotação [15].

3.4.3 Interactive Target Transformation Factor Analysis (ITTFA)

Dentre os diversos algorítimos quimiométricos para resolução de um conjunto de dados de um processo acompanhado com espectroscopia, o ITTFA se destaca por ser um método de *self-modeling*, podendo ser aplicado após a análise de componentes principais (PCA) para obter a representação física das componentes abstratas obtidas no PCA. A partir das concentrações e espectros abstratos ($\mathbf{S_{abs}}$ e $\mathbf{C_{abs}}$), cujo sentido é puramente matemático, ou seja, não possuindo sentido espectroscópico, é possível realizar transformações no espaço para obter soluções. A ideia matemática da transformação pode ser expressa pelas seguintes equações [7, 9]:

$$\begin{aligned} \mathbf{D} &= \mathbf{S_{abs}}\mathbf{C^{T}_{abs}} \\ \mathbf{D} &= (\mathbf{S_{abs}}\mathbf{T})(\mathbf{T}^{-1}\mathbf{C^{T}_{abs}}) \\ \mathbf{D} &= \mathbf{S}\mathbf{C}^{\mathbf{T}} \end{aligned} \tag{3.4.5}$$

Desse modo, soluções abstratas podem ser transformadas através de uma matriz de transformação **T**, os critérios que envolvem a rotação das componentes variam de método para método, sendo os mais utilizados são: Varimax, Quartimax e Equimax. Dentre estes, o mais popular é o Varimax, que pertence ao grupo de rotações ortogonais, onde a dependência angular entre os autovetores é preservada [7, 9].

3.5 MCR-ALS

A resolução multivariada pode ser entendida como um conjunto de métodos e técnicas estatísticas capazes de lidar com múltiplas variáveis ao mesmo tempo. Dessa forma, permitem uma interpretação mais simples dos dados obtidos, podendo ser utilizadas na análise de misturas para encontrar informações desconhecidas, como por exemplo a concentração relativa ou os espectros puros de um conjunto de dados.

De maneira geral, o modelo do MCR consegue obter informações da mistura a partir de uma decomposição da matriz de dados experimentais em duas outras matrizes. Tal proposta é baseada na decomposição em mínimos quadrados:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^T + \mathbf{E} \tag{3.5.1}$$

sendo $\mathbf{D}$ uma matriz $M \times N$ de espectros medidos experimentalmente (QEXAFS ou outra técnica), $\mathbf{C}$ uma matriz $M \times P$ de contribuições relativas, $\mathbf{S}$ uma matriz $N \times P$ de espectros puros e $\mathbf{E}$ a matriz dos resíduos. De maneira geral, $\mathbf{C}$ pode ser vista como a matriz de combinações lineares de $\mathbf{S}$ com relação a $\mathbf{D}$. Além disso, como a equação 3.5.1 é a Lei de Beer para misturas com múltiplas componentes, a matriz $\mathbf{C}$ é comumente conhecida como concentrações relativas [16].

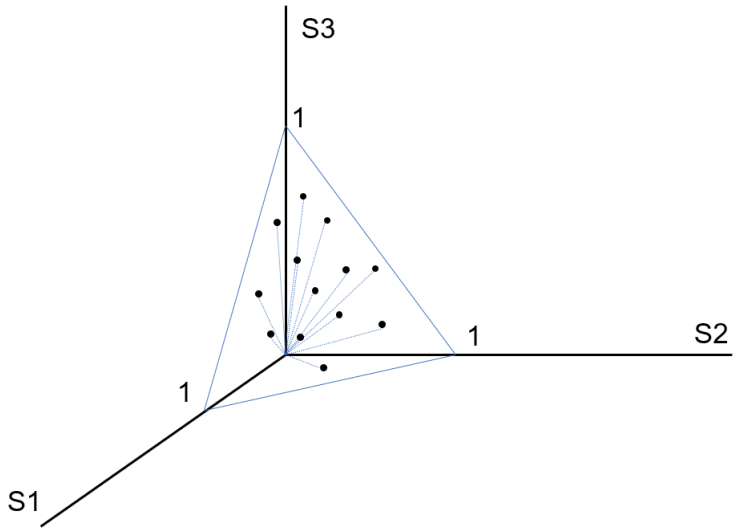
3.5.1 Estimativas iniciais

As estimativas iniciais são uma etapa crítica para o MCR-ALS, pois a probabilidade de obtenção de resultados significativos depende dos valores supostos para componentes ou concentrações. Nesse contexto, existem dois algoritmos que são comumente usados em análises de dados de absorção de raios-x: o SIMPLISMA (ou Pure) e EFA, ambos visam estimativas automáticas de linhas espectrais puras e concentrações.

3.5.2 SIMPLe-to-use Interactive Self-modeling Mixture Analysis

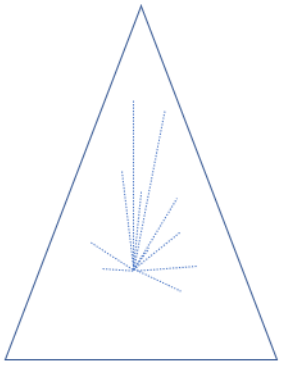
O método SIMPLISMA é conhecido por utilizar uma aproximação de componentes puras, ou seja, as respostas espectroscópicas mais puras ou mais diferentes de um conjunto de dados. Este método pode determinar quais das variáveis é formada por apenas uma componente, sendo que uma forma de visualizar o processo é imaginar que as concentrações de cada espécie estão em um hiperplano. Por exemplo, em um caso com três componentes, o vetor que liga a origem ao plano representa as concentrações relativas de cada espécie, e a soma das concentrações deve ser igual a um e pertencer ao plano da figura 3.5.1 [17]:

Figura 3.5.1: Exemplo visual do perfil de concentrações em uma análise multivariada. [2]



Ao realizar a projeção dos vetores de cada variável no plano triangular é possível determinar a pureza de uma variável através da norma ou tamanho do vetor. Uma representação da projeção está disposta na figura 3.5.2. Nesse caso, variáveis que são puras vão coincidir com os eixos de espécies puras (S1, S2, S3), dessa forma, os maiores vetores serão as componentes mais puras do conjunto de dados [17].

Figura 3.5.2: Projeção dos vetores das variáveis no plano de concentrações. [2]



Para calcular o comprimento das variáveis em termos da média e do desvio padrão de cada variável basta calcular como:

$$\lambda_i^2 = \mu_i^2 + \sigma_i^2 \tag{3.5.2}$$

Como o vetor μ é a distância entre a origem da variável e o plano triangular, e o vetor σ é a contribuição da variável na mistura. Uma maneira de garantir que a variável encontra-se, de fato, no plano triangular é limitar o comprimento do vetor λ, desse modo é garantida a projeção tal como na figura 3.5.2 [2, 17].

$$\mathbf{D} = \mathbf{C}^T\mathbf{P} + \mathbf{E} \tag{3.5.3}$$

Sendo **D** a matriz de dados originais, a matriz **C** é a das contribuições (não é exatamente uma matriz de concentrações), P^T as componentes puras e E o erro residual da decomposição. A resolução ocorre com a equação dos mínimos quadrados, até que os espectros ou contribuições puras são encontrados no conjunto de dados original [2].

O valor de pureza de uma variável é definido como a tangente entre os ângulos dos vetores μ e σ, dessa forma, para a i-ésima variável cujo índice 1 representa a primeira variável pura, a pureza é definida como:

$$p_{i,1} = \frac{\sigma_i}{\mu_i + \alpha} \tag{3.5.4}$$

Na equação da pureza é necessário considerar um erro α, pois para casos em que μ tende a zero é necessário corrigir o cálculo para evitar indefinições matemáticas.

Para as demais variáveis o cálculo é semelhante, contudo a pureza é multiplicada por um fator $w_{i,(2,3,4...)}$, este fator é o determinante da matriz de correlação ao redor da origem, que pode ser calculada utilizando a matriz de dados D, que é uma matriz E x S, onde E são as variáveis e S os espectros:

$$\mathbf{C} = \frac{\mathbf{1}}{\mathbf{S}}\mathbf{D}\mathbf{D}^\mathbf{T} \tag{3.5.5}$$

Desse modo, com os coeficientes calculados a matriz *w* é calculada baseado nos coeficientes de C:

$$w_{i,2} = \begin{bmatrix} c_{i,i} & c_{i,p_1} \\ c_{p_1,i} & c_{p_1,p_1} \end{bmatrix}$$

Assim, a segunda variável pura será:

$$p_{i,2} = w_{i,2}\frac{\sigma_i}{(\mu_i + \alpha)} \tag{3.5.6}$$

Para encontrar as demais variáveis basta repetir o método, expandindo a matriz *w* com os termos da segunda componente pura, até que o sistema seja composto por todas as componentes previstas [2, 17].

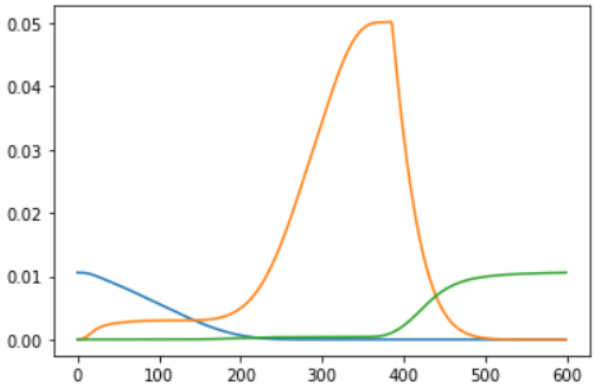
Caso a estimativa inicial seja os espectros puros (**S**), há uma matriz de concentração **C** que pode ser calculada utilizando o método de mínimos quadrados. Em outras palavras, minimizando o quadrado dos resíduos dos dados experimentais e dos espectros puros:

$$\mathbf{C} = \mathbf{D}^\mathbf{T}\mathbf{S}(\mathbf{S}^\mathbf{T}\mathbf{S})^{-\mathbf{1}} \tag{3.5.7}$$

3.5.3 Evolving Factor Analysis (EFA)

A Análise de Fatores Evolucionários é um método de análise multivariada capaz de gerar estimativas iniciais das concentrações, desde que os dados tenham uma ordem definida, ou seja, caso os dados de absorção sejam resultado de um experimento com resolução temporal. A ideia central do EFA é seguir a mudança no posto da matriz de dados em função dos dados ordenados. Um exemplo de estimativa inicial obtida está representado na figura 3.5.3, a qual apresenta concentrações não normalizadas.

Figura 3.5.3: Exemplo de concentração obtida via EFA, em ordem: Azul, laranja e verde.

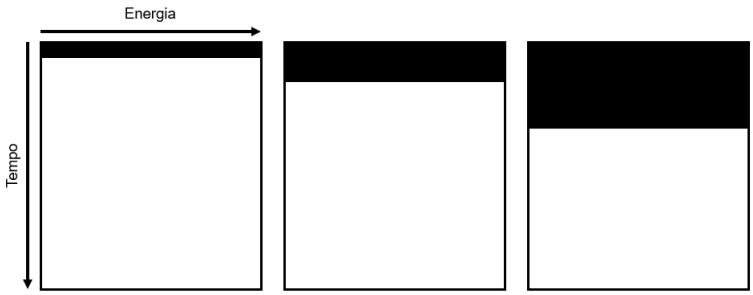


Como intuito de obter uma boa descrição matemática de um sistema, a matriz **D** deve ser decomposta em uma matriz **C** de concentrações, uma matriz de absorção molar **A** e uma matriz de ruido **E**, assim a matriz original é composta por:

$$\mathbf{D} = \mathbf{CA} + \mathbf{E} \tag{3.5.8}$$

A análise começa com a primeira linha da matriz **D** com o cálculo dos autovalores da primeira linha, em seguida a primeira e a segunda linha são utilizadas para calcular os autovalores, assim o processo continua até que chegue até a última linha, essa análise é conhecida como análise direta (forward analysis). A mesma análise é feita começando da última linha até chegar na primeira linha, esta segunda análise é denominada análise inversa (backward analysis). A ideia principal é realizar um tipo de Análise de Componentes Principais linha a linha, até que se faça com a matriz completa. Por fim, as duas análises são combinadas para obter a concentração inicial, assumindo que a primeira componente a surgir é a primeira a desaparecer, a segunda será a segunda a desaparecer e assim por diante. [3]

Figura 3.5.4: Exemplo visual do processo de *forward analysis*, autovetores são calculados de maneira progressiva na matriz de dados $\mathbf{D}^{\mathbf{T}}$. [3]



3.6 Ambiguidades

A decomposição matricial, da forma com que foi apresentada, é sujeita a dois tipos de ambiguidades. Tais ambiguidades são responsáveis primariamente pelo amplo conjunto de soluções possíveis para o sistema. Nesse sentindo, é possível descrever as ambiguidades das soluções em termos de uma matriz **T** *n x n* quaisquer. Se **T** for inversível é possível obter os mesmos resíduos **E** com:

$$\mathbf{X} = (\mathbf{CT})(\mathbf{T^{-1}S^T}) + \mathbf{E} \tag{3.6.1}$$

Se a matriz **T** for diagonal a ambiguidade é dita como multiplicativa, desse modo as soluções serão multiplicadas por um fator de **T**. No caso especial em que a matriz de espectros é tal que: $\mathbf{S_p S_p^T} = 1$ é possível que apenas os perfis de concentração sejam alterados pela ambiguidade multiplicativa. Contudo, o uso de restrições como fechamento e não negatividade (que serão destacadas nos próximos tópicos), reduzem o espaço de soluções e resolvem o problema multiplicativo [18].

Contudo, se **T** não for uma matriz diagonal, a ambiguidade é dita rotacional, neste caso o uso de outras restrições vão reduzir as possíveis soluções, porém não há como eliminar de fato a ambiguidade rotacional.

3.6.1 Mínimos Quadrados Alternantes (ALS)

Com as estimativas iniciais, seja das concentrações ou espectros puros é possível realizar a otimização baseada em mínimos quadrados alternantes. Esse "passo"evidencia a importância de uma boa estimativa inicial para o início de um processo de otimização. Quando aplicado com restrições o que ocorre é o descrito pelo algorítimo simplificado:

1. Computar estimativa inicial das concentrações $\mathbf{C} = \mathbf{C_0}$, utilizar o contador como i = 0;
2. Estimar os espectros puros com a relação: $\mathbf{S_{i+1}^T} = (\mathbf{C_i^T C_i})^{-1}\mathbf{C_i^T D}$;
3. Aplicar as restrições nos espectros obtidos;
4. Estimar $\mathbf{C_{i+1}}$ com a relação: $\mathbf{C_{i+1}} = \mathbf{DS_{i+1}}(\mathbf{S_{i+1}^T S_{i+1}})^{-1}$
5. Aplicar as restrições nas concentrações obtidas;
6. Avaliar os parâmetros estatísticos, como: desvio padrão entre duas interações e R^2, somar uma unidade no contador i;
7. Avaliar se o sistema convergiu ou divergiu baseado nas escolhas dos usuários;
8. Repetir o ciclo.

