Universidade de Aveiro Departamento de Matemática,
2021

Francisco João
Correia Silva

Co-Tucker: aplicação do modelo estatístico em espectros metabolíticos

Universidade de Aveiro Departamento de Matemática,
2021

Francisco João Correia Silva

Co-Tucker: aplicação do modelo estatístico em espectros metabolíticos

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, realizada sob orientação científica de Adelaide de Fátima Baptista Valente Freitas, Professora Associada do Departamento de Matemática da Universidade de Aveiro, e co-orientação da Professora Associada com Agregação Ana Maria Pissarra Coelho Gil.

O júri / The jury

Presidente / President

Prof. Doutor Eugénio Alexandre Miguel Rocha
Professor Associado da Universidade de Aveiro

Vogais / Committee

Prof. Doutora Susana Luísa da Custódia Machado Mendes
Professora Adjunta da Escola Superior de Turismo e Tecnologia do Mar - Politécnico de Leiria

Prof. Doutora Adelaide de Fátima Baptista Valente Freitas
Professora Associada da Universidade de Aveiro (orientadora)

Agradecimentos / Acknowledgements

Em primeiro lugar, quero agradecer à minha família, principalmente à mãe, ao pai e irmão, por me proporcionarem este trajeto académico de 5 anos e pelo constante apoio e incentivo em cumprir os objetivos traçados. Sem eles não tinha sido possível.

Em segundo lugar, agradecer à professora Adelaide Freitas por esta oportunidade, e, ao mesmo tempo, pela disponibilidade, atenção e apoio que demonstrou durante todo este último ano. Agradecer também à Daniela Duarte e à professora e coorientadora Ana Gil pela ajuda e disponibilidade na concretização desta dissertação.

Não é fácil estudar longe do conforto de casa e dos amigos. Um agradecimento a eles, pela força, ajuda e paciência que sempre tiveram comigo. Mesmo nas muitas vezes que estive ausente, sempre estiveram disponíveis quando foi preciso.

Agradecer também aos colegas de curso que sempre me acompanharam neste percurso. Sem dúvida que sem eles não seria possível chegar a esta etapa final. Ao Bernardo, Diogo, João e à minha patroa Inês, um muito obrigado.

Sem mencionar nomes porque poderia deixar alguém ofendido, não podia deixar de agradecer aos meus colegas de casa e de quarto. Sem dúvida uma 2º família que sempre apoiaram, mesmo nos piores momentos. É fundamental um lar saudável quando estamos a mais de 100km de casa. Marcaram para sempre o meu percurso académico e por isso um gigante obrigado a todos.

Por último, queria deixar um agradecimento muito especial a uma pessoa muito importante para mim e que me deixou durante o percurso universitário. A minha avó. Um enorme obrigado por tudo.

Palavras-chave Tucker3, análise da co-inércia, biplot, modelo Co-Tucker, ortogonalidade, análise de componentes principais, dados tridimensionais, modo

Resumo O modelo Co-Tucker é uma metodologia estatística, recentemente desenvolvida, que aplica a decomposição de Tucker a matrizes do covariâncias resultantes de pares de dados tridimensionais (descritos em cubos), onde os objetos são medidos por diferentes variáveis ao longo de eventos/tempos/espaços comuns para os dois cubos de dados. Esta dissertação apresenta, na componente teórica, o desenvolvimento da metodologia Co-Tucker complementada com a implementação de um código construído em linguagem de programação R para a sua aplicação. São analisados dois cubos emparelhados de dados relativos a medidas obtidas por espectroscopia de ressonância magnética nuclear (NMR) de dois bioflúidos (saliva e urina), de 7 mulheres grávidas, observadas em três instantes de tempo. Os cubos analisados são, um relativo à espectroscopia salivar de quatro metabolitos, e o outro cubo relacionado com a espetroscopia urinária de vinte e quatro metabolitos. Explorar interações entre os metabolitos salivares e urinários e nos três trimestres de gestação é a contribuição aplicada desta dissertação. Este processo decorre através da realização de dois modelos Co-Tucker, construídos de forma diferente, com o âmbito de abordar a exploração dos dados de duas formas diferentes: uma por trimestre e outra por diferença entre trimestres.

Keywords Tucker3, co-inertia analysis, biplot, Co-Tucker model, orthogonality, principal component analysis, three-way data analysis, mode.

Abstract The Co-Tucker Model is a recently developed statistical methodology that applies the Tucker decomposition to covariance matrices resulting from the three-dimensional data pairs (described in cubes), where objects are measured by different variables over common events/times/spaces for the two data cubes. This dissertation presents, in the theoretical component, the development of the Co-Tucker methodology complemented with the implementation of a code built in R programming language. Two paired of data cubes, relating to measurements obtained by nuclear magnetic resonance (NMR) spectroscopy are analyzed in two biofluids (saliva and urine) from 7 pregnant women, observed in three instants of time. The cubes analyzed are one related to salivary spectroscopy of four metabolites, and the other cube related to urinary spectroscopy of twenty-four metabolites. Explore interactions between salivary and urinary metabolites in the three trimesters of pregnancy is the applied contribution of this dissertation. This process takes place through the realization of two Co-Tucker models, built differently, with the scope of approaching data exploration in two different ways: one by trimester and another by difference between trimesters.

Conteúdo

1 Introdução **1**

2 Metodologia: Co-Tucker e os seus principais conceitos **3**
2.1 Introdução . . . 3
2.2 Principais conceitos . . . 3
2.2.1 Análise de co-inércia . . . 3
2.2.2 Modelo Tucker3 . . . 4
2.3 Modelo Co-Tucker . . . 6
2.3.1 Construção do modelo Co-Tucker . . . 6
2.3.2 Construção das matrizes $\mathbf{A}^X$, $\mathbf{B}^Y$ e $\mathbf{C}$ . . . 10
2.4 Escolha do modelo . . . 12
2.5 Interpretação de resultados . . . 13
2.5.1 Interpretação numérica . . . 13
2.5.2 Interpretação gráfica . . . 15

3 Aplicação de modelos Co-Tucker **21**
3.1 Introdução . . . 21
3.2 Identificação de metabolitos . . . 21
3.3 Modelo Co-Tucker . . . 22
3.3.1 Tratamento de dados . . . 22
3.3.2 Interações trimestrais com a análise da co-inércia . . . 24
3.3.3 Escolha do modelo . . . 25
3.3.4 Construção do modelo e as suas caraterísticas . . . 26
3.3.5 Interpretações . . . 28
3.3.6 Análise de correlações . . . 31
3.3.7 Conclusões . . . 34
3.4 Modelo Co-Tucker diferencial . . . 35
3.4.1 Tratamento de dados . . . 35
3.4.2 Interações entre trimestres com a análise da co-inércia . . . 36
3.4.3 Escolha do modelo diferencial . . . 37
3.4.4 Construção do modelo diferencial . . . 38
3.4.5 Interpretações do modelo diferencial . . . 40
3.4.6 Análise de correlações no método diferencial . . . 41
3.4.7 Conclusões do modelo diferencial . . . 44

4 Considerações finais **45**

Lista de Tabelas

2.1 Matrizes exemplo $\mathbf{A}^X$, $\mathbf{B}^Y$ e $\mathbf{C}$ do 1º, 2º e 3º modo, respetivamente. 14
2.2 Cubo exemplo $\mathbb{G}$. 15

3.1 Identificação das variáveis do cubo $\mathbb{X}$. 21
3.2 Identificação das variáveis do cubo $\mathbb{X}$. 22
3.3 Critério DifFit. 25
3.4 Pesos das entradas das componentes do primeiro, segundo e terceiro modos. . 26
3.5 Cubo $\mathbb{G}$ 27
3.6 Proporção da variância explicada por cada uma das componentes construídas. 28
3.7 Correlação de Spearman de NAG com os metabolitos urinários 33
3.8 Correlação de Spearman de ureia com os metabolitos urinários com um valor-p$< \alpha = 0.05$ 34
3.9 Critério DifFit aplicado ao modelo Co-Tucker diferencial. 37
3.10 Pesos das entradas das componentes do primeiro, segundo e terceiro modo do modelo Co-Tucker diferencial. 38
3.11 Cubo $\mathbb{G}^\star$ do modelo diferencial. 39
3.12 Proporção da variância explicada por cada uma das componentes construídas no modelo diferencial. 40
3.13 Correlação de Spearman da acetoína, ureia e NAG com os metabolitos urinários. 43

Lista de Figuras

2.1 Esquema do modelo Tucker3. Decomposição do cubo $\mathbb{X}$ com as matrizes $\mathbf{A}$ $(I \times P)$, $\mathbf{B}$ $(J \times Q)$ e $\mathbf{C}$ $(K \times R)$ e com o cubo $\mathbb{G}$. 5

2.2 Passo 1 do Co-Tucker: realização de K tabelas cruzadas a partir dos dados longitudinais $\mathbb{X}$ e $\mathbb{Y}$ através de K análises co-inércia. 7

2.3 Passo 2 do Co-Tucker: decomposição do cubo $\mathbb{W}$ via ACP. 8

2.4 Interação positiva entre variáveis 17

2.5 Interação negativa entre variáveis 17

2.6 Interação nula entre variáveis 18

2.7 Biplot com a projeção das componentes do primeiro e segundo modo na primeira componente do terceiro modo. 18

3.1 Variância dos metabolitos salivares, a vermelho, e dos metabolitos urinários, a azul, no primeiro (a), segundo (b) e terceiro (c) trimestres. 23

3.2 Co-inérica medida no 1º, 2º e 3º trimestres a partir de dados $\mathbb{X}$ centrados e $\mathbb{Y}$ normalizados. 24

3.3 Co-inérica medida no 1º, 2º e 3º trimestres a partir de dados $\mathbb{X}$ e $\mathbb{Y}$ centrados (a), $\mathbb{X}$ e $\mathbb{Y}$ normalizados (b), e $\mathbb{X}$ normalizado e $\mathbb{Y}$ centrado (c). 24

3.4 Coordenadas do primeiro modo (a), segundo modo (b) e terceiro modo (c). . 27

3.5 Projeção das componentes dos modos da saliva e urina em R_1 (a) e em R_2 (b). 28

3.6 Projeção na primeira componente do terceiro modo R_1, com dados $\mathbb{X}$ e $\mathbb{Y}$ centrados (a), $\mathbb{X}$ e $\mathbb{Y}$ normalizados (b), e $\mathbb{X}$ normalizado e $\mathbb{Y}$ centrado (c). . . 31

3.7 Espectro do NMR de NAG, a preto, e U2, a azul, em cada grávida e ao longo dos 3 trimestres 32

3.8 Espectro do NMR de NAG, a preto, e GAA, a azul, em cada grávida no primeiro e terceiro trimestres. 32

3.9 Espectro do NMR de ureia, a preto, e X2.KG, a vermelho, em cada grávida em cada um dos trimestres. 33

3.10 Espectro do NMR de ureia, a preto, e U2, a vermelho, em cada grávida no primeiro e terceiro trimestres. 34

3.11 Variância dos metabolitos salivares, a vermelho, e dos metabolitos urinários, a azul, na primeira (a), segunda (b) e terceira (c) condição. 36

3.12 Co-inérica medida na 1º, 2º e 3º condição a partir de dados $\mathbb{X}^\star$ e $\mathbb{Y}^\star$ estandardizados. 36

3.13 Co-inérica medida na 1º, 2º e 3º condições a partir de dados $\mathbb{X}^\star$ e $\mathbb{Y}^\star$ centrados (a), $\mathbb{X}^\star$ normalizado e $\mathbb{Y}^\star$ centrado (b), e $\mathbb{X}^\star$ centrado e $\mathbb{Y}^\star$ normalizado (c). . 37

3.14 Coordenadas do primeiro (a) e segundo (b) modo. 39

3.15 Projeção das componentes do primeiro e segundo modo em R_1 do modelo diferencial. 40

3.16 Espectro do NMR de acetoína, a preto, e X4.DTA, a azul, em cada uma das condições. 42

3.17 Espectro do NMR de ureia, a preto, e X4.DTA, a azul, em cada uma das condições. 42

3.18 Espectro do NMR de NAG, a preto, e N5AC, a azul, em cada uma das condições. 43

Capítulo 1

Introdução

A gravidez é um período dinâmico no qual a mulher grávida passa por adaptações metabólicas indispensáveis para garantir o crescimento e o desenvolvimento fetal adequado [9]. A metabolómica é definida como a análise quantitativa e qualitativa de todos os metabolitos presentes num sistema biológico [11]. As técnicas analíticas mais usadas são a espetrometria de massa e a espectroscopia de ressonância magnética nuclear (RMN), complementares entre si [12]. Neste projeto, a espetroscopia de RMN foi usada para analisar a urina e a saliva de uma coorte de sete gestantes, durante a gravidez e ao longo de três trimestres. Devido à quantidade de dados gerados por RMN, a análise multivariada é de elevada importância [13]. Os espetros resultantes foram manuseados de forma a diferenciar cada trimestre da gravidez (do primeiro para o terceiro) através da análise multivariada. Concomitantemente, a análise univariada desses dados permitiu identificar quatro metabolitos urinários e vinte e quatro metabolitos salivares, responsáveis pela progressão da gravidez. Foram assim obtidos dois conjuntos de dados metabólicos longitudinais, onde temos a avaliação de espectros sobre os mesmos objetos e durante os três instantes de tempo.

Ao longo dos anos, várias metodologias estatísticas têm sido desenvolvidas com o objetivo de analisar tabelas tridimensionais, denominadas por "cubos", que avaliam simultaneamente três dimensões: objetos, variáveis e eventos. Estas metodologias ganham cada vez mais relevo nas áreas científicas mais aplicadas, com destaque para o estudo da interação entre fatores no decorrer de vários espaços temporais. Um modelo destinado a interpretar dados com estas caraterísticas é o modelo Co-Tucker. Neste sentido, e numa vertente exploratória, a realização desta dissertação tem como objetivo implementar o modelo estatístico Co-Tucker sobre estes dados. Tendo em conta que os espectros de metabolitos são medidos nas mesmas sete grávidas, no decorrer de três trimestres, consideremos dois cubos de dados a estudar: o primeiro, envolvendo o espectro de saliva que contém quatro metabolitos, denominado por cubo $\mathbb{X}$ e de dimensão $(7 \times 4 \times 3)$. O segundo, relacionado com o espectro da urina e que contém vinte e quatro metabolitos, denominado por cubo $\mathbb{Y}$ e de dimensão $(7 \times 24 \times 3)$. A metodologia Co-Tucker foi implementada em linguagem de programação de código em R, e encontra-se disponível no github em `https://github.com/Francisjcs1997/CT`.

No Capítulo 2 temos a metodologia descrita com uma análise detalhada da implementação do modelo Co-Tucker. Inicialmente, são introduzidos os conceitos básicos sendo, os mais importantes, a análise da co-inércia e a noção de decomposição tridimensional proposta por Tucker. De seguida, é mostrada a implementação de um modelo Co-Tucker e as noções numéricas que esta envolve. A escolha do melhor modelo também é mencionada com a descrição do método DifFit. O capítulo termina fornecendo as diferentes interpretações que

o modelo pode oferecer. Todo este material é exposto numa vertente teórica, com exemplos e observações, pelo que os dados correspondentes aos espectros de urina e saliva que iremos estudar, apenas vão ser trabalhados no Capítulo 3.