De modo geral, o programa busca resolver o sistema matricial com mínimos quadrados:

$$\mathbf{C} = \mathbf{DS}(\mathbf{S^T S})^{-1} \tag{3.6.2}$$

$$\mathbf{S^T} = (\mathbf{C^T C})^{-1}\mathbf{C^T D} \tag{3.6.3}$$

3.7 Restrições do MCR-ALS (Constraints)

Restrições, como estimativas iniciais, são críticas para alcançar a convergência, algumas dessas restrições podem ser aplicadas tanto às concentrações quanto aos espectros puros obtidos, no entanto, o uso excessivo de restrições pode levar a problemas de convergência e espectros sem significado físico. As restrições reduzem o número de soluções possíveis no processo mínimos quadrados alternantes, sendo estas essenciais para a redução/eliminação de ambiguidades, tanto multiplicativas quanto rotacionais.

3.7.1 Não negatividade (Non-Negativity)

Não negatividade pode ser aplicada a perfis de concentração e espectrais, impedindo que valores negativos sejam incluídos na resolução de mínimos quadrados em geral. Diferentes implementações, como mínimos quadrados no negativos (NNLS e FNNLS) e force-to-zero (FTZ), podem ser usadas.

3.7.2 Unimodalidade (Unimodality)

A unimodalidade é geralmente aplicada a perfis de concentração e está relacionada com uma moda única no perfil aplicado, isto é, os perfis de concentração possuem apenas um máximo local, dentro de uma certa tolerância.

3.7.3 Fechamento (Closure)

O fechamento está relacionado com o balanço de massa de uma reação, o algorítimo faz com que a soma de todas as concentrações seja igual a um valor pre-definido, geralmente igual a 100% ou 1, assim os perfis de concentração ficam sempre entre 0 e o valor escolhido para a soma.

3.7.4 Igualdade (Equality)

A igualdade é uma restrição que assume uma hipótese de que os perfis de concentração devem ser menores ou iguais a um certo valor, dessa maneira é necessário um conhecimento prévio acerca do sistema estudado para que possa ser aplicada a igualdade.

3.8 Otimização

Após o processo de seleção de componentes, estimativas iniciais e perfis de restrição a otimização é realizada com o algorítimo de ALS (já descrito anteriormente). Para avaliar a convergência ou divergência de uma otimização basta verificar a variação do desvio padrão de interações consecutivas, nesse caso:

- Caso a alteração no desvio padrão seja menor que 0.1% o sistema convergiu;
- Se a alteração no desvio padrão for negativa por mais de 20 vezes o sistema divergiu;
- Se o sistema não convergir no número de interações desejadas o algorítimo encontra o melhor R^2 e retorna o resultado para o usuário;

Além disso, outros parâmetros estatísticos também são de importância para avaliar a convergência do sistema, o %LOF que é a porcentagem de falta de ajuste e a variância explicada (R^2) são dois indicadores que podem revelar mais informações acerca da decomposição [16, 18].

$$LOF = 100 \times \sqrt{\frac{\Sigma e_{ij}^2}{\Sigma d_{ij}^2}} \tag{3.8.1}$$

$$R^2 = 100 \times (1 - \frac{\Sigma e_{ij}^2}{\Sigma d_{ij}^2})$$

3.9 Matrizes de transformação

No contexto das ambiguidades de soluções o método de matriz de transformação leva em conta que as decomposições não são únicas, isto é:

$$\mathbf{D} = \mathbf{C_{abs}}\mathbf{T}\mathbf{T}^{-1}\mathbf{S}^{\mathbf{T}}_{\mathbf{abs}} \tag{3.9.1}$$

Onde **T** é uma matriz de transformação quadrada *n* x *n*, onde *n* é o número de componentes puras do sistema. Por ser uma matriz tipicamente inversível a multiplicação $\mathbf{T}\mathbf{T}^{-1}$ não afeta em nada o resultado estatístico do sistema. Contudo, as decomposições espectrais se tornam diferentes, pois a multiplicação de $\mathbf{C_{abs}}$ por **T** gera uma nova matriz $\mathbf{C} = \mathbf{C_{abs}}\mathbf{T}$. Os elementos da matriz **T** podem ser rearranjados para se obter as componentes puras do sistema e buscar pelas melhores soluções. Apesar da versatilidade de soluções, utilizar o método tende a ser trabalhoso, pois o número de parâmetros da matriz cresce com n^2 [9, 19, 20].

Devido a normalização é possível restringir o número de elementos mutáveis de **T**, isto é, apenas $n^2 - n$ elementos da matriz serão responsáveis pela rotação. Para sistemas com duas componentes a restrição possibilita uma análise acerca das ambiguidades rotacionais do sistema estudado [9, 19, 20].

3.9.1 Restrições e espaço solução

As matrizes de transformação podem revelar as mais diversas soluções de um sistema de misturas, pois tais matrizes são responsáveis por transformar as componentes abstratas de um sistemas em componentes com sentido espectroscópico. Desse modo, é possível utilizar as matrizes de transformação para analisar como as restrições afetam as soluções do sistema, ou seja, analisar de modo quantitativo as como as restrições limitam o espaço de soluções [21].

3.10 Linha de luz QUATI

A linha de luz QUATI terá foco em experimentos de espectroscopia de absorção de raios-x com alta qualidade, resolução espacial e temporal em condições *in situ*. Devido a resolução temporal na escala de milissegundos, é esperado uma grande quantidade de dados advindas de medidas de processos cinéticos mais duradouros. Além disso, o intervalo de energia engloba energias de 4,5 até 35 keV, com um feixe de tamanho variável que pode ir desde 15 por 10 μm^2 na posição focal até 4 por 0,5 mm^2 a três metros dela.

Devido ao alto brilho, além das câmaras de ionização rápidas para a amostragem, a linha de luz QUATI permitirá aos usuários estudar cinéticas de reações, transições de fase e especies em superfície em condições de reação e com isso entender os sítios ativos das amostras, tendo para isso uma boa compreensão estrutural ou o que está acontecendo com a mesma quimicamente falando. Sendo assim é uma linha de luz de amplo interesse para a área de catálise, por exemplo. A tabela 3.10.1 representa os parâmetros esperados para a linha de luz.

Tabela 3.10.1: Parâmetros da linha de luz QUATI.

Parâmetro	Valor	Condição
Faixa de Enegia	4.5-12 keV | 8-35 keV	Si(111) | Si(311)
Resolução de energia ($\frac{\Delta E}{E}$)	10^{-4} | 10^{-5}	Si(111) | Si(311)
Tamanho do feixe	$15 \times 10\ \mu m^2$ até $5 \times 0.6\ mm^2$	45-48m

Fonte: Adaptado de: `https://lnls.cnpem.br/facilities/quati` . Acesso em: 09/02/2021 às 16:48.

3.11 Quick EXAFS

A técnica de Quick-EXAFS (QEXAFS), permite realizar medidas de espectroscopia de absorção de raios-x em escalas temporais na faixa de milissegundos. Assim, permite medidas rápidas de XANES e EXAFS, isto é, o XANES é primariamente responsável por revelar informações eletrônicas e estruturais da amostra, enquanto o EXAFS está relacionado a informações geométricas, como distância de primeiros vizinhos e desordem estrutural, por exemplo. Assim sendo, as medidas de QEXAFS fornecem ao pesquisador um método de investigação de reações *in situ* em condições *operando*, permitindo acompanhar a evolução de fases intermediárias em reações catalíticas [8, 9]

3.12 Jupyter

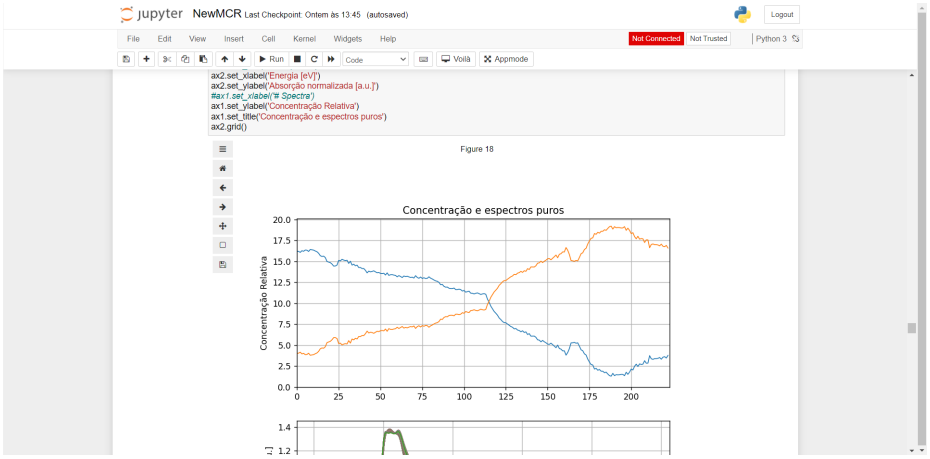
O Jupyter é um projeto de código-aberto, sem fins lucrativos que teve início com o desenvolvimento do projeto IPython em 2014. O principal objetivo do Jupyter é fornecer uma maneira interativa para a análise de dados da área de ciência dos dados. Apesar das bases em Python, o Jupyter suporta também outras linguagens de programação com ênfase em ciência computacional.

Por ser de uso gratuito e de fácil manipulação, o Jupyter tem sido adotado em diversas áreas, incluindo aceleradores de partículas. Sua interface gráfica permite ao usuário rodar códigos em Python de maneira rápida, fácil e reprodutiva. Por essas razões, o Jupyter é visto como ferramenta essencial na ciência de dados, permitindo usuários criarem cadernos experimentais que ajudam a organizar o fluxo de pensamento, os conhecidos Jupyter Notebooks [22].

3.12.1 Jupyter Notebook

O Jupyter Notebook é um ambiente que simula um caderno experimental comum, exceto que no Jupyter é possível atrelar as anotações com o poder do processamento computacional. Assim, é possível compilar ou, no caso do Python, interpretar pequenos trechos de um código, habilitando a experimentação com diferentes linguagens de programação. Um exemplo de resultado um Jupyter Notebook está na figura 3.12.1.

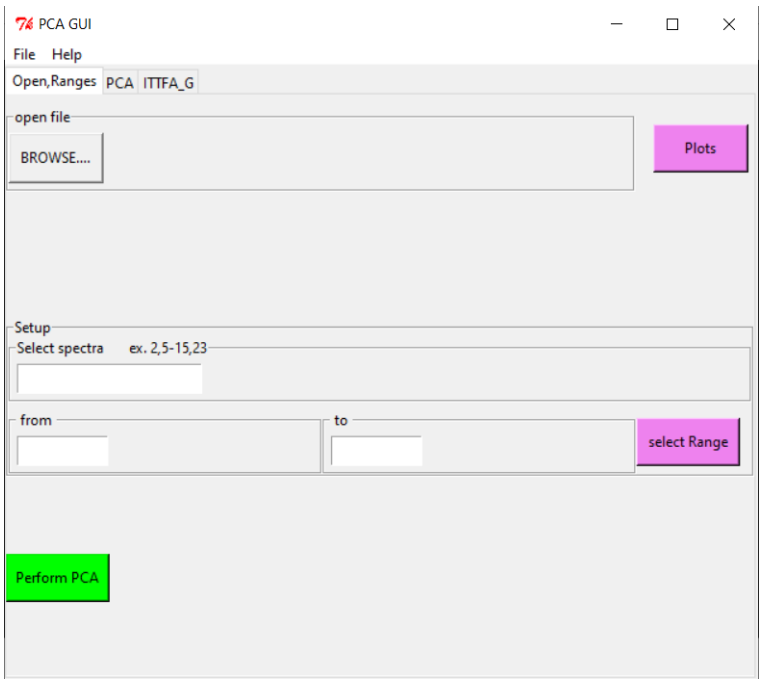
Figura 3.12.1: Otimização do MCR-ALS do CORAL em andamento no Jupyter Notebook.



3.13 PrestoPronto

PrestoPronto é um aplicativo baseado em Python e na plataforma Tkinter para analisar dados de linhas dispersas e arquivos provenientes de QEXAFS. Devido a implementação de uma interface gráfica que permite ao usuário interagir com os dados sem conhecimento prévio de programação seu uso é simples e objetivo. Nesse sentido, o programa permite analisar diversos espectros com Análise de Componentes Principais, Varimax e ITTFA[23]. Na figura 3.13.1 está apresentado a interface de usuário desenvolvida no Tkinter.

Figura 3.13.1: Interface de usuário do PCA oferecido pelo PrestoPronto.



Devido a sua implementação em Python e seu objetivo inicial de tratar grandes conjuntos de dados o Prestopronto é um programa extremamente versátil, além disso o paradigma de programação utilizado (orientação a objetos) facilita o bom entendimento de cada função do programa. Contudo, devido as limitações da época e o passar do tempo os produtos matriciais em cadeia não levam em conta o custo do produto matricial, desse modo uma otimização é recomendada para tratar matrizes de dados maiores. Além disso, como o programa foi pensado para dados de absorção de raios-x não há tanta flexibilidade da interface para tratamento de imagens, por exemplo.

Para melhor entender o funcionamento do programa um conjunto de espectros de uma amostra de níquel (figura 3.13.2) foi utilizada no programa. Com os espectros carregados é possível realizar a análise com diversos índices que indicarão quantas componentes o sistema possui, assim como indicado na figura 3.13.3. Para maiores informações acerca dos índices é possível consultar a referência [9].

Figura 3.13.2: Espectros utilizados para realizar a análise de componentes principais.

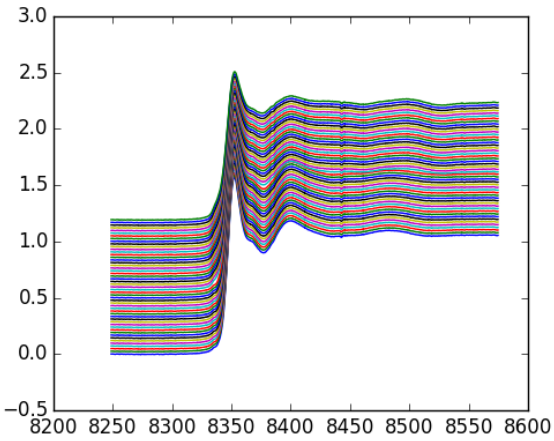
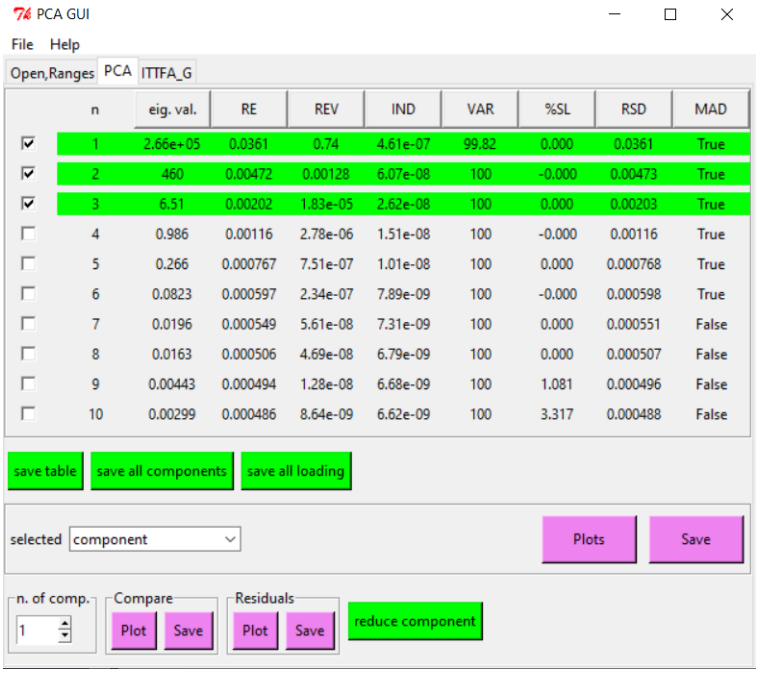


Figura 3.13.3: Janela PCA do aplicativo PCA_GUI do Prestopronto, os índices em verde representam as componentes selecionadas.