Com a componente teórica estudada, o próximo passo é a aplicação da mesma sobre os dados em análise. No Capítulo 3, todos as etapas da metodologia vão ser aplicadas sobre os conjuntos de dados acima referidos, no sentido de extrair o máximo de interpretações possíveis. De interesse prático, e recorrendo ao modelo Co-Tucker, seria bastante enriquecedor determinar marcadores com um possível valor diagonóstico. Para isso, é crucial responder a algumas questões que esta dissertação terá como objetivo clarificar: quais são os trimestres que revelam um maior número de interações de metabolitos? Em que passagem de tempo acontecem mais interações, do 1º para o 2º trimestre ou do 2º para o 3º? Que metabolitos presentes na saliva e urina apresentam uma interação positiva? Que metabolitos apresentam interações negativas? Quais as diferenças entre os espaços temporais em estudo? De forma a responder a estas questões, temos como motivação a criação de dois modelos Co-Tucker. O primeiro com o objetivo de analisar as interações em cada um dos trimestres. O segundo modelo com o objetivo de analisar as interações dos metabolitos entre os trimestres em estudo. Mais concretamente, este último modelo pretende fornecer informação do que acontece na passagem de tempo do 1º para o 2º trimestre, do 2º para o 3º trimestre e do 1º para o 3º trimestre.

Com a componente aplicada desenvolvida nesta dissertação, foi construído um resumo alargado (extended abstract) com o título "Co-Tucker method for analysis of urine and saliva interactions during pregnancy", o qual foi submetido para ser apresentado sob a forma de um poster no encontro "*3rd Statistics on health decision making: Public Health*", a realizar no dia 22/julho na Universidade de Aveiro. O trabalho foi sujeito a avaliação científica (cega).

Capítulo 2

Metodologia: Co-Tucker e os seus principais conceitos

2.1 Introdução

Para compreender e aplicar o modelo Co-Tucker num determinado conjunto de dados, é importante assinalar alguns pontos que são essenciais à construção do mesmo. Neste sentido, este capítulo serve para mostrar aos leitores todos os processos que permitem obter um modelo confiável e que nos permita extrair conclusões que de outra forma seria difícil ou até impossível.

Das noções mais básicas como conceitos de ortogonalidade, produto de Kronecker ou co-inércia, passando pela construção do modelo Co-Tucker e os seus principais componentes, e terminando na interpretação de resultados, esta secção vai revelar, de uma forma geral, como este método exploratório funciona. Todos os temas aqui tratados são explicados de uma forma genérica e com notações também genéricas. Foram também adicionados alguns exemplos práticos sempre com o objetivo de possibilitar, ao leitor desta dissertação, uma maior facilidade de compreensão.

2.2 Principais conceitos

Nesta secção, e num sentido mais detalhado, vamos estudar a análise da co-inércia com a construção de matrizes cruzadas entre dois conjuntos de dados. De seguida, é mencionada a construção do modelo proposto por Tucker que envolve uma redução de dimensionalidade.

2.2.1 Análise de co-inércia

A análise da co-inércia [4][7] é essencial à construção do modelo Co-Tucker. Esta é responsável pela associação das variáveis dos conjuntos de dados em estudo, medindo, as mesmo tempo, a co-estrutura e/ou concordância entre eles. Para uma fácil compreensão deste assunto, é apresentada de seguida uma descrição matemática da análise da co-inércia.

Consideremos $\mathbf{X}$ a primeira matriz de dados, com N linhas (objetos) e I colunas (variáveis) e $\mathbf{Y}$ a segunda, com os mesmos N objetos e J colunas (segundo grupo de varíaveis). Estas matrizes estão ligadas pelo mesmo número de objetos. Sejam ainda $\mathbf{X}^T$ e $\mathbf{Y}^T$ as matrizes transportas de $\mathbf{X}$ e $\mathbf{Y}$, respetivamente, $\mathbf{D}_N$ =Diag$(\omega_1, ..., \omega_N)$ uma matriz diagonal $(N \times N)$ de pesos ω_i dos objetos, e $\mathbf{D}_I$ e $\mathbf{D}_J$ matrizes diagonais $I \times I$ e $J \times J$ do hiperespaço da primeira e segunda matriz, respetivamente. Neste processo, uma análise em componentes

principais (ACP) generalizada pode ser realizada em cada matriz de dados através de uma decomposição espetral das matrizes $\mathbf{X}^T\mathbf{D}_N\mathbf{X}D_I$ e $\mathbf{Y}^T\mathbf{D}_N\mathbf{Y}\mathbf{D}_J$. Existem muitas formas de realizar esta ACP, contudo, a ACP usual em matrizes é caraterizada por uma matriz $\mathbf{D}_N$ de pesos uniformes, isto é, $\omega_1 = ... = \omega_N = 1/N$, e $\mathbf{D}_I$ e $\mathbf{D}_J$ matrizes identidade. Mais ainda, no caso dos dados centrados as matrizes anteriormente referidas serão proporcionais às correspondentes matrizes de covariância, enquanto que com dados normalizados coinciderão com as matrizes de correlação.

Todo este procedimento tem como objetivo maximizar a covariância entre os pesos das linhas das duas tabelas de dados [4]. Nesta análise é essencial a construção das tabelas cruzadas que são obtidas através da expressão

$$\mathbf{X}^*_{I\times J} = \mathbf{X}^T\mathbf{D}_N\mathbf{Y} \tag{2.1}$$

O número de variáveis tanto em $\mathbf{X}$ como em $\mathbf{Y}$ vai definir a dimensão da matriz $\mathbf{X}^*$. Para o caso de termos colunas centradas em ambas as tabelas, o total da inercia de cada uma delas corresponde simplesmente à soma das variâncias, ou seja, $\text{Iner}_{\mathbf{X}} = \text{traço}(\mathbf{X}\mathbf{D}_I\mathbf{X}^T\mathbf{D}_N)$ e $\text{Iner}_{\mathbf{Y}} = \text{traço}(\mathbf{Y}\mathbf{D}_J\mathbf{Y}^T\mathbf{D}_N)$ e a co-inercia entre $\mathbf{X}$ e $\mathbf{Y}$ corresponde à soma do quadrado das covariâncias, isto é, $\text{CoIner}_{\mathbf{XY}} = \text{traço}(\mathbf{X}\mathbf{D}_I\mathbf{X}^T\mathbf{D}_N Y\mathbf{D}_J\mathbf{Y}^T\mathbf{D}_N)$.

Em suma, o significado da co-estrutura entre duas matrizes pode ser definida do seguinte modo: a co-inércia é alta quando os valores de ambas as matrizes são simultaneamente altos ou quando variam inversamente. Por outro lado, a co-inércia é baixa quando os valores variam independentemente ou quando simplesmente não variam [7].

2.2.2 Modelo Tucker3

Modelo apresentado por Tucker [1], que pode ser encarado como uma redução tridimensional de um conjunto de dados. O método associado ao modelo de Tucker3 é simplesmente uma solução algébrica que, caso exista uma solução exata, esta metodologia produzi-la-á [10].