Após a seleção do número de componentes é possível utilizar técnicas de rotação como o Varimax e o ITTFA para transformar as componentes abstratas em componentes com sentido físico, isto é, trasformar os autovetores da decomposição em valores singulares em espectros de absorção interpretáveis e perfis de concentração.

3.14 Otimizações do CORAL

A implementação do PCA do programa PrestoPronto no CORAL levou em conta as atualizações necessárias para que o programa tivesse compatibilidade com as novas versões de Python. As principais mudanças foram meramente semânticas e matemáticas, sem mudanças significativas no algorítimo utilizado. Também foram consideradas novas variáveis como o custo de produto matricial e a utilização de GPU para acelerar as decomposições matriciais.

3.14.1 Custo do produto matricial

Como citado anteriormente, o Prestopronto não levava em conta os produtos matriciais em cadeia, ou seja, o produto de três ou mais matrizes. Desse modo, no caso de uma matriz com 180.000 espectros e 1281 pontos de energia o programa gerava uma matriz 180.000x180.000, matrizes dessa ordem podem ocupar muito mais de 240GB de RAM. Além disso, tal matriz é apenas um passo intermediário para o cálculo das componentes reais. Para fins de comparação, uma matriz com 180.000 espectros representa 5 horas de medidas na linha QUATI, desse modo o limite do programa pode ser facilmente alcançado no servidor utilizado.

Para realizar a otimização, a implementação do CORAL levou em consideração a ordem multiplicativa e o custo das operações matricialistas. Como resultado, o programa passou a aceitar um número muito maior de espectros (considerando o servidor utilizado), enquanto o limite anterior era de cerca de 82.000 espectros, com a otimização, a aplicação conseguiu aceitar mais de 1.000.000 espectros, completando a análise em cerca de 15 minutos. Com medidas a cada 100ms, um milhão de espectros equivale a mais de 27 horas de medições simultâneas na QUATI. Nas versões anteriores do software, uma matriz como esta exigia 7 terabytes de RAM, tornando impossível fazê -lo mesmo em clusters mais poderosos. No entanto, após a otimização, o mesmo cálculo poderia ser feito com apenas aproximadamente 50 terabytes de RAM.

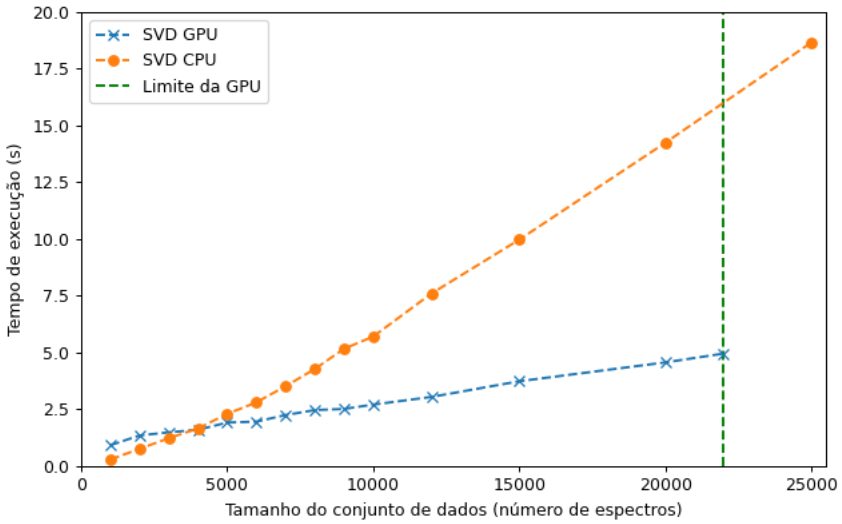
3.14.2 Aceleração com placa de vídeo

A plataforma *CUDA* foi criada pela NVIDIA para permitir o paralelismo computacional em GPUs. Usando CUDA, é possível acelerar o processamento usando a capacidade rápida de processamento de matrizes da GPU . Em Python, existe uma implementação em código aberto de uma biblioteca que usa bibliotecas CUDA como cuSPARSE , cuBLAS, cuDNN , cuSOLVER, cuRAND e NCCL para fornecer aceleração de processamento de GPU, o CuPy . Além disso, o CuPy possui uma sintaxe semelhante à da biblioteca de processamento numérico NumPy , aceitando uma ampla gama de dados e métodos , incluindo métodos de álgebra linear e álgebra tensorial. Devido a semelhança com o NumPy, é comum que muitas funções implementadas em NumPy, e até algumas do SciPy, estejam disponíveis em CuPy.

No entanto, embora a biblioteca forneça acesso a processamentos mais rápidos, as placas de vídeo só oferecem vantagens sobre os processadores quando o número de pontos de dados é significativamente maior do que o número de pontos de dados comumente manipulados pela CPU, em termos de espectros: cerca de 3400. Como resultado , acima de um certo número de espectros, é possível observar um aumento de velocidade em uma variedade de implementações, desde operações element - wise (elemento a elemento) até decomposições mais complexas , como Decomposição em Valores Singulares (SVD, do inglês *Singular Value Decomposition*).As acelerações, por exemplo, podem resultar em uma melhoria de 200% no tempo de processamento. A Figura 3.14.1 é uma representação gráfica do tempo de processamento do algoritmo SVD para vários tamanhos de conjuntos de dados.

Apesar das placas de vídeo apresentarem grande vantagem em questões temporais, ainda há um fator limitante: a quantidade de memória de vídeo. Enquanto um processador depende da memória RAM para um bom funcionamento, as placas de vídeo possuem memórias integradas e, de maneira geral, não expansíveis. Assim, um dos fatores limitantes para o uso de GPUs é a quantidade de memória disponível. No mesmo servidor em que os testes ocorreram o limite foi de 22.000 espectros, mesmo com a otimização do produto matricial.

Figura 3.14.1: Em laranja o tempo de processamento da CPU e em azul o da GPU, ambos utilizando o algoritmo de SVD.

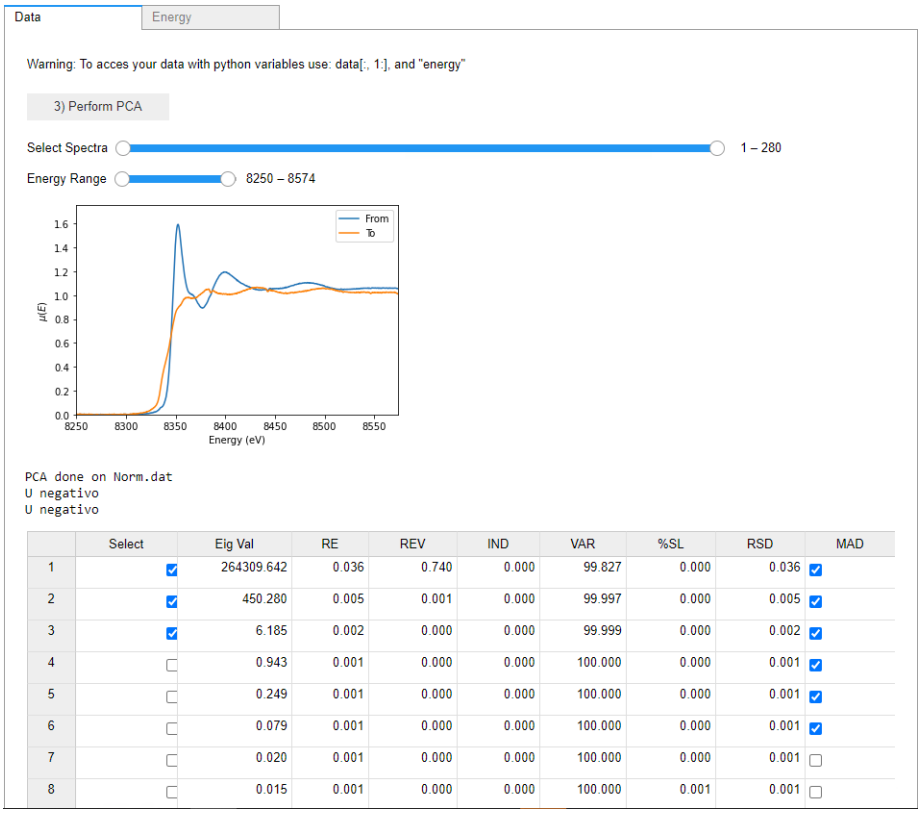


3.14.3 Interface de usuário para Jupyter

As interfaces de usuário (GUI, ou *Graphical User Interface*) são um elemento da experiência do usuário (UX, ou *User Experience*), que é definida como um conjunto de fatores relacionados ao tipo de interação que um usuário tem com um produto. As experiências podem ser classificadas como positivas, negativas ou neutras, portanto, ter uma boa interface é fundamental para aumentar fatores como produtividade e engajamento.

Em termos de implementação do PrestoPronto PCA no Jupyter, foi possível fornecer uma interface gráfica de usuário para facilitar o uso do programa. O principal objetivo da interface é familiarizar os usuários com o Jupyter Notebook sem remover a possibilidade de implementação de código após uma análise com as interfaces Jupyter . Tal exemplo de interface está disposto na Figura 3.14.2.

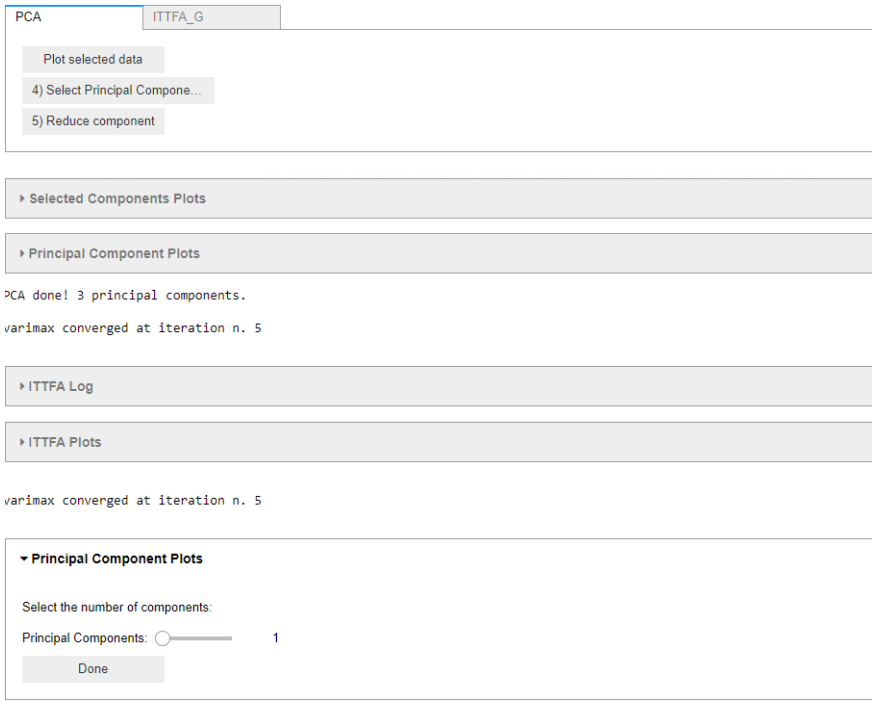
Figura 3.14.2: Interface gráfica do PCA no Jupyter Notebook.



	Select	Eig Val	RE	REV	IND	VAR	%SL	RSD	MAD
1	☑	264309.642	0.036	0.740	0.000	99.827	0.000	0.036	☑
2	☑	450.280	0.005	0.001	0.000	99.997	0.000	0.005	☑
3	☑	6.185	0.002	0.000	0.000	99.999	0.000	0.002	☑
4	☐	0.943	0.001	0.000	0.000	100.000	0.000	0.001	☑
5	☐	0.249	0.001	0.000	0.000	100.000	0.000	0.001	☑
6	☐	0.079	0.001	0.000	0.000	100.000	0.000	0.001	☑
7	☐	0.020	0.001	0.000	0.000	100.000	0.000	0.001	☐
8	☐	0.015	0.001	0.000	0.000	100.000	0.001	0.001	☐

A interface modularizada garante ao usuário flexibilidade para realizar múltiplas análises. Desse modo, é possível escolher o número de componentes, analisar os autovalores e reduzir o espaço utilizando o Varimax e o ITTFA em um mesmo caderno experimental. Além disso, também é possível realizar múltiplas análises no mesmo caderno. Para implementar as interfaces em Jupyter a biblioteca IPyWidgets foi de fundamental importância, a modularização está representada na figura 3.14.3.

Figura 3.14.3: Modularização do PCA no mesmo caderno experimental.



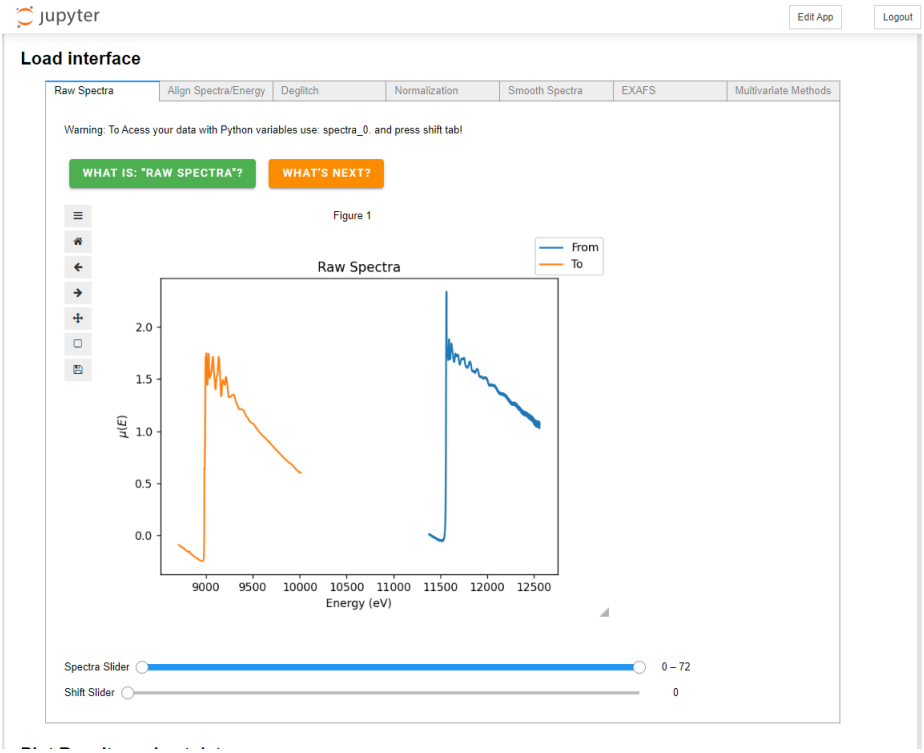
3.15 CORAL

O Coral também possuí uma interface única para tratamento de dados, isto é, integrada com o X-ray Larch, desenvolvida por Matt Newville, a biblioteca fornece ao usuário uma maneira simples de tratar os dados provenientes da linha de luz diretamente no caderno experimental, possibilitando o desenvolvimento de uma análise única, de modo a otimizar o tempo de linha do usuário. [24]

3.15.1 Proposta de interfaces

Para evitar problemas com usuários que não estão familiarizados com programação, foi desenvolvida uma interface gráfica utilizando IPyWidgets. Além disso, a interface é simples de usar e semelhante às já disponíveis em programas semelhantes. Ainda assim, há uma implementação compatível com a extensão AppMode , que é responsável por gerar um aplicativo *Web* a partir de um Jupyter Notebook, por questões estéticas e de proteção de código fonte.