Definição 2.2.1 *Seja $\mathbb{X}$ um cubo, isto é, uma matriz tridimensional de dados. Definimos cada dimensão do cubo $\mathbb{X}$ como entidade ou modo. Vamos ter o primeiro modo associado aos objetos (linhas), segundo modo associado às variáveis (colunas) e o terceiro modo associado às condições (espaços temporais).*

O grande objetivo do modelo passa por transformar as três dimensões (objetos, variáveis e condições) do cubo, de tal forma que, as principais informações nelas contidas sejam expressas num número limitado de componentes para cada uma delas. Desta forma, o modelo Tucker3 é uma decomposição de dados tridimensionais $\mathbb{X} = (x_{ijk})$, onde assumimos que o primeiro modo está associado a uma matriz de componentes $\mathbf{A}$, o segundo modo, associado a uma matriz de componentes $\mathbf{B}$, e o terceiro e último modo, referente a uma matriz de componentes $\mathbf{C}$. Este conceito de modo é manuseado na maioria dos artigos que envolvem modelos de Tucker, pelo que esta designação vai sobrepor-se ao conceito de entidade, como implícito na Definição 2.2.1. A decomposição pode-se traduzir na forma:

$$x_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} a_{ip}b_{jq}c_{kr}g_{pqr} + e_{ijk}, \quad i = 1, ..., I; j = 1, ..., J; k = 1..., K \tag{2.2}$$

onde os coeficientes a_{ip}, b_{jq} e c_{kr} representam, respetivamente, o i-ésimo elemento da componente p da matriz $\mathbf{A}$ $(I \times P)$, o j-ésimo elemento da componente q da matriz $\mathbf{B}$ $(J \times Q)$ e o k-ésimo elemento da componente r da matriz $\mathbf{C}$ $(K \times R)$. O elemento g_{pqr} expressa a interação entre as três componentes de cada modo e encontra-se presente num cubo $\mathbb{G}$ $(P \times Q \times R)$ que vai ser descrito posteriormente. Os valores P, Q e R são o número de componentes selecionados para descrever o primeiro, segundo e terceiro modo, respetivamente. Obviamente que P, Q e R são, respectivamente, menores ou iguais a I, J e K, dado que o número de componentes escolhido não pode ser superior ao número de variáveis em estudo. Notar que não é necessário serem em igual número. Por último, e_{ijk} são erros existentes que são colocados no cubo $\mathbb{E}$. A Figura 2.1 mostra o esquema deste modelo com a decomposição de $\mathbb{X}$.

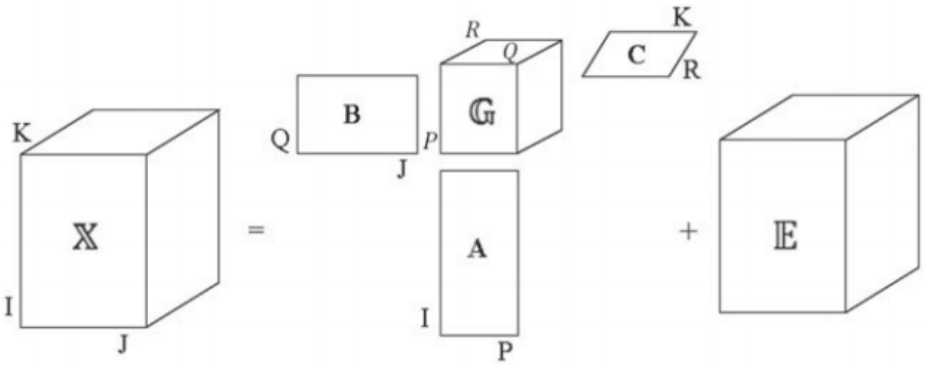

Figura 2.1: Esquema do modelo Tucker3. Decomposição do cubo $\mathbb{X}$ com as matrizes $\mathbf{A}$ $(I \times P)$, $\mathbf{B}$ $(J \times Q)$ e $\mathbf{C}$ $(K \times R)$ e com o cubo $\mathbb{G}$.

Definição 2.2.2 *Sejam* $\boldsymbol{A} = [a_{mn}]$ *e* $\boldsymbol{B} = [b_{pq}]$ *duas matrizes. O produto de Kronecker* $A \otimes B$ *é uma matriz de dimensão* $(mp \times nq)$ *da forma:*

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} a_{11}\boldsymbol{B} & \cdots & a_{1n}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{pmatrix} \tag{2.3}$$

Duas propriedades importantes:

1. $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$, sendo $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ e $\mathbf{D}$ matrizes;
2. $\mathbf{A}^T \otimes \mathbf{B}^T = (\mathbf{A} \otimes \mathbf{B})^T$.

Exemplo 2.2.3 *O produto de Kronecker entre* $\boldsymbol{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ *e* $\boldsymbol{B} = \begin{pmatrix} 0 & 5 \\ 6 & 7 \end{pmatrix}$ *é:*

$$\begin{aligned} \boldsymbol{A} \otimes \boldsymbol{B} &= \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \otimes \begin{pmatrix} 0 & 5 \\ 6 & 7 \end{pmatrix} \\ &= \begin{pmatrix} 1\begin{pmatrix} 0 & 5 \\ 6 & 7 \end{pmatrix} & 2\begin{pmatrix} 0 & 5 \\ 6 & 7 \end{pmatrix} \\ 3\begin{pmatrix} 0 & 5 \\ 6 & 7 \end{pmatrix} & 4\begin{pmatrix} 0 & 5 \\ 6 & 7 \end{pmatrix} \end{pmatrix} \\ &= \begin{pmatrix} 1\times 0 & 1\times 5 & 2\times 0 & 2\times 5 \\ 1\times 6 & 1\times 7 & 2\times 6 & 2\times 7 \\ 3\times 0 & 3\times 5 & 4\times 0 & 4\times 5 \\ 3\times 6 & 3\times 7 & 4\times 6 & 4\times 7 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 5 & 0 & 10 \\ 6 & 7 & 12 & 14 \\ 0 & 15 & 0 & 20 \\ 18 & 21 & 24 & 28 \end{pmatrix} \end{aligned}$$

Outra forma de apresentar o cubo $\mathbb{X}$ é em termos matriciais:

$$\underline{\mathbf{X}} = \mathbf{A}\underline{\mathbf{G}}(\mathbf{C}^T \otimes \mathbf{B}^T) + \underline{\mathbf{E}} \tag{2.4}$$

em que $\underline{\mathbf{X}}$ $(I \times JK)$, $\underline{\mathbf{G}}$ $(P \times QR)$ e $\underline{\mathbf{E}}$ $(I \times JK)$ representam, respetivamente, matrizes bidimensionais dos cubos $\mathbb{X}$, $\mathbb{G}$ e $\mathbb{E}$. Ainda neste capítulo vamos demonstrar a construção das matrizes $\mathbf{A}$, $\mathbf{B}$ e $\mathbf{C}$ tal como o cubo dos pesos $\mathbb{G}$ de cada combinação de componentes.

Observação 2.2.4 *Cada componente de* $\boldsymbol{A}$*,* $\boldsymbol{B}$ *e* $\boldsymbol{C}$ *representa uma combinação linear de* I *objetos,* J *variáveis e* K *condições, respetivamente. Os* I *objetos,* J *variáveis e* K *condições podem ser vistos como níveis do primeiro, segundo e terceiro modo, respetivamente.*

Observação 2.2.5 *As matrizes que apresentem uma linha em baixo da respetiva letra de identificação são matrizes que foram obtidas de um cubo, isto é, apresentam uma dimensão a menos. Isto procede-se mantendo-se o mesmo número de linhas e multiplicando o número de colunas ao número de condições.*

2.3 Modelo Co-Tucker

Esta secção responsabiliza-se por mostrar a construção do Co-Tucker e os seus principais pressupostos. Para além disso, é também apresentada uma componente algorítmica que é essencial à construção e eficiência do mesmo.

2.3.1 Construção do modelo Co-Tucker

Explicado como se trabalha a análise da co-inércia e a construção do Tucker3, estamos em condições de partir para a análise do modelo Co-Tucker. Este modelo não é mais nem menos que uma combinação dos anteriores [7]. Desta forma, Co-Tucker traduz-se na junção da análise da co-inércia que é efetuada K vezes, uma vez que temos K matrizes de covariâncias cruzadas, com o modelo Tucker3, que analisa estas novas K matrizes. Não esquecer que o valor de K corresponde ao número de condições em estudo.