Figura 3.15.1: Representação da interface desenvolvida em Jupyter Notebook.



Dentre as diversas funcionalidades propiciadas pelo aliciamento de ambas as bibliotecas (IPyWidgets e AppMode), é a possibilidade de mostrar avisos aos usuários. Assim, é possível guiar, explicar funcionalidades ou explicitar mensagens de erro de maneira interativa. Essa funcionalidade visa responder perguntas que podem ocorrer de maneira frequente acerca do uso e funcionamento do programa.

Figura 3.15.2: Interface de avisos, desenvolvida visando alertar o usuário acerca de problemas ou erros.

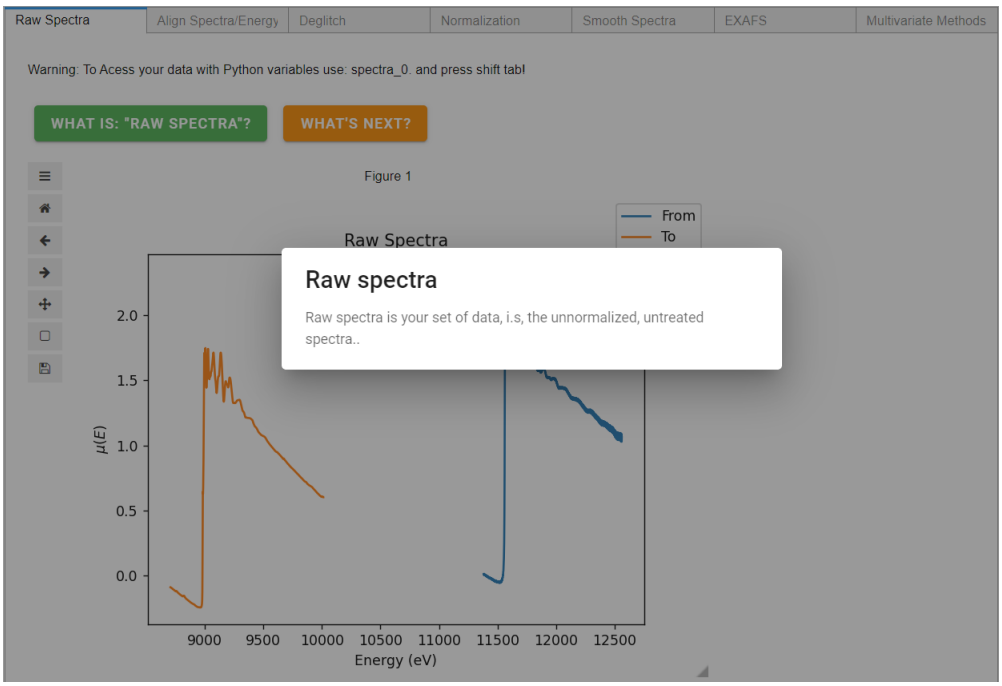
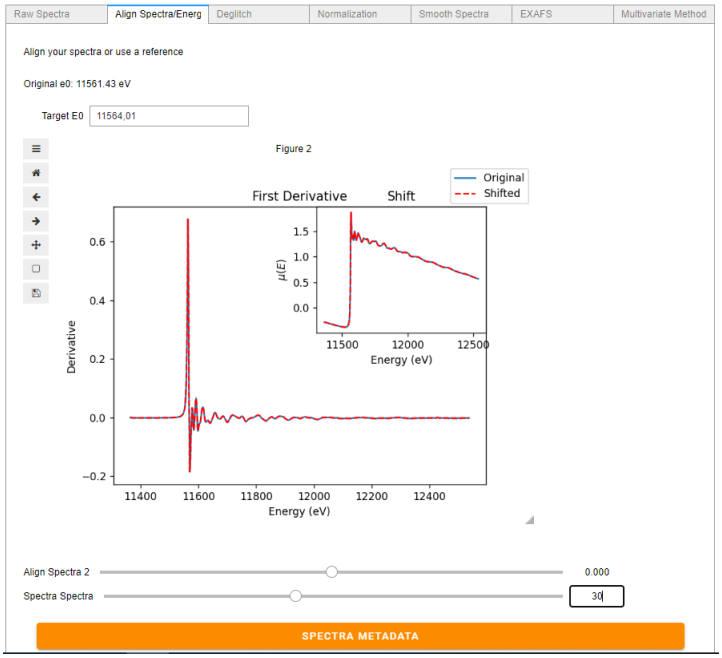


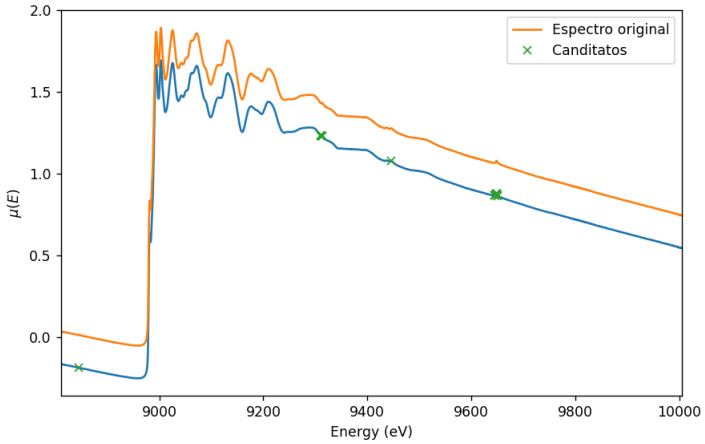
Figura 3.15.3: Interface de alinhamento de espectros, em azul a referência e em vermelho o dado que será alinhado.



3.15.2 Remoção de efeitos espúrios no sinal ou *Glitches*

Dentre as novas implementações para o CORAL está o algorítimo de Deglitch proposto por Wallace, onde um filtro de Savitzky-Golay é aplicado aos espectros, os pontos com maior divergência do filtro são considerados como *glitches*, ou melhor, são considerados como pequenas variações que não foram corrigidas devidamente pela compensação usual das câmaras de ionização ou fotodiodos utilizados na linha de luz [25].

Figura 3.15.4: Algorítimo de *deglitch* em funcionamento, em azul com marcações os principais candidados a *glitches*.



O algorítimo original elimina os pontos que considera ser *glitches*, enquanto o algoritmo modificado remove o ponto, faz uma interpolação com um polinômio e adiciona o novo ponto

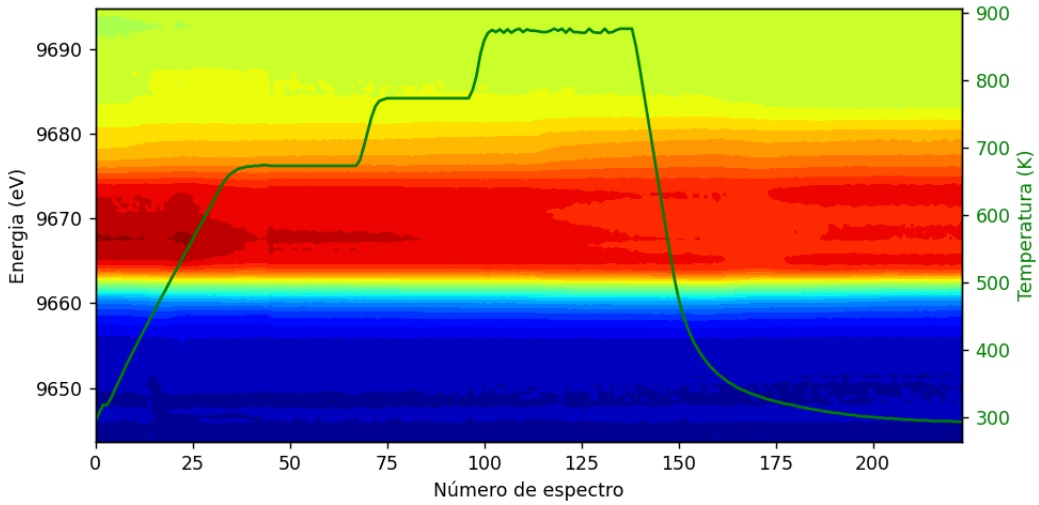
ao gráfico, assim evitando problemas futuros com a resolução multivariada, que é extremamente dependente de um conjunto de dados com a mesma quantidade de pontos experimentais.

4 Resultados e Discussões

4.1 Conjuntos de dados

Para o presente estudo foram utilizados dados de absorção de raios-x provenientes da linha de luz DXAS do antigo acelerador de partículas brasileiro, o UVX, medidos em 2005. O primeiro conjunto de dados $\mathbf{D_1}$ é um conjunto de medidas na borda K do zinco. Os espectros foram utilizados para estudos da evolução do estado de uma ferrita de zinco de tamanho nanométrico. O conjunto de dados foi utilizado devido às informações advindas de técnicas complementares como espectroscopia Mössbauer e simulações computacionais para validar o modelo. Além disso o conjunto de dados é composto unicamente por duas componentes, o que facilita as primeiras análises [4].

Figura 4.1.1: Mapa de curvas de nível dos espectros de absorção de raios-x em função da temperatura. Enquanto a cor azul corresponde à pré-borda, à linha branca (borda de absorção) é vermelha. Os aquecimentos ocorreram a 10K/min (em verde) e a pequena descontinuidade na linha branca na região de 40-50 (número de espectro) ocorre devido à saturação de intensidade do sinal.

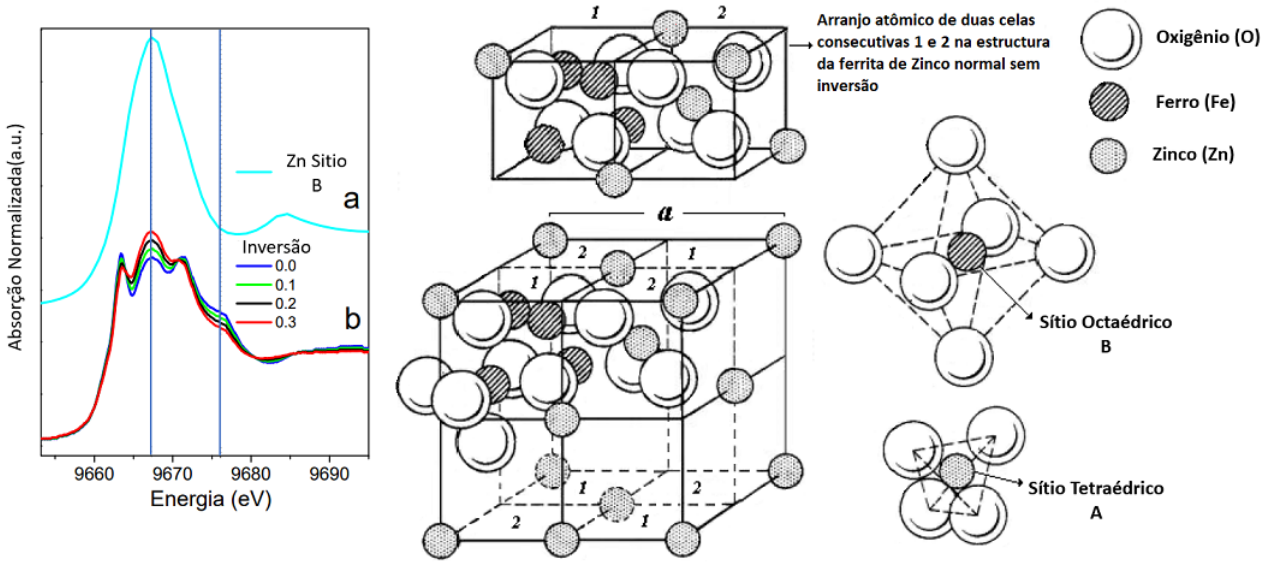


A figura 4.1.2 traz um esquema dos sítios A e B das ferritas de zinco e seus espectros simulados com o programa FEFF [4]. Em azul claro o espectro XANES normalizado das ferritas em sítio B, ou ferrita invertida, já em azul escuro o espectro XANES normalizado das ferritas em sítio A, ou ferrita normal, e em diferentes cores a combinação linear (com pesos: 0; 0,1; 0,2 e 0,3) do espectro do sítio B com o de $ZnFe_2O_4$ em seu estado normal. Como a síntese do composto completamente invertido (Zn em sítio B em toda a estrutura) não é possível, não há referências em bases de dados, sendo necessário recorrer a simulação computacional para validar a estrutura apresentada pelo XANES da amostra, principalmente ao longo do processo experimental previamente descrito 4.1.1.

Além disso, a falta de referência do estado totalmente invertido impede que o MCR-ALS seja iniciado com espectros puros de cada componente da amostra, sendo necessário buscar outros parâmetros da reação química, como taxas de reação ou valores de concentração para que o modelo de análise multivariada possa convergir para um resultado com um maior sentido espectroscópico. Caso isso não seja possível, os erros associados às ambiguidades serão predominantes na decomposição espectral, logo, é possível que os pesquisadores percam interesse

nessa técnica de análise de dados.

Figura 4.1.2: Esquema das ferritas de zinco em estado normal e seus espectros de absorção. Na estrutura normal os átomos de Zinco ocupam os sítios tipo A tetraédricos e os Ferros os tipo B octaédricos. Quando há inversão, o Zinco migra para o sítio B e o Ferro para o sítio A. No caso os espectros de absorção se alteram como mostrado em (a) e (b) na figura, sendo em azul escuro o correspondente à ferrita de zinco em estado normal sem inversão. Os espectros do conjunto de dados, idealmente são combinações lineares dos espectros do Zn em sítio A e B.[4]



4.1.1 Problema proposto

Devido a dificuldade de síntese de uma das espécies puras, não foi possível utilizar o espectro de cada uma delas como estimativa inicial do MCR-ALS. Desse modo, haviam algumas soluções possíveis para entender a reação em termos dos espectros e das concentrações provenientes da decomposição de $\mathbf{D_1}$. Uma das possibilidades para a resolução do problema é simular espectros puros e comparar com o conjunto de dados (como os simulados da figura 4.1.2), a outra é utilizar algum método multivariado, como PCA ou MCR-ALS. Contudo, o uso de tais métodos é limitado, pois em uma decomposição há intrinsecamente uma ambiguidade rotacional, desse modo, utilizar poucas restrições pode gerar resultados que não descrevem corretamente o sistema físico. Assim, avaliar como os espaços de solução variam conforme diferentes restrições são aplicadas, pode ajudar a entender o efeito das restrições em sistemas de múltiplas componentes.