O primeiro passo decorre da transformação de dois conjuntos de dados longitudinais $\mathbb{X}_{N\times I\times K}$ e $\mathbb{Y}_{N\times J\times K}$ num único. Depois do tratamento dos dados (centralização ou normalização), são calculadas as matrizes de covariâncias cruzadas $\mathbf{W}_K$, isto é, a execução da análise de co-inércia K vezes, onde $\mathbf{W}_K = \mathbf{Y}_K^T\mathbf{D}_N\mathbf{X}_K$, e $\mathbf{X}_K$ e $\mathbf{Y}_K$ representam, respetivamente, o primeiro e segundo conjunto de dados na condição K e $\mathbf{D}_N$ é uma matriz diagonal de pesos $1/N$. Desta forma, cada matriz $\mathbf{W}_K$ tem dimensão $(I\times J)$. Esta sequência de matrizes dá origem ao cubo $\mathbb{W}$ $(I\times J\times K)$. Todo este processo encontra-se esquematizado na Figura 2.2:

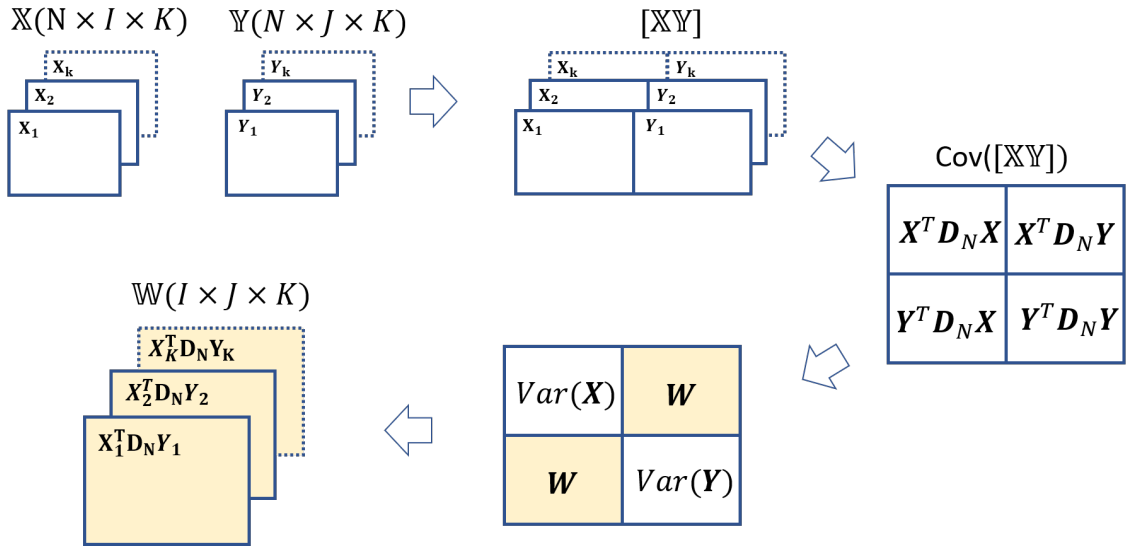

Figura 2.2: Passo 1 do Co-Tucker: realização de K tabelas cruzadas a partir dos dados longitudinais $\mathbb{X}$ e $\mathbb{Y}$ através de K análises co-inércia.

O segundo passo ocorre da aplicação do modelo Tucker3 sobre o cubo $\mathbb{W}$, resultando na sua decomposição como em 2.2:

$$w_{ijk} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} a_{ip}^X b_{jq}^Y c_{kr} g_{pqr} + e_{ijk}, \quad i = 1,...,I; j = 1,...,J; k = 1...,K \tag{2.5}$$

onde os coeficientes a_{ip}^X, b_{jq}^Y e c_{kr} representam, respetivamente, o i-ésimo elemento da componente p da matriz $\mathbf{A}^X$ $(I\times P)$, o j-ésimo elemento da componente q da matriz $\mathbf{B}^Y$ $(J\times Q)$ e o k-ésimo elemento da componente r da matriz $\mathbf{C}$ $(K\times R)$. A notação $\mathbf{A}^X$ e $\mathbf{B}^Y$ apenas serve de ajuda informativa no sentido de esclarecer que a matriz $\mathbf{A}$ está relacionada com o primeiro conjunto de dados $\mathbb{X}$ e a matriz $\mathbf{B}$ relacionada com o segundo conjunto de dados $\mathbb{Y}$. Para uma melhor compreensão, a Figura 2.3 mostra esta decomposição esquematizada através da construção das matrizes $\mathbf{A}^X$, $\mathbf{B}^Y$ e $\mathbf{C}$.

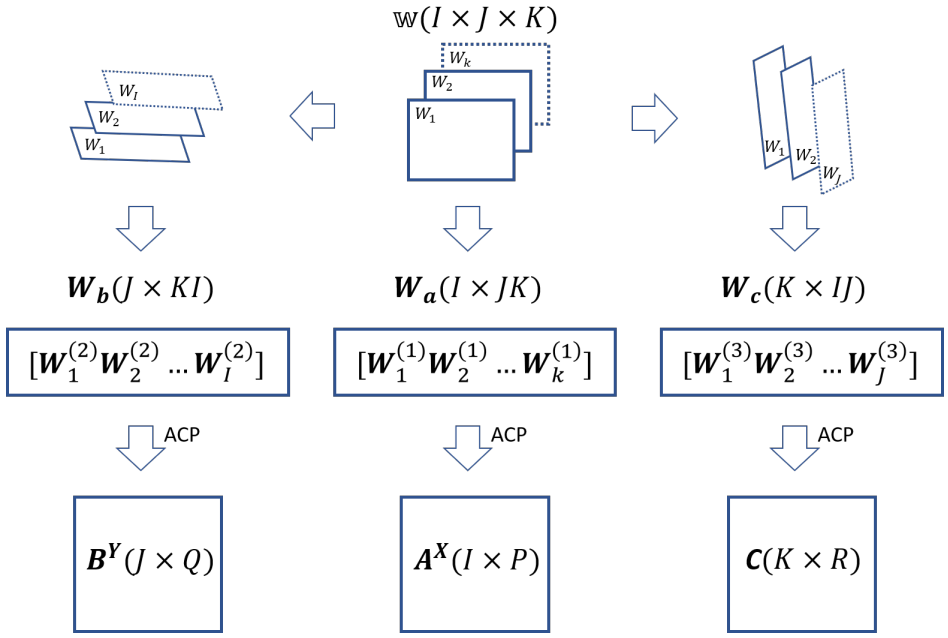

Figura 2.3: Passo 2 do Co-Tucker: decomposição do cubo $\mathbb{W}$ via ACP.

Como já referido anteriormente, cada coluna destas matrizes correspondem a componentes, isto é, uma combinação linear das variáveis em estudo. Neste sentido, a matriz $\mathbf{A}^X$, pertencente ao primeiro modo, vai ter P combinações lineares envolvendo as I variáveis presentes em $\mathbb{X}$, com $P \leq I$. A matriz $\mathbf{B}^Y$, inserida no segundo modo, tem Q combinações lineares envolvendo as J variáveis presentes em $\mathbb{Y}$, com $Q \leq J$. Por último, a matriz $\mathbf{C}$, pertencente ao terceiro modo, vai ter R combinações lineares envolvendo as K condições, com $R \leq K$. A fatorização em 2.5 pode ser escrita da forma:

$$w_{ijk} = \widehat{w}_{ijk} + e_{ijk}, \quad i = 1, ..., I; j = 1, ..., J; k = 1..., K \tag{2.6}$$

tal que $\widehat{w}_{ijk}$ corresponde aos coeficientes expressos no modelo $\widehat{\mathbb{W}}$ que queremos construir e que é dependente do número de componentes escolhidas P, Q e R. Obviamente que esta redução de dimensionalidade vai levar a uma perda de informação, sendo que estas mesmas informações vão ser agora expressas em combinações lineares. Um número mais elevado de componentes escolhidas em cada modo possibilita um modelo $\widehat{\mathbb{W}}$ com maior variância explicada, isto é, a informação lá expressa aproxima-se da informação contida em $\mathbb{W}$. Obviamente que vamos ter um modelo com maior complexidade, porém a dificuldade de interpretação do mesmo vai ser maior. No sentido contrário, e utilizando um número de componentes mais reduzido, a variância explicada por $\widehat{\mathbb{W}}$ vai ser menor, contudo a dificuldade de interpretação vai ser menor também, dada a sua reduzida complexidade.

De forma a compreendermos a potencialidade do modelo $\widehat{\mathbb{W}}$ construído, é essencial determinar a variância explicada por este. Este é um simples cálculo que envolve a divisão entre a soma do quadrado dos coeficientes em $\mathbb{G}$ e a soma do quadrado dos coeficientes em $\mathbb{W}$:

$$\mathfrak{V}_{\widehat{\mathbb{W}}} = \frac{\sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr}^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} w_{ijk}^2} \tag{2.7}$$

Definição 2.3.1 *Sejam u e v dois vetores. Os vetores u e v são ortonormais se forem ortogonais entre si e ambos com norma unitária, isto é:*

1. $\langle u, v\rangle = 0$
2. $\|u\| = 1$
3. $\|v\| = 1$

Definição 2.3.2 *Seja $\boldsymbol{A}_{I\times J}$ uma matriz. $\boldsymbol{A}$ é ortonormal (em coluna) se $\boldsymbol{A}^T\boldsymbol{A} = I_J$ onde I_J é uma matriz identidade de dimensão $(J \times J)$.*

Exemplo 2.3.3 *A matriz* $\boldsymbol{A}_{3\times 2} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}$ *é ortonormal em coluna:*

$$\boldsymbol{A}^T\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Observar que as colunas da matriz $\boldsymbol{A}$ são ortonormais.

Genericamente, cada uma das matrizes $\mathbf{A}^X$, $\mathbf{B}^Y$ e $\mathbf{C}$ podem ser vistas como um conjunto de combinações lineares obtidas através de uma ACP. Para além disso, são construídas de forma a serem ortogonais e cada coluna tem norma unitária. Com estas restrições e com o cubo $\mathbb{G}$ obtido da forma bidimensional

$$\underline{\mathbf{G}}_{P\times QR} = \mathbf{A}^T\underline{\mathbf{W}}_{I\times JK}(\mathbf{C}\otimes\mathbf{B}) \tag{2.8}$$

fazem com que o quadrado de cada elemento de $\mathbb{G}$, g^2_{pqr}, indique a importância da combinação das componentes p, q e r no modelo $\widehat{\mathbb{W}}$. Interessante notar que a informação contida em $\mathbb{W}$ representa o valor ou peso da combinação entre os níveis dos modos originais, e, de forma similar, a informação contida em $\mathbb{G}$ representa o valor ou interação da combinação entre as componentes dos modos. Um valor absoluto em $\mathbb{G}$ que seja grande comparativamente aos outros merece uma espacial atenção na sua interpretação, uma vez que daqui conseguimos extrair as combinações p, do primeiro modo, q do segundo e r do terceiro modo, que traduzem mais informação sobre o cubo $\mathbb{W}$. Neste sentido, a grandeza

$$\mathfrak{V}_{pqr} = \mathfrak{V}_{\widehat{\mathbb{W}}} \times \frac{g^2_{pqr}}{\sum_{p=1}^{P}\sum_{1=1}^{Q}\sum_{r=1}^{R} g^2_{pqr}} \tag{2.9}$$

indica-nos a fração da variabilidade que é explicada por esta combinação de componentes. Se quisermos este valor expresso em percentagem, basta multiplicar $\mathfrak{V}_{pqr}$ por 100. É também a partir de $\mathbb{G}$ que conseguimos obter a variabilidade que é capturada por uma específica componente. Por exemplo, para determinar a importância da componente p no primeiro modo, o cálculo resulta na seguinte expressão:

$$\mathfrak{V}_p = \frac{\sum_{q=1}^{Q}\sum_{r=1}^{R} g^2_{qpr}}{\sum_{p=1}^{P}\sum_{1=1}^{Q}\sum_{r=1}^{R} g^2_{pqr}} \tag{2.10}$$

O mesmo pode ser realizado para outra qualquer componente de outro qualquer modo.

2.3.2 Construção das matrizes $\mathbf{A}^X$, $\mathbf{B}^Y$ e $\mathbf{C}$

Nesta secção, será mostrada a construção das matrizes $\mathbf{A}^X$, $\mathbf{B}^Y$ e $\mathbf{C}$. Na literatura existem muitas formas de construção, no entanto todas elas têm o mesmo objetivo - aproximar o máximo possível o modelo $\widehat{\mathbb{W}}$ ao cubo $\mathbb{W}$ tendo em conta o número de componentes escolhidas para cada modo. Esta aproximação pode ser encarada como um problema de minimização da seguinte forma:

$$\min_{\mathbf{A}^X,\mathbf{B}^Y,\mathbf{C}} \quad \|\mathbb{W} - \widehat{\mathbb{W}}\|_F^2 = \|\underline{\mathbf{W}} - \mathbf{A}^X\underline{\mathbf{G}}(\mathbf{C}^T \otimes \mathbf{B}^T)\|_F^2 \tag{2.11}$$

com $\underline{\mathbf{W}}$ e $\underline{\mathbf{G}}$ matrizes bidimensionais $(I \times JK)$ e $(P \times QR)$, respetivamente, $\mathbf{A}^X$, $\mathbf{B}^Y$ e $\mathbf{C}$ encontram-se sob a restrição de ortonormalidade e $\|.\|_F$ representa a norma de Frobenius.

Definição 2.3.4 *A norma de Frobenius de uma matriz $A_{m\times n}$, também denominada por norma euclideana, calcula-se da seguinte forma:*

$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} = \sqrt{Traço(\boldsymbol{A}\boldsymbol{A}^T)} \tag{2.12}$$

No R, uma função que permite este cálculo é norm(A,"F").

Definição 2.3.5 *A norma de Frobenius de um cubo $\mathbb{A}_{m\times n\times c}$, calcula-se da seguinte forma:*

$$\|\mathbb{A}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{c}|a_{ijk}|^2} \tag{2.13}$$

Definição 2.3.6 *A decomposição em valores singulares (singular value decomposition - SVD) de uma matriz $\boldsymbol{A}_{m\times n}$ ocorre da seguinte expressão:*

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T \tag{2.14}$$

onde $\boldsymbol{U}$ e $\boldsymbol{V}$ são matrizes ortogonais de dimensão $(m \times n)$ e $(n \times n)$, respetivamente, e $\boldsymbol{\Lambda}$ é uma matriz diagonal com valores próprios e de dimensão $(n \times n)$.

Nesta dissertação, o método que utilizamos para a minimização da função em 2.11, e de forma a encontrar os melhores coeficientes das matrizes em questão, é o método dos mínimos quadrados alternados, em inglês, alternating least squares - ALS [6][8]. Neste processo a estimação dos coeficientes de uma matriz ocorre através da estimação fixa dos coeficientes das restantes duas matrizes, daí advém o nome "alternados". Todo este processo dá-se num determinado número de iterações até que o critério de convergência falhe. Este algorítmo é definido nos seguintes passos:

1. Inicializar $\mathbf{B}^Y$ e $\mathbf{C}$ e definir o valor de convergência ϵ.

2. Tranformação de $\mathbb{W}$ numa matriz $\mathbf{W}_a$ $(I \times JK)$.
   Decomposição através de SVD em $\mathbf{W}_a(\mathbf{C} \otimes \mathbf{B}^Y) = \mathbf{H}_a\mathbf{\Lambda}_a\mathbf{H}_a^T$.
   Matriz do primeiro modo $\mathbf{A}^X = \mathbf{H}_{a_{(1:P)}}$ onde $1:P$ representam as P colunas escolhidas de $\mathbf{H}_a$.