4.2 Espaço de soluções

Na presença de ambiguidades rotacionais ou ARs, há um conjunto de soluções prováveis denotado pelas matrizes $\mathbf{C_{sol}}$ e $\mathbf{S_{sol}^T}$. No caso da decomposição em valores singulares, a ambiguidade é vista como as possíveis combinações lineares das componentes abstratas do sistema. Seja $\mathbf{D}$ uma matriz proveniente de dados de absorção contendo M linhas (resolução energética) e N colunas (espectros de absorção), a decomposição em valores singulares pode ser feita por:

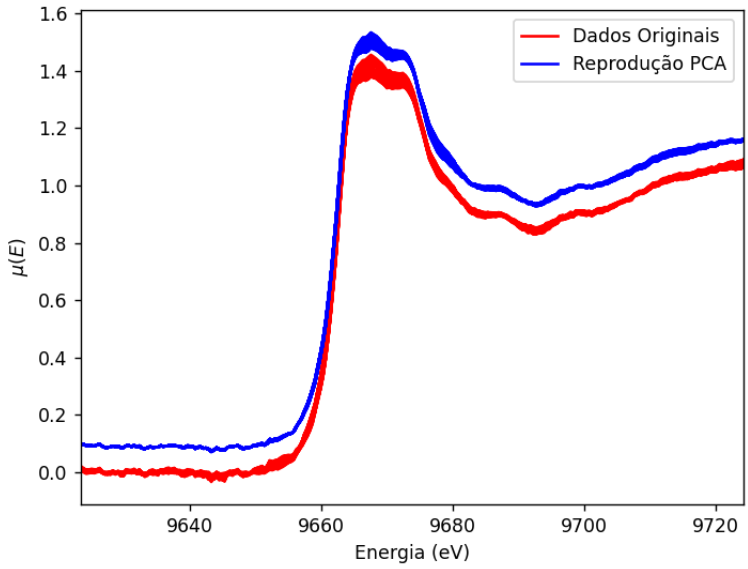
$$\mathbf{svd}(\mathbf{D^T}) = \mathbf{U\Sigma V^T}$$

Em primeira análise, é possível reduzir o número de componentes utilizando as matrizes reduzidas. Ao diminuir a dimensionalidade do sistema, parte do ruído experimental também é reduzido, pois este é considerado como uma das componentes que formam a matriz de dados $\mathbf{D}$. Esta processo é dado é dada por:

$$\mathbf{U_r} = \mathbf{U}[:,:2]\mathbf{S_r} = \Sigma[:2]\mathbf{V_r^T} = \mathbf{V^T}[:2,:]$$

onde o índice " : " representa todas as linhas ou colunas. Por exemplo: $\mathbf{U_r}$ é a matriz reduzida das duas primeiras colunas para todas as linhas de $\mathbf{U}$. Com isso, a reprodução dos dados com as matrizes reduzidas causa uma pequena diferença no conjunto, pois há uma minimização de ruído experimental. A figura 4.2.1 representa a redução de ruído utilizando PCA, isto ocorre pois, como explicado anteriormente, o erro é uma das componentes do conjunto de dados, pois ao restringir-se o número de componentes para apenas duas, o ruído é suprimido.

Figura 4.2.1: Espectros de absorção normalizados e reprodução do conjunto de dados com PCA.



Para avaliar as ambiguidades em sistemas multicomponentes é possível utilizar uma matriz de transformação $\mathbf{T}$ $(n \times n)$, onde n é o número de componentes puras que constituem o sistema. De modo geral, o objetivo das matrizes $\mathbf{T}$ e $\mathbf{T^{-1}}$ é gerar um conjunto de combinações lineares das componentes abstratas, que são espectros gerados pelo PCA sem sentido físico, mas com grande concordância matemática em relação ao conjunto de dados. Devido a normalização dos espectros, é possível inferir que ao menos um elemento de cada linha da matriz de transformação será igual a uma constante α. Para duas componentes a matriz de transformação será:

$$\mathbf{T} = \begin{bmatrix} \alpha & T2 \\ \alpha & T4 \end{bmatrix}$$

Dessa forma, há apenas $n^2 - n$ parâmetros variáveis, isto é, apenas T2 e T4. Assim, as ambiguidades do sistema devem estar relacionadas aos parâmetros livres da matriz de transformação. Contudo, é necessário estabelecer os valores α que tornam a matriz completamente normalizada. Para tanto é possível utilizar a estimativa inicial previamente implementada no MCR-ALS, o SIMPLISMA.

O SIMPLISMA é capaz de estimar quais são os espectros puros que compõe o sistema e, com esses espectros, calcular a matriz de transformação ($\mathbf{T_0}$) com *Target Transformation* e os resultados do SIMPLISMA ($\mathbf{S^T}$). Desse modo, a matriz $\mathbf{T_0}$ que gera a melhor transformação de $\mathbf{V_r}$ em $\mathbf{S}$ é dada por:

$$\mathbf{VrT_0} = \mathbf{S^T}$$
$$\mathbf{T_{inicial}} = \mathbf{T_0^T}$$

Como a equação matricial acima é do tipo $\mathbf{Ax} = \mathbf{B}$, é possível encontrar $\mathbf{T_0}$ com mínimos quadrados, ou utilizar espectros de referências ou resultados do ITTFA e MCR-ALS para calcular a matriz de transformação inicial. Para dados bem normalizados, é possível impor a condição de igualdade do parâmetro α. Além disso, para avaliar a qualidade do resultado das rotações é possível calcular a soma dos quadrados dos resíduos e gerar um *Grid Map* que é um mapa do logarítimo do *ssq* (com relação ao PCA ou ao experimento) em função de T2 e T4, em termos matemáticos:

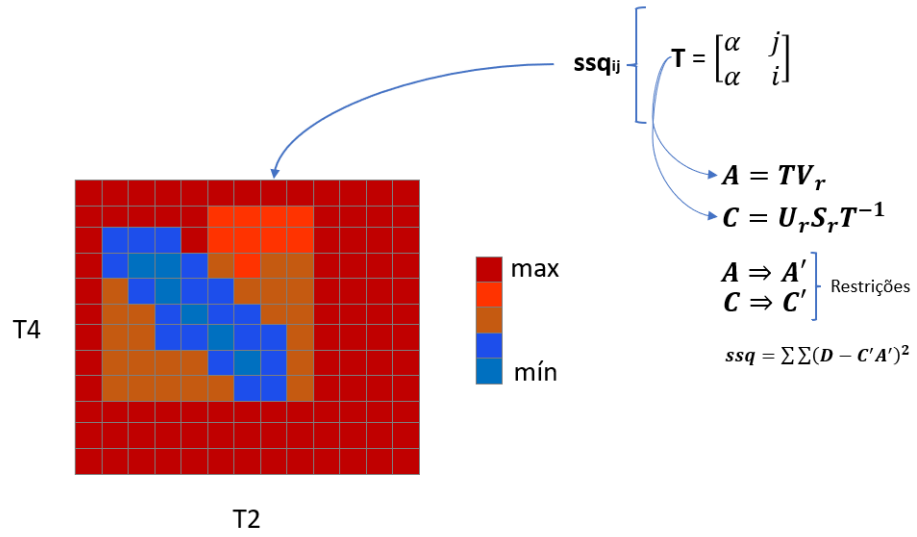
$$ssq[PCA] = \sum\sum(\mathbf{U_r S_r V_r} - \mathbf{C'A'})^2$$
$$ssq[EXP] = \sum\sum(\mathbf{D_1} - \mathbf{C'A'})^2$$

onde $\mathbf{C'}$ é a matriz de concentrações com restrições aplicadas (não negatividade, fechamento, unimodalidade, igualdade, dentre outras) e $\mathbf{A'}$ é a matriz de espectros restrita, que podem ser calculadas utilizando as definições já discutidas na seção de introdução teórica, a reprodução de $\mathbf{A}$ e $\mathbf{C}$ é dada pelas seguintes relações:

$$\mathbf{A} = \mathbf{TV_r^T}$$
$$\mathbf{C} = \mathbf{U_r S_r T^{-1}}$$

A figura 4.2.2 traz uma breve representação do método utilizado para gerar mapas de soluções com a soma dos quadrados dos resíduos.

Figura 4.2.2: Esquema de geração do mapa de soluções com *ssq* em relação ao experimento. Para melhorar a visualização de áreas de solução, é possível utilizar escala logarítmica.



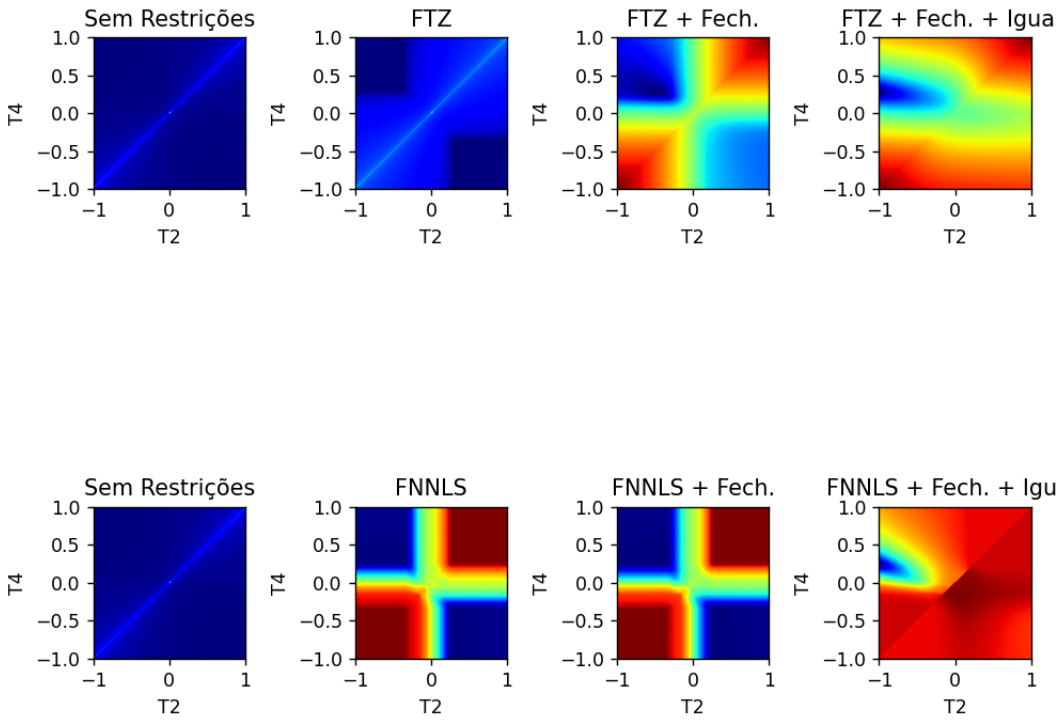
No caso em que o *ssq* calculado para o sistema $\mathbf{A}'$ e $\mathbf{C}'$ é mínimo, é possível aproximar as matrizes $\mathbf{A}' = \mathbf{S}_{\mathbf{sol}}^{\mathbf{T}}$ e $\mathbf{C}' = \mathbf{C}_{\mathbf{sol}}$ e utilizar soluções próximas como estimativa do erro associado à decomposição. Geralmente não há apenas um ponto mínimo, mas um conjunto de soluções que define uma área de solução plausível. Os somatórios do *ssq* não possuem índice, pois a soma é feita tanto nas linhas quanto nas colunas.

4.2.1 Efeito das restrições nos mapas de solução

Para avaliar os efeitos das restrições (*constraints*) no espaço de solução T2T4 foram aplicadas duas estratégias, a primeira consiste na avaliação do sistema sem restrições, seguido pela aplicação de *constraints* uma a uma, utilizando o método de forçar para zero (FTZ, do inglês *Force to zero*). Já segunda consiste em uma avaliação semelhante à primeira, sendo aplicada uma outra restrição de não negatividade. Para tanto foi utilizado o *FNNLS* (do inglês, Fast Non-Negative Least Squares), que é um método de mínimos quadrados com coeficientes da matriz $\mathbf{x}$ ($\mathbf{Ax} = \mathbf{B}$), diferentes de zero.

O efeito das restrições fica claro na figura 4.2.3, onde as zonas azuis representam soluções com baixo *ssq*, portanto boas soluções. Enquanto o sistema sem restrições apresenta diversas soluções ótimas (zona azul), os sistemas com mais restritos constituem um espaço mais seleto de combinações de T2 e T4. Apesar da condição de não-negatividade ser uma imposição natural, na maioria dos sistema químicos sua aplicação não apresenta localidade de solução, isto é, não apresenta um ou dois poços de soluções ótimas. O mesmo vale para sistemas com fechamento (*closure*) cujas soluções são reduzidas, mas apresentam amplo espaço para ambiguidades rotacionais.

Figura 4.2.3: Efeito da aplicação de restrições em um conjunto de dados com duas componentes puras.



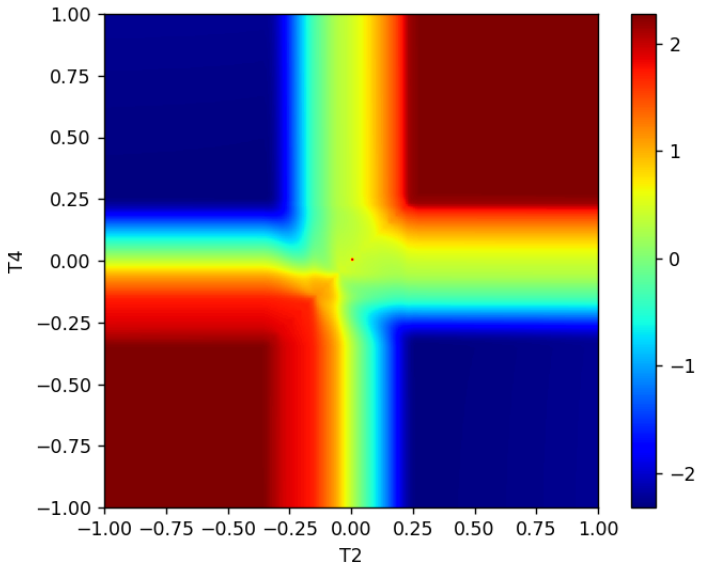
Por outro lado, a condição de igualdade, também conhecida como igualdade de posto local, reduz significativamente o número de soluções, isolando completamente uma região do mapa de soluções. Assim, é possível observar que, para o conjunto de dados utilizado, restrições naturais como não negatividade, unimodalidade e fechamento não são suficientes para uma boa redução do espaço de soluções. Isto indica que soluções ótimas com menor ambiguidade podem ser encontradas aplicando *hard-modeling*, isto significa a aplicação de conhecimento prévio na otimização do MCR-ALS.

Também é possível notar que em todos os casos, as soluções cuja condição T2 = T4 é satisfeita, são incongruentes, em razão das matrizes singulares. Portanto, caso T2 seja igual a T4, a matriz de transformação **T** é singular, então não é possível calcular a matriz inversa, apenas sua pseudo inversa de Moore-Penrose. Em regiões onde o determinante da matriz **T** é próximo de zero, os valores da soma dos quadrados dos resíduos tendem a ser altos para determinados tipos de restrição.