3. Tranformação de $\mathbb{W}$ numa matriz $\mathbf{W}_b$ $(J \times IK)$.
   Decomposição através de SVD em $\mathbf{W}_b(\mathbf{C} \otimes \mathbf{A}^X) = \mathbf{H}_b\mathbf{\Lambda}_b\mathbf{H}_b^T$.
   Matriz do primeiro modo $\mathbf{B}^Y = \mathbf{H}_{b(1:Q)}$ onde $1:Q$ representam as Q colunas escolhidas de $\mathbf{H}_b$.

4. Tranformação de $\mathbb{W}$ numa matriz $\mathbf{W}_c$ $(K \times IJ)$.
   Decomposição através de SVD em $\mathbf{W}_c(\mathbf{B}^Y \otimes \mathbf{A}) = \mathbf{H}_c \mathbf{\Lambda}_c \mathbf{H}_c^T$.
   Matriz do primeiro modo $\mathbf{C} = \mathbf{H}_{c(1:R)}$ onde $1:R$ representam as R colunas escolhidas de $\mathbf{H}_c$.

5. Calcular matriz $\underline{\mathbf{G}} = \mathbf{A}^{X^T} \mathbf{W}_a (\mathbf{C} \otimes \mathbf{B}^Y)$

6. Calcular $f_n = \|\underline{\mathbf{W}} - \mathbf{A}^X \underline{\mathbf{G}} (\mathbf{C}^T \otimes \mathbf{B}^{Y^T})\|_F^2$ na iteração n. Se $f_{n-1} - f_n > \epsilon$ ir novamente para o passo 2. Caso contrarírio $\mathbf{A}^X$, $\mathbf{B}^Y$, $\mathbf{C}$ e $\mathbb{G}$ estão determinados.

Este algorítmo produz as matrizes ortogonais em cada modo e colunas de norma unitárias, como era suposto. Observar que, para uma aproximação ótima e uma ortonormalização perfeita temos:

$$\begin{aligned}
\min_{\mathbf{A}^X, \mathbf{B}^Y, \mathbf{C}} \quad & \|\mathbb{W} - \widehat{\mathbb{W}}\|_F^2 \\
&= \|\underline{\mathbf{W}} - \mathbf{A}^X \underline{\mathbf{G}} (\mathbf{C}^T \otimes \mathbf{B}^{Y^T})\|_F^2 \\
&= \|\underline{\mathbf{W}} - \mathbf{A}^X \mathbf{A}^{X^T} \underline{\mathbf{W}} (\mathbf{C} \otimes \mathbf{B}^{Y^T})(\mathbf{C}^T \otimes \mathbf{B}^{Y^T})\|_F^2 \quad \text{substituir } \underline{G} \text{ dada em 2.8} \\
&= \|\underline{\mathbf{W}} - (\mathbf{A}^X \mathbf{A}^{X^T}) \underline{\mathbf{W}} (\mathbf{C}\mathbf{C}^T \otimes \mathbf{B}^Y \mathbf{B}^{T^Y})\|_F^2 \quad \text{prop. do prod. de Kronecker} \\
&= \|\underline{\mathbf{W}} - I_d \underline{\mathbf{W}} (I_d \otimes I_d)\|_F^2 \quad \text{ortogonalidade} \\
&= \|\underline{\mathbf{W}} - \underline{\mathbf{W}}\|_F^2 \\
&= 0
\end{aligned} \tag{2.15}$$

Minimar a função $\|\mathbb{W} - \widehat{\mathbb{W}}\|_F^2$ é idêntico a minimizar $f_{\mathbb{E}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} e_{ijk}^2$, dada a expressão em 2.6. Obviamente que quanto menores, em valor absoluto, forem os coeficientes no cubo dos erros $\mathbb{E}$, mais próximo estará o modelo $\widehat{\mathbb{W}}$ de $\mathbb{W}$.

A inicialização das matrizes $\mathbf{B}^Y$ e $\mathbf{C}$ no primeiro passo poderá ser efetuada de várias formas. Uma delas é uma inicialização onde ambas as matrizes têm coeficientes nulos. Outra alternativa seria uma inicialização com valores aleatórios. Outro procedimento de inicialização que é um pouco diferente e mais complexo é o seguinte:

- Transformação de $\mathbb{W}$ numa matriz $\mathbf{W}_b$ $(J \times IK)$;
  Aplicar SVD em $\mathbf{W}_b \mathbf{W}_b^T$ de forma a produzir a igualdade $\mathbf{W}_b \mathbf{W}_b^T = \mathbf{H}_b \mathbf{\Lambda}_b \mathbf{H}_b^T$;
  Para a iteração $n = 1$ temos a matriz do primeiro modo $\mathbf{B}^Y = \mathbf{H}_{b(1:Q)}$, onde $1:Q$ representam as Q colunas escolhidas de $\mathbf{H}_b$.

- Transformação de $\mathbb{W}$ numa matriz $\mathbf{W}_c$ $(K \times IJ)$;
  Aplicar SVD em $\mathbf{W}_c \mathbf{W}_c^T$ de forma a produzir a igualdade $\mathbf{W}_c \mathbf{W}_c^T = \mathbf{H}_c \mathbf{\Lambda}_c \mathbf{H}_c^T$;
  Para a iteração $n = 1$ temos a matriz do primeiro modo $\mathbf{C} = \mathbf{H}_{c(1:R)}$, onde $1:R$ representam as R colunas escolhidas de $\mathbf{H}_c$.

Todos os métodos de inicialização convergem para um mínimo local ou global, no entanto este último descrito tem tendência a apresentar menos iterações e mais facilidade tem em atingir o mínimo global. Existe também a função *tucker* no R da biblioteca *CA3variants* que utiliza outro tipo de inicialização e vai ser também utilizada nesta dissertação. A ideia é utilizar todos estes procedimentos de fácil implementação e obter o melhor modelo.

A obtenção de valores ótimos para as matrizes em estudo através do método dos mínimos quadrados alternados poderá ser bastante demorada e computacionalmente exigente. Contudo, dado que, posteriormente, vamos trabalhar com dados de dimensões relativamente não elevadas, vamos considerar este ponto um não problema na aplicação proposta. Por curiosidade, e de forma a evitar este método computacionalmente exigente, um método proposto por Kiers e Kinderens [2], é a obtenção de $\mathbf{B}^Y$ e de $\mathbf{C}$ através dos dois pontos descritos anteriormente e a obtenção de $\mathbf{A}^X$ de forma semelhante, mas devidamente adaptada. Neste caso não existe qualquer iteração e a aproximação do modelo construído é considerada relativamente boa [2].

Exemplo 2.3.7 *De forma a comparar a performance entre estes procedimentos, vamos ter em conta a base de dados longitudinal presente no R data(meau) da biblioteca ade4. Depois de construído o cubo* $\mathbb{W}_{10\times13\times4}$ *procedemos ao método dos mínimos quadrados alternados sobre* 278 *modelos com três diferentes inicializações e com um critério de convergência* $\epsilon = 10^{-6}$. *A inicialização de* $\boldsymbol{B}^Y$ *e* $\boldsymbol{C}$ *com valores nulos apresenta um número médio de iterações de* 7.40. *Para a inicialização envolvendo o SVD de* $\boldsymbol{W}_b\boldsymbol{W}_b^T$ *e* $\boldsymbol{W}_c\boldsymbol{W}_c^T$ *temos um número médio de* 4.82. *A função tucker do R é a que apresenta um maior número médio de iterações -* 13.42. *A variância explicada pelos* 278 *modelos construídos é exatamente igual entre os três métodos. Observa-se assim, neste exemplo, que a principal diferença está relacionada com a velocidade de convergência. A programação de código deste exemplo está disponível no github em* `https://github.com/Francisjcs1997/CT/blob/main/Exemplo_ALS.R`.