4.2.2 Simetria de espaços de solução

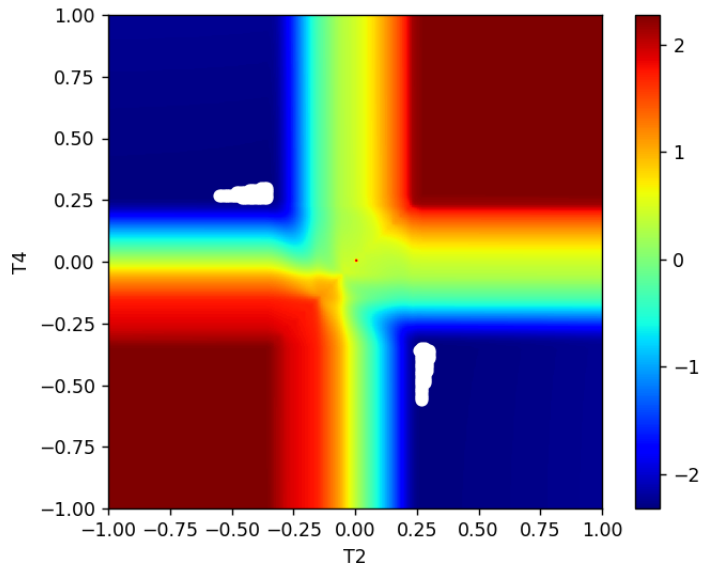
Quando pouco restringidos, os resultados do MCR-ALS tendem a apresentar maior ambiguidade, assim como demonstrado pelas curvas de SSQ em função de T2 e T4. Contudo, é possível notar que em boa parte das soluções há uma linha de simetria, isto é, há soluções ótimas para valores em que ssq(T2, T4) = ssq(T4, T2). Por mais restrito que o sistema da figura 4.2.4 aparente ser, quando comparado com os modelos com menos restrições, apresenta regiões cujas soluções serão invertidas, tal qual na figura 4.2.6. A vantagem em utilizar mapas de solução, como os apresentados, é a possibilidade de avaliar se o número de restrições impostas são suficientes para descrever o modelo.

Figura 4.2.4: Mapa de soluções ótimas do conjunto de dados **D1** quando FNNLS e fechamento são aplicados.



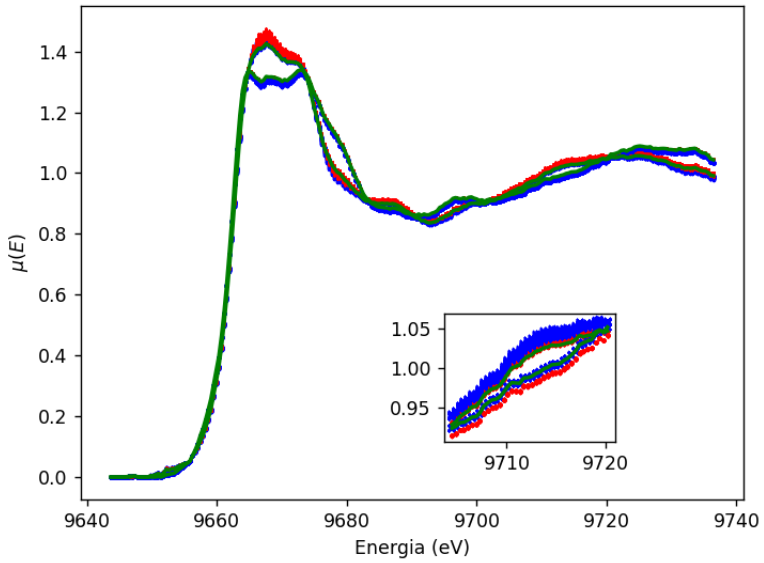
Os mapas de soluções em função de T2 e T4 auxiliam na seleção de resultados ótimas do sistema. Além disso, com a análise da evolução dos mapas é possível concluir se as restrições aplicadas são ou não suficientes para descrever o conjunto de dados $\mathbf{D_1}$. Apesar das soluções ótimas estarem concentradas em regiões específicas, tipicamente próximas a 0, esses mapas possuem uma simetria que possibilita a "troca de componentes", ou seja, ao selecionar os menores valores de *ssq* no mapa, tal como na figura 4.2.5, há uma simetria de valores que geram componentes iguais, que mudam apenas quanto à ordem.

Figura 4.2.5: Mapa de soluções ótimas do conjunto de dados **D1** quando FNNLS e fechamento são aplicados. As regiões em branco indicam as soluções ótimas do sistema, cujo ssq é 0.01% acima do mínimo.



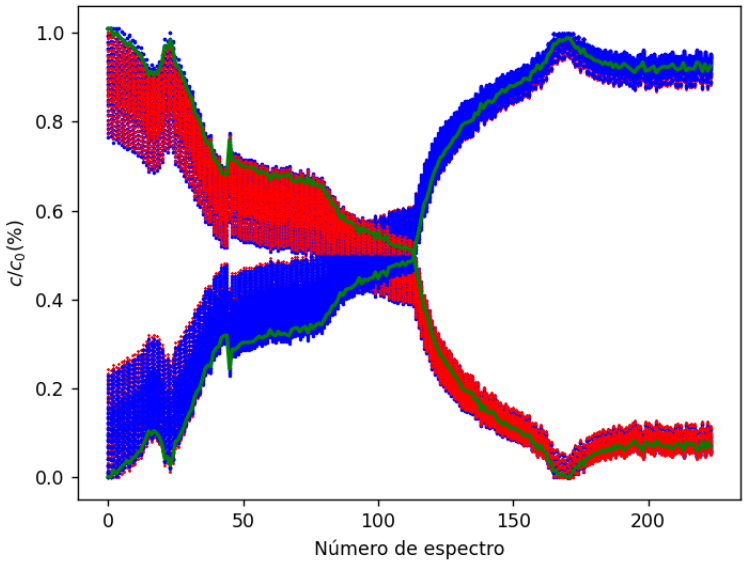
Como é possível notar na figura 4.2.6, há regiões onde as curvas em azul (primeira componente) e em vermelho (segunda componente) são sobrepostas, indicando que as soluções do sistema são simétricas, neste caso há trocas de componentes. A melhor solução do sistema (em verde), também apresenta simetria. Isso ocorre pois o número de restrições do sistema não foi suficiente para definir as concentrações iniciais. Na prática, não há uma diferença entre os resultados, apenas uma troca, o que era azul transforma-se vermelho e vice-versa.

Figura 4.2.6: Conjunto de soluções ótimas. Em azul, está primeira componente pura, em vermelho a segunda e em verde a solução cujo *ssq* é mínimo. Há sobreposição das curvas vermelha e azul.



Diferentemente dos espectros, os perfis de concentração apresentam um maior erro associado às ambiguidades. Contudo, não é possível concluir que isso é um efeito comum nos conjuntos de dados, pois a análise foi feita com apenas um experimento. Ainda assim, é possível notar que, do mesmo modo que os espectros, as concentrações apresentam o mesmo efeito de troca de componentes devido a simetria do espaço de soluções. Além disso, em razão do erro associado à ambiguidade, é difícil estimar parâmetros da reação por conta da incerteza da decomposição.

Figura 4.2.7: Conjunto de soluções ótimas. Em azul, está primeira componente pura, em vermelho a segunda e em verde a solução cujo *ssq* é mínimo. Há sobreposição das curvas vermelha e azul.



Para resolver o problema de trocas, é possível utilizar a condição de fechamento. Além de impor o balanço de massa, garante-se a não simetria do espaço de solução, fazendo com que as soluções deixem de ser perfeitamente simétricas. Como resultado, as componentes vermelhas e azuis se distinguem e não são sobrepostas, como explicitado na figura 4.2.8. Contudo, o erro associado às ambiguidades rotacionais nos espectros aparenta ser igual em ambos os casos, enquanto nas concentrações, a única diferença é a redução das trocas de componentes, assim como na figura 4.2.9. Uma possível explicação, é que os espectros estão normalizados, com isso, o balanço de massa, por ser uma restrição natural, não afetaria significativamente a forma com que o sistema é decomposto, apenas a ordem em que a reação química ocorre.

Figura 4.2.8: Conjunto de soluções ótimas. Em azul, está primeira componente pura, em vermelho a segunda e em verde a solução cujo *ssq* é mínimo. Há sobreposição das curvas vermelha e azul.

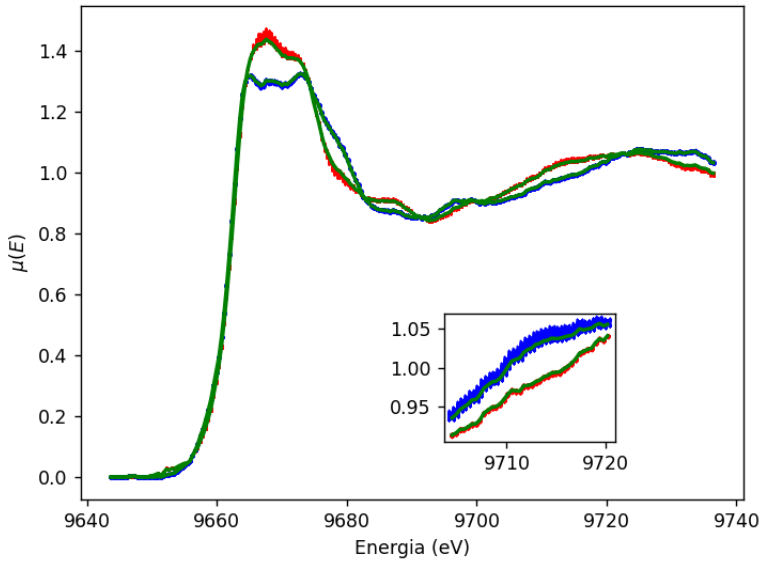
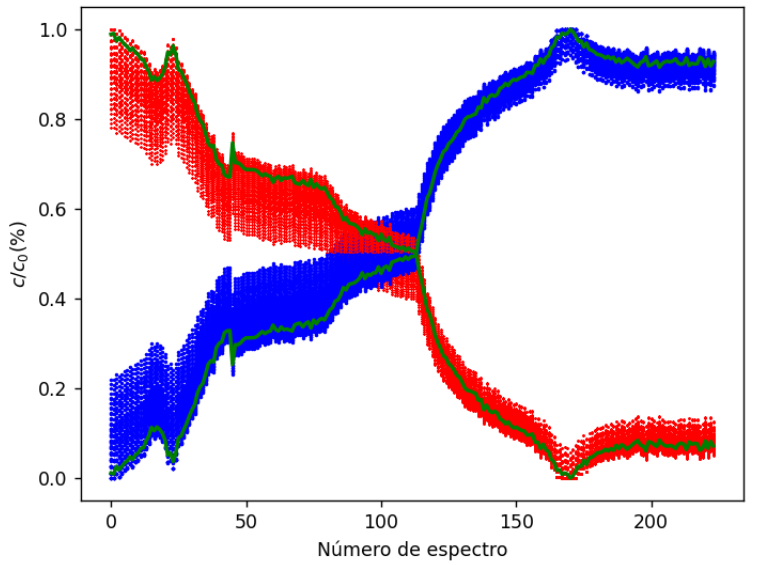


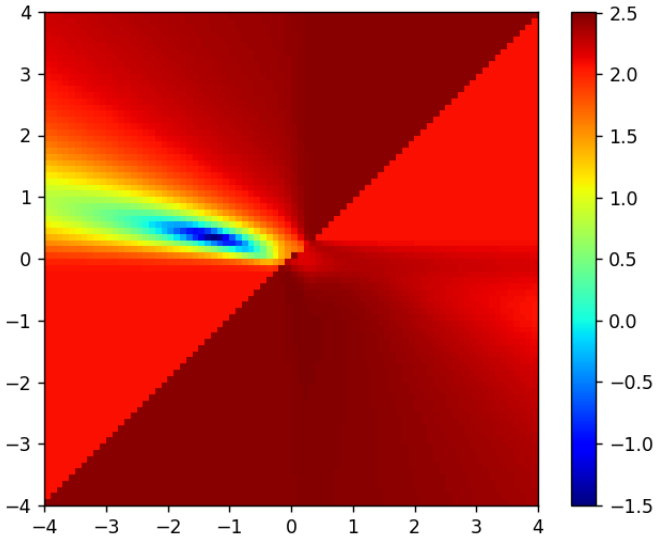
Figura 4.2.9: Conjunto de soluções de concentração ótimas em azul a primeira componente pura, em vermelho a segunda e em verde a solução cujo ssq é mínimo. Há sobreposição das curvas vermelhas e azuis.



4.2.3 Ambiguidades em sistemas mais restritos

No caso em que a otimização do MCR é mais limitada, o mapa de soluções pode restringir-se. Para o caso da condição de igualdade a simetria do problema pode ser resolvida, gerando um "poço" de soluções ótimas. Nesse tipo de análise, devido a restrição de concentração, não há como haver troca de componentes como na figura 4.2.6, pois soluções com simetria, nesse caso, encontram-se nas regiões em vermelho da figura 4.2.10 e, consequentemente, não são soluções ideais do ponto de vista estatístico.

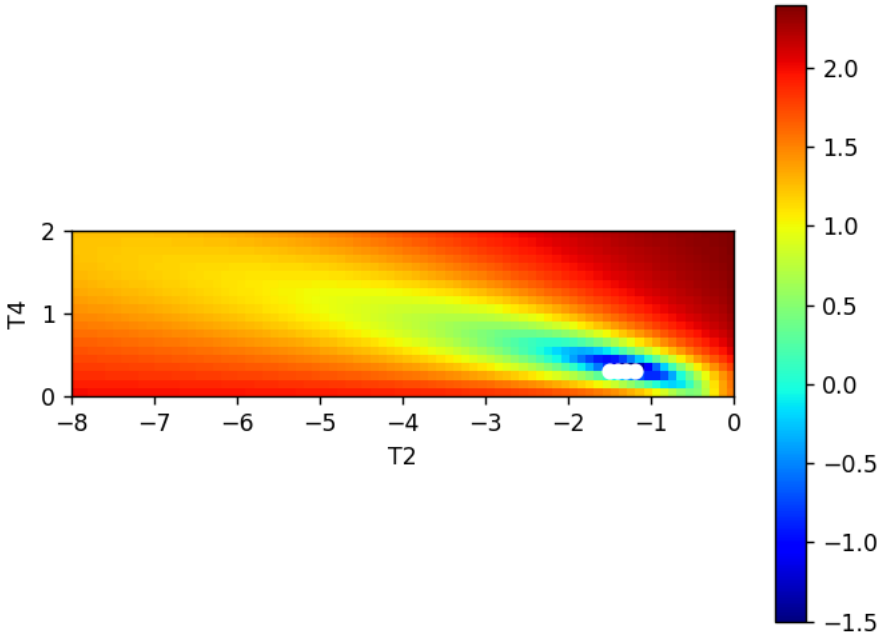
Figura 4.2.10: Mapa de soluções ótimas do conjunto de dados **D1** quando *FNNLS*, fechamento e *equality* são aplicados. Regiões em vermelho são soluções com alto *ssq* e em azul com baixo *ssq*.



Assim como na análise anterior, é possível selecionar a região cujas soluções são 30% maiores que o mínimo de *ssq* para avaliar as ambiguidades rotacionais (na análise anterior considera-se apenas 0.01%). Nesse caso, onde não há trocas de componentes, é possível utilizar as soluções provenientes de ambiguidade rotacional para estimar o erro associado a estas. Isto significa que é possível estimar qual erro da decomposição está associado à ambiguidade rotacional. Contudo, não foi possível definir um parâmetro de seleção de soluções mínimas.

Quando comparada com o modelo com menos restrições, a figura 4.2.10 apresenta uma região menor de soluções. Devido às limitações computacionais (e também temporais), a resolução dos mapas não foi aumentada. Contudo, é possível, ainda assim, aproximar o espaço soluções para um conjunto contínuo de pontos para calcular os espectros da figura 4.2.12.

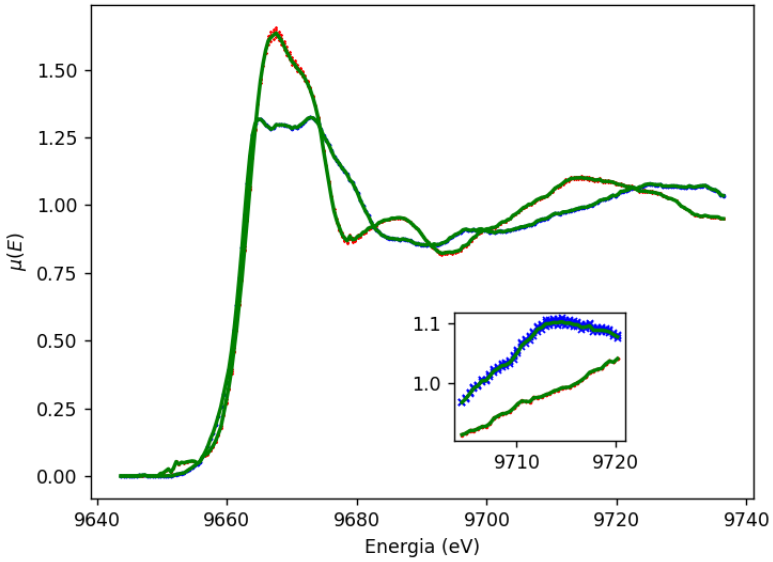
Figura 4.2.11: Mapa de soluções ótimas do conjunto de dados **D1** quando *FNNLS*, fechamento e igualdade são aplicados. As regiões em branco indicam as soluções ótimas do sistema, cujo *ssq* é 30% acima do mínimo.



Assim, diferentemente da figura 4.2.6, o caso com maiores restrições não apresenta sobreposição de espectros, apenas um *range* de soluções. Também é possível notar que a solução ótima (em verde) está aproximadamente no centro das soluções relacionadas à ambiguidade. Este fato indica que é possível estimar os erros associados com um algorítimo de maximização e minimização de espectros, assim como o proposto por Tauler et al.[26].

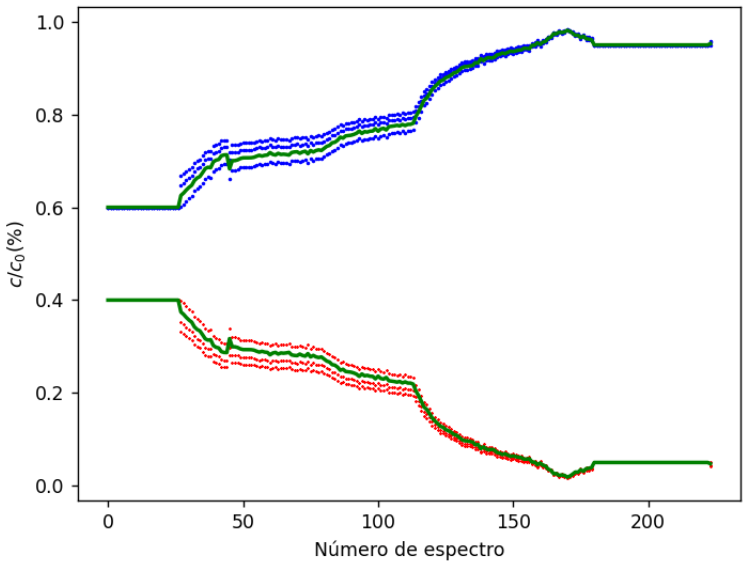
Caso seja possível definir uma área sob o “menor”e o “maior”espectro azul ou vermelho, isto é, um parâmetro de seleção de soluções ótimas, é viável calcular um erro associado à decomposição e, desse modo, garantir que a o sistema não possua um ruído na ordem da própria medida.

Figura 4.2.12: Conjunto de soluções ótimas. Em azul a primeira componente pura, em vermelho a segunda e em verde a solução cujo *ssq* é mínimo. Não há sobreposição.



Quando a condição de igualdade é utilizada, é possível notar que os erros associados às ARs são reduzidos, pois, mesmo selecionando as soluções 30% maiores que o mínimo, as ambiguidades dos espectros ficam bem definidas. Apesar das concentrações apresentarem uma maior ambiguidade, os erros são menores quando comparados com os resultados com menos restrições. Esse resultado indica que o sistema é bem resolvido quando está mais restrito.

Figura 4.2.13: Conjunto de soluções de concentrações ótimas. Em azul a primeira componente pura, em vermelho a segunda e em verde a solução cujo *ssq* é mínimo. Não há sobreposição.



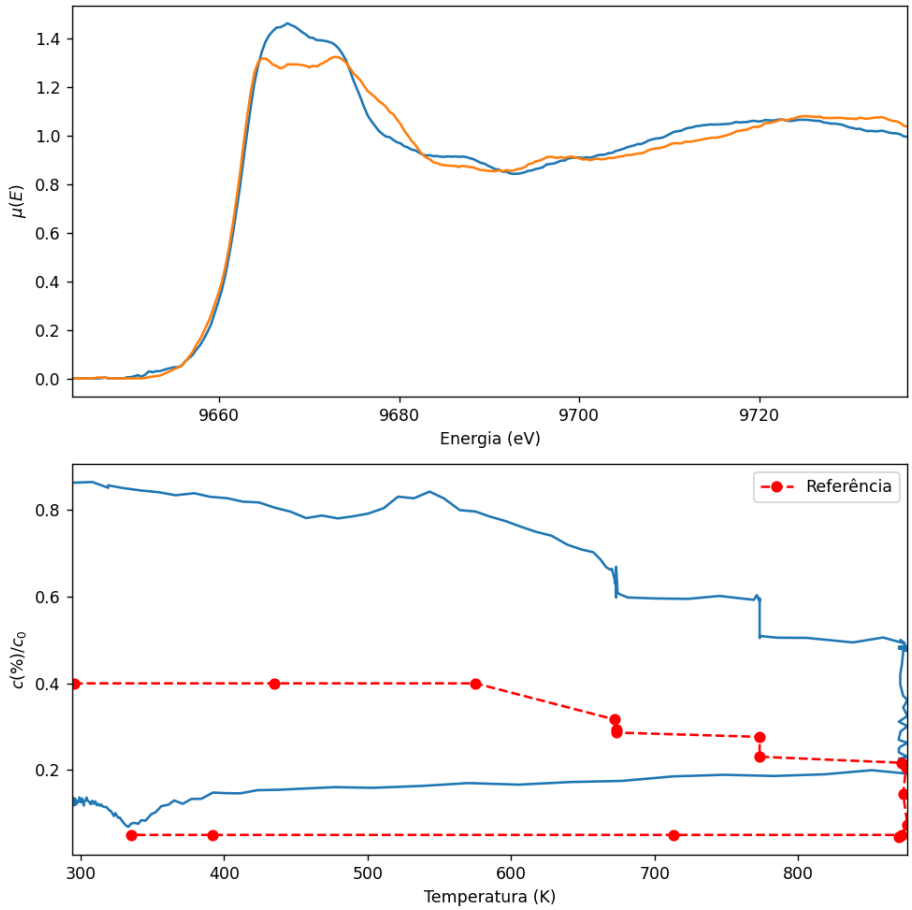
4.3 Resultados do MCR-ALS

Para analisar o conjunto de dados, foi utilizado o MCR-ALS do CORAL. Devido a implementação em Python, é possível utilizar os Jupyter Notebooks para montar um caderno experimental de análise de dados, de modo a facilitar a criação de uma análise completa de dados multivariados. Nesse caso, os dados utilizados foram da matriz **D1**, onde não haviam referências para aplicar o método de combinação linear. Por esta razão a análise foi realizada com técnicas multivariadas.

4.3.1 Restrições naturais

Na ocasião em que apenas restrições naturais são aplicadas, isto é, não-negatividade, unimodalidade e fechamento, o sistema tende a achar uma resposta cuja ambiguidade é vista como um alongamento. Isso ocorre pois os perfis de concentração são alongados (no eixo y) por um fator definido pela otimização, como já discutido previamente na aplicação de *FNNLS* e fechamento, cujo resultado resulta em modelos simétricos não restritos no espaço de soluções. Assim, resultados como o da figura 4.3.1 são comuns em modelos com poucas restrições para descrever a mistura. Desse modo, a solução é múltipla da referência obtida através de espectroscopia Mossbauer (em vermelho).

Figura 4.3.1: Resultados do MCR-ALS com FNNLS e fechamento.



4.3.2 Restrições de *hard-modeling*

Como visto no resultado da figura 4.3.1, a solução ótima do MCR pode possuir forma semelhante a referência, mas valores diferentes. Além disso, os mapas de soluções indicam a necessidade de introduzir informações prévias no modelo para obter respostas mais próximas às reais. Para tanto, é possível utilizar *hard-modeling*, ou seja, introduzindo mais considerações no sistema através de igualdade (*equality*) de posto local. Nesse tipo de restrição, as concentrações são limitadas a 40% até aproximadamente 580K. Esse tipo de restrição força a otimização do MCR-ALS a encontrar soluções alternando-se a concentração e o espectro, levendo aos perfis de concentração a serem próximos da referência, assim como na figura 4.3.2.

Figura 4.3.2: Soluções com imposição de igualdade no MCR-ALS.

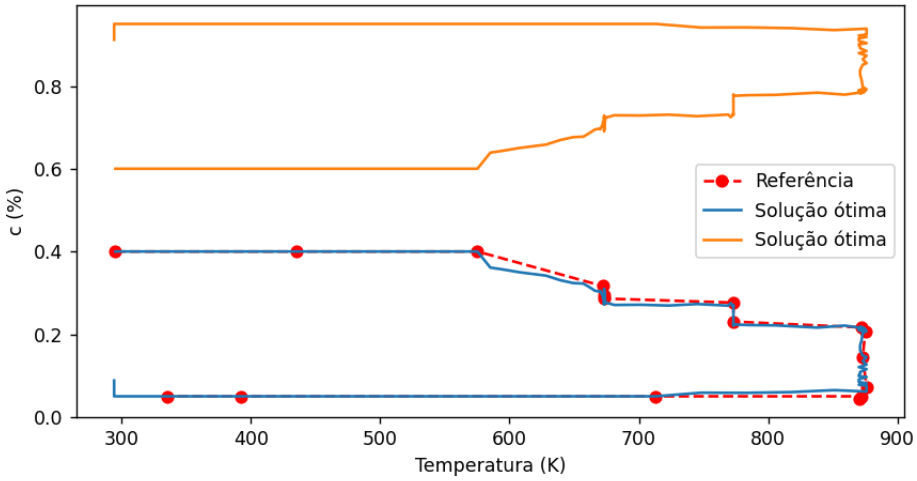
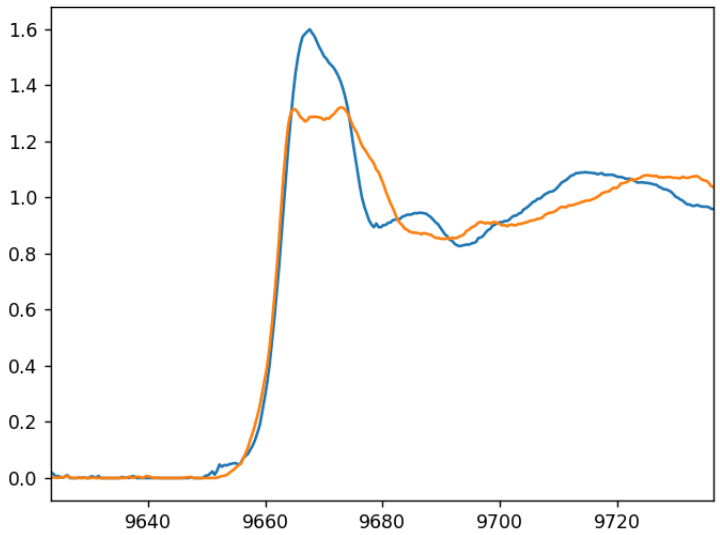


Figura 4.3.3: Espectros soluções com imposição de igualdade no MCR-ALS.



4.3.3 Evolução dos modelos

Ao comparar a evolução de ambos os modelos (*FNNLS*, Fechamento - modelo 1, e *FNNLS*, Fechamento, Igualdade - modelo 2) é possível notar uma clara diferença. Enquanto no modelo menos restrito há pouca ou quase nenhuma modificação da primeira solução (figura 4.3.4), no mais restrito (figura 4.3.5), tanto espectros quanto concentrações são forçados a se modificarem para descrever a mistura.

Figura 4.3.4: Modelo 1 com restrições naturais. As curvas em vermelho são resultados de concentrações e espectros, já as curvas intermediárias (passos da otimização) são as coloridas. Não há escala de energia nos espectros e as concentrações estão função do número de espectros.

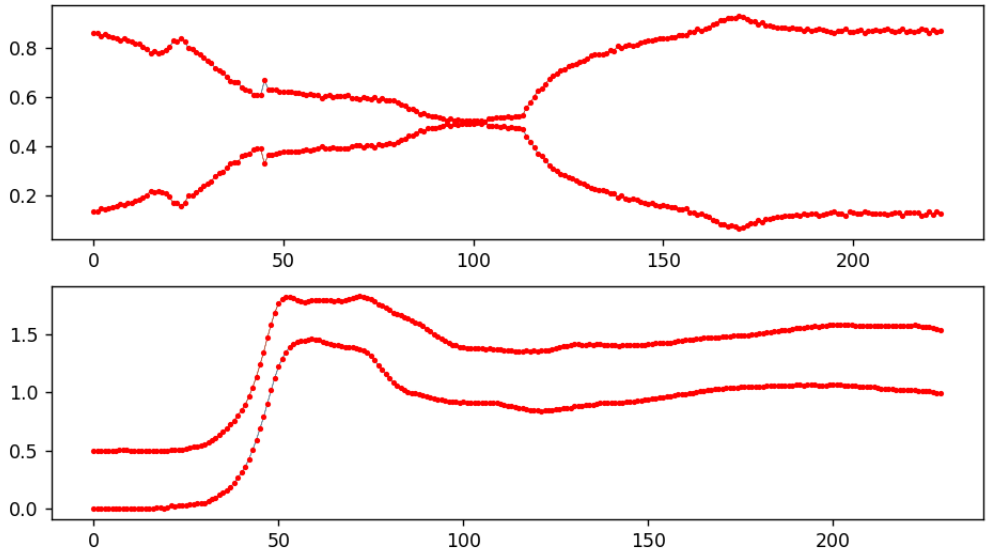
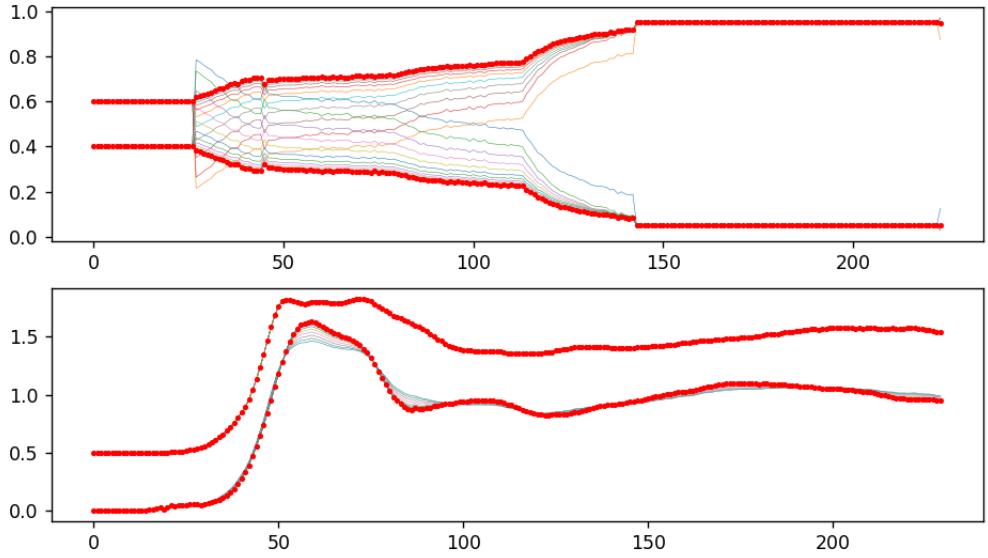


Figura 4.3.5: Modelo 2 com *hard-modeling*. As curvas em vermelho são os resultados de concentrações e espectros, as curvas intermediárias (passos da otimização) são as coloridas. Não há escala de energia nos espectros e as concentrações estão em função do número de espectros.



Apesar de apresentar diferentes otimizações, ambos os modelos possuem parâmetros estatísticos semelhantes. A variância explicada pelos dois modelos é igual (R^2), enquanto a falta de ajuste (*Lack of Fit*) é ligeiramente melhor para as restrições naturais. Vale ressaltar que tais diferenças são pequenas e na ordem dos erros da medida, na prática, isto quer dizer que ambos os modelos convergem igualmente (em termos estatísticos) para descrever o sistema, sendo este um caso claro de ambiguidade rotacional.

Tabela 4.3.1: Parâmetros estatísticos da decomposição com MCR-ALS.

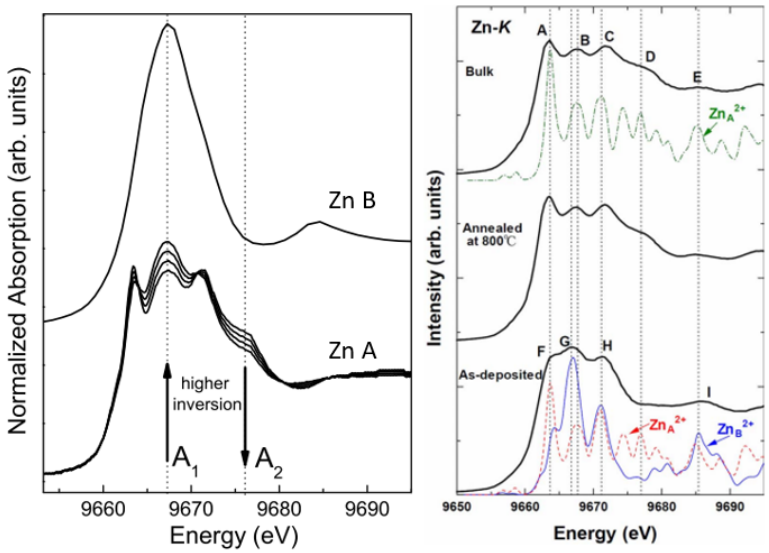
Modelo	R^2	Lack of Fit PCA (LOF) [%]	Lack of Fit Exp (LOF) [%]
Hard Modeling	99.992	0.238	0.903
Restrições Naturais	99.992	0.217	0.897

O resultado apresentado na figura 4.3.5 é semelhante ao obtido pelas matrizes de transformação (figura 4.2.12 e figura 4.2.13). Isso indica que estas soluções advindas das matrizes de transformação podem ser boas estimativas iniciais para o MCR-ALS, pois são soluções próximas de uma região com o *ssq* mínimo. Para tanto é possível utilizar um algorítimo para encontrar mínimos, pois podem reduzir o tempo de cálculo e apresentar apenas o melhor resultado.

4.3.4 Concordância com simulações

Além disso, também é possível comparar os espectros puros com referências teóricas, sendo o ideal é comparar posições de picos para validar o modelo experimental. Não é possível comparar os espectros apenas visualmente, pois a largura dos picos pode mudar por diferentes efeitos do cálculo teórico e confundir o experimentador. Nesse caso, é possível notar que a solução mais restrita apresenta picos sobrepostos na região logo após a linha branca. Ainda assim, é possível notar que os picos em 9667 eV e 9673 eV são os mesmos apresentados na figura 4.3.3. Contudo, em ambas as simulações apresentadas, na figura 4.3.6 os picos encontram-se com larguras diferentes do resultado experimental, indicando que os parâmetros utilizados no modelo teórico não foram os ideais para determinar um espectro semelhante ao encontrado pelo MCR-ALS.

Figura 4.3.6: Simulação computacional do Zinco no sítio B e os efeitos da inversão no espectro calculados por dois grupos de pesquisa independentes. Na figura a esquerda, para Zn B a largura de pico sobrepõe os picos menores próximos a 9673 eV, que podem se evidenciar na simulação teórica a direita em azul, calculados com outros parâmetros iniciais. Como se pode observar, os parâmetros teóricos utilizados afetam significativamente a largura dos picos. [5, 6]



Vale ressaltar, que diferentemente da simulação, o MCR-ALS não necessita de parâmetros da estrutura cristalina do composto estudado, além de valores que variam a largura dos picos do cálculo teórico. Assim, é possível ter resultados próximos aos das referências sem conhecer a estrutura que está sendo estudada em um tempo bem menor que o de cálculo de uma simulação computacional. Além disso, o CORAL é um programa de uso simples, possibilitando análises rápidas durante o uso da linha de luz, enquanto simulações podem possuir parâmetros que necessitam de outras técnicas complementares para ajustar o modelo. Cabe aqui a possibilidade de propor o uso das facilidades desenvolvidas neste trabalho para otimizar os parâmetros utilizados nas simulações teóricas.

Quando comparado com o método de matriz de transformação, o MCR-ALS também apresenta diversas vantagens, pois a regressão com mínimos quadrados precisa de pouca informação inicial, enquanto o método de matrizes de transformação necessita da visualização do mapa de soluções para o bom entendimento de quais soluções são estatisticamente ideais. Apesar disso,

o método de matriz de transformação apresenta vantagens quando aplicado em conjuntos de dados onde há baixa variância entre os espectros, isto é, cujas curvas são semelhantes.

5 Conclusões e perspectivas

Portanto, foi possível desenvolver um método para análise de ambiguidade de soluções para métodos multivariados. Devido à sua generalidade, este pode ser estendido para incluir mais componentes, desde que certas aproximações sejam feitas. Os mapas de soluções, quando usados em conjunto com o MCR-ALS, são ferramentas críticas para validar dados e garantir a unicidade das soluções.

Embora a implementação possa ser facilmente estendida para incluir um número maior de componentes, as consequências do aumento desse número em sistemas com misturas devem ser estudadas. Isso ocorre pois não há estudos que mostrem claramente como uma mudança no número de componentes puras afeta o espaço de soluções do sistema com dados de absorção de raios-x. Além disso, é necessário um estudo da inicialização do sistema, ou seja, de como as estimativas iniciais afetam o comportamento das soluções. Nesse aspecto, como resultado das várias metodologias utilizadas, é razoável esperar que as estimativas baseadas em referências experimentais produzam as melhores soluções.

Apesar de promissores, os mapas de soluções espaciais são grandes coleções de dados devido à dependência com a boa resolução. Mapas com baixa resolução podem produzir resultados que não são tão próximos do mínimo de *ssq*, em contraponto os com alta resolução podem levar dias para serem calculados completamente. Como resultado, é necessário estabelecer critérios para os cálculos de mapas sem deixar de lado pontos menores devido à baixa resolução escolhida.

Ainda assim, o método da matriz de transformação pode ser uma alternativa viável para as estimativas iniciais do MCR-ALS. Pois, diferentemente do SIMPLISMA, os resultados das matrizes de rotação não necessariamente são os espectros que estão contidos no conjunto de dados. Nesse sentido, a análise parte para estimativas que não apenas estão fora do conjunto de dados, mas seguem as mesmas restrições impostas à otimização do MCR-ALS. Este desenvolvimento poderá ser um tópico de investigação futura na área de métodos de inicialização para minimização com mínimos quadrados alternados.

Por fim, o CORAL ganha novas ferramentas de análise de dados, dando margens para o pesquisador examinar com maior profundidade as soluções para o sistema de interesse. Como resultado, é possível quantificar como as ambiguidades afetam o sistema. Além disso, se um critério for usado para selecionar áreas de solução, é possível utilizar os espectros como uma estimativa do erro associado às ambiguidades da decomposição do sistema. Ainda assim, as soluções obtidas por este método podem vir a ser a base para aprimorar os parâmetros das simulações computacionais das componentes puras.

Referências

[1] J. E. Penner-Hahn *et al.*, "X-ray absorption spectroscopy," *Comprehensive Coordination Chemistry II*, vol. 2, pp. 159–186, 2003.

[2] W. Windig and J. Guilment, "Interactive self-modeling mixture analysis," *Analytical chemistry*, vol. 63, no. 14, pp. 1425–1432, 1991.

[3] H. Keller and D. Massart, "Evolving factor analysis," *Chemometrics and intelligent laboratory systems*, vol. 12, no. 3, pp. 209–224, 1991.

[4] S. J. A. Figueroa, *Propiedades asociadas a la estructura local en sistemas nanométricos: Estudio mediante el empleo de técnicas basadas en el uso de luz de sincrotrón*. PhD thesis, Universidad Nacional de La Plata, 2009.

[5] S. J. Stewart, S. Figueroa, J. R. López, S. G. Marchetti, J. F. Bengoa, R. Prado, and F. G. Requejo, "Cationic exchange in nanosized zn fe 2 o 4 spinel revealed by experimental and simulated near-edge absorption structure," *Physical Review B*, vol. 75, no. 7, p. 073408, 2007.

[6] S. Nakashima, K. Fujita, K. Tanaka, K. Hirao, T. Yamamoto, and I. Tanaka, "First-principles xanes simulations of spinel zinc ferrite with a disordered cation distribution," *Physical Review B*, vol. 75, no. 17, p. 174443, 2007.

[7] E. R. Malinowski and D. G. Howery, *Factor analysis in chemistry*, vol. 3. Wiley New York, 1980.

[8] A. Rochet, B. Baubet, V. Moizan, C. Pichon, and V. Briois, "Co-k and mo-k edges quick-xas study of the sulphidation properties of mo/al2o3 and como/al2o3 catalysts," *Comptes Rendus Chimie*, vol. 19, no. 10, pp. 1337–1351, 2016.

[9] A. Martini and E. Borfecchia, "Spectral decomposition of x-ray absorption spectroscopy datasets: Methods and applications," *Crystals*, vol. 10, no. 8, p. 664, 2020.

[10] H. Bornebusch, B. Clausen, G. Steffensen, D. Lutzenkirchen-Hecht, and R. Frahm, "A new approach for qexafs data acquisition," *Journal of synchrotron radiation*, vol. 6, no. 3, pp. 209–211, 1999.

[11] S. J. Figueroa, D. C. d. Oliveira, A. Rochet, J. C. Mauricio, C. Doro Neto, A. P. S. Levinsky, and H. Westfahl Junior, "Quati: time-resolved xas beamline at sirius," 2019.

[12] M. Newville, "Fundamentals of xafs," *Reviews in Mineralogy and Geochemistry*, vol. 78, no. 1, pp. 33–74, 2014.

[13] S. Calvin, *XAFS for Everyone*. CRC press, 2013.

[14] J. Timoshenko and A. I. Frenkel, ""inverting" x-ray absorption spectra of catalysts by machine learning in search for activity descriptors," *Acs Catalysis*, vol. 9, no. 11, pp. 10192–10211, 2019.

[15] H. F. Kaiser, "Computer program for varimax rotation in factor analysis," *Educational and psychological measurement*, vol. 19, no. 3, pp. 413–420, 1959.

[16] J. Jaumot, A. de Juan, and R. Tauler, "Mcr-als gui 2.0: New features and applications," *Chemometrics and Intelligent Laboratory Systems*, vol. 140, pp. 1–12, 2015.

[17] W. Windig, N. B. Gallagher, J. M. Shaver, and B. M. Wise, "A new approach for interactive self-modeling mixture analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 77, no. 1-2, pp. 85–96, 2005.

[18] P. H. Março, P. Valderrama, G. L. Alexandrino, R. J. Poppi, and R. Tauler, "Resolução multivariada de curvas com mínimos quadrados alternantes: descrição, funcionamento e aplicações," *Química Nova*, vol. 37, no. 9, pp. 1525–1532, 2014.

[19] A. C. Olivieri, "A down-to-earth analyst view of rotational ambiguity in second-order calibration with multivariate curve resolution- a tutorial," *Analytica Chimica Acta*, vol. 1156, p. 338206, 2021.

[20] M. Vosough, C. Mason, R. Tauler, M. Jalali-Heravi, and M. Maeder, "On rotational ambiguity in model-free analyses of multivariate data," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 20, no. 6-7, pp. 302–310, 2006.

[21] A. Golshan, H. Abdollahi, and M. Maeder, "The reduction of rotational ambiguity in soft-modeling by introducing hard models," *Analytica chimica acta*, vol. 709, pp. 32–40, 2012.

[22] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. B. Hamrick, J. Grout, S. Corlay, *et al.*, *Jupyter Notebooks-a publishing format for reproducible computational workflows.*, vol. 2016. 2016.

[23] S. Figueroa and C. Prestipino, "Prestopronto: a code devoted to handling large data sets," in *Journal of Physics: Conference Series*, vol. 712, p. 012012, IOP Publishing, 2016.

[24] M. Newville, "Larch: an analysis package for xafs and related spectroscopies," in *Journal of Physics: Conference Series*, vol. 430, p. 012007, IOP Publishing, 2013.

[25] S. M. Wallace, M. A. Alsina, and J.-F. Gaillard, "An algorithm for the automatic deglitching of x-ray absorption spectroscopy data," *Journal of Synchrotron Radiation*, vol. 28, no. 4, 2021.

[26] A. C. Olivieri and R. Tauler, "N-bands: A new algorithm for estimating the extension of feasible bands in multivariate curve resolution of multicomponent systems in the presence of noise and rotational ambiguity," *Journal of Chemometrics*, vol. 35, no. 3, p. e3317, 2021.