Observação 2.3.8 *Dada a expressão da norma de Frobenius na Definição 2.12 e 2.13, facilmente se observa que a norma da diferença tridimensional* $\mathbb{W} - \widehat{\mathbb{W}}$ *é exatamente igual à norma da diferença bidimensional* $\underline{\boldsymbol{W}} - \underline{\widehat{\boldsymbol{W}}}$.

2.4 Escolha do modelo

Para o investigador é essencial produzir um modelo capaz de transmitir, da melhor forma possível, as informações retidas nos dados que está a tratar. Um modelo mais simples provavelmente não nos transmitirá o maior número de informação. Um modelo mais complexo eventuamente vai sofrer um sobreajuste e, porventura, será também difícil de o interpretar. É portanto fundamental determinar o melhor modelo com base na sua complexidade, isto é, o número de componentes escolhidas, de forma a encontrar-mos um equilíbrio de acordo com os problemas aqui mencionados.

Vários métodos e heurísticas foram propostas ao longo dos anos. Um dos grandes problemas de muitos destes métodos é que envolvem uma análise univariada em cada modo, não tendo em consideração o que acontece nos restantes. Um exemplo é a determinação do número de componentes de cada modo através do "scree-test"envolvendo os valores próprios presentes em $\boldsymbol{\Lambda}_a$, $\boldsymbol{\Lambda}_b$ e $\boldsymbol{\Lambda}_c$. Para contornar estes problemas, a escolha dos modelos apresentados nesta dissertação vão ser obtidos através do método DifFit [3][5].

O método DifFit carateriza-se por analisar a variância que é explicada por modelos com diferentes dimensões $S = P+Q+R$, tendo em consideração a alteração desta mesma variância através da adição de componentes. Geralmente, a complexidade ideal do modelo Co-Tucker é aquela que requer o menor número de componentes e que captura, ao mesmo tempo, uma alta variância dos dados. Com isto em conta, o método DifFit segue os seguintes passos:

1. Considerar todos os modelos Co-Tucker $\widehat{\mathbb{W}}$ com $P = P_1, ..., P_I$, $Q = Q_1, ..., Q_J$ e $R = R_1, ..., R_K$ e calcular a sua variância explicada $\mathfrak{V}_{\widehat{\mathbb{W}}}$. Ter em conta a restrição $P \leq QR$, $Q \leq PR$ e $R \leq PQ$ de forma a eliminar soluções redundantes;

2. Para cada valor de $S = P + Q + R$, determinar o modelo $\widehat{\mathbb{W}}_S$ com maior variância explicada ($S = 3, 5, 6, ..., P_I + Q_J + R_K$);

3. Comparar os modelos anteriormente escolhidos:

   (a) Calcular $dif_s = \mathfrak{V}_{\widehat{\mathbb{W}}_S} - \mathfrak{V}_{\widehat{\mathbb{W}}_{S-1}}$

   (b) Considerar apenas as diferenças que são sequencialmente mais altas;

   (c) Calcular $b_s = dif_s/dif_{s^*}$, onde dif_{s^*} é o próximo valor mais alto depois de dif_s.

   (d) Calcular o valor crítico $V_c = 1/(S(min) - 3)$, onde $S(min) = min(P, QR) + min(Q, PR) + min(R, PQ)$;

   (e) Entre os modelos com $V_c \leq dif_s$, escolher aquele com maior valor b_s.

Para este método não foi encontrado qualquer comando específico no R. Desta forma foi realizada a programação de cada um destes passos no mesmo software que se encontra disponível no github.

Depois da construção de cada uma das matrizes com as respetivas combinações lineares das variáveis de $\mathbb{X}$ e $\mathbb{Y}$ e das K condições e depois de escolhido o melhor modelo, estamos preparados para analisar as interações obtidas entre as variáveis dos dois cubos de dados nos diferentes espaços temporais. Nesta análise entra também o cubo $\mathbb{G}$ que nos vai indicar as combinações de componentes mais relevantes.

2.5 Interpretação de resultados

A interpretação dos resultados obtidos por parte do investigador pode ser efetuada por duas vertentes: uma via numérica, onde é analisada a interação das variáveis envolvendo as matrizes $\mathbf{A}^X$, $\mathbf{B}^Y$ e $\mathbf{C}$ de cada modo e o cubo $\mathbb{G}$, ou por uma via gráfica, envolvendo a projeção de componentes e a construção de biplots. Obviamente que uma análise que incluí estas duas vias será mais enriquecedora do ponto de vista prático.

2.5.1 Interpretação numérica

Depois de construído o modelo com o número de componentes selecionado, a análise de $\mathbf{A}^X$, $\mathbf{B}^Y$, $\mathbf{C}$ e $\mathbb{G}$ vai permitir investigar a interação entre as variáveis de $\mathbb{X}$ e $\mathbb{Y}$ nas diferentes condições avaliadas. Ao investigador é essencial determinar quando existe uma interação positiva, isto é, as variáveis que apresentam um crescimento ou decrescimento simultâneo nos dois conjuntos de dados, e determinar quando existe uma interação negativa, ou seja, conhecer que variáveis mostram crescimentos ou decrescimentos em sentidos contrários.

O primeiro elemento a ser estudado vai ser o cubo $\mathbb{G}$ ($P \times Q \times R$). Como já mencionado anteriormente, este cubo é responsável por nos apresentar os pesos de cada combinação de componentes entre os três modos. Obviamente que as primeiras análises vão ser efetuadas nos coeficientes que apresentam maiores valores absolutos, no sentido que aqui estão expressas as combinações que extraem maior variabilidade explicada. Seja g_{pqr} o coeficiente que apresenta maior valor absoluto. Significa isto que vamos estudar a p-ésima componente de $\mathbf{A}^X$ do

primeiro modo, a q-ésima componente de $\mathbf{B}^Y$ do segundo modo e a r-ésima componente de $\mathbf{C}$ do terceiro modo. Para além de analisarmos o valor absoluto de g_{pqr}, é fundamental também ter em consideracão o seu sinal. Este sinal é responsável pela determinação do tipo de interação, positiva ou negativa. Vamos considerar, por agora, este coeficiente positivo, $sinal(g_{ijk}) = (+)$. Posteriormente, vamos focar e explicar melhor a sua importância.

Com as componentes selecionadas vamos agora observar cada uma delas nas respetivas matrizes. Nestas, vamos olhar também para os valores absolutos e para os sinais de cada coeficiente tal como anteriormente. Consideremos o i-ésimo elemento da p-ésima componente da matriz $\mathbf{A}^X$, o j-ésimo elemento da q-ésima componente da matriz $\mathbf{B}^Y$ e o k-ésimo elemento da r-ésima componente da matriz $\mathbf{C}$ aqueles que apresentam maiores valores absolutos e sinais positivos. Significa isto que vamos estudar a interação da i-ésima variável de $\mathbb{X}$ com a j-ésima variável de $\mathbb{Y}$ na k-ésima condição. A considerável amplitude do valor absoluto destes coeficientes em cada componente expressa que existe uma interação entre estas variáveis. Resta determinar se este tipo de interação é positiva ou negativa. Nesse ponto vamos ter em conta a multiplicação dos quatro sinais obtidos nos quatro coeficientes em estudo das matrizes $\mathbf{A}^X$, $\mathbf{B}^Y$ e $\mathbf{C}$ e do cubo $\mathbb{G}$, tal como a seguinte expressão indica:

$$sinal(I) = sinal(a_{ip}) \times sinal(b_{jq}) \times sinal(c_{kr}) \times sinal(g_{pqr}) \tag{2.16}$$

onde $sinal(I)$ representa o tipo de interação, ou seja, positiva se $sinal(I) = (+)$ ou negativa se $sinal(I) = (-)$. Uma vez que considerámos anteriormente coeficientes positivos nas matrizes de cada modo e visto que considerámos também g_{pqr} um coeficiente positivo, então a i-ésima variável de $\mathbb{X}$ apresentará uma interação positiva com a j-ésima variável de $\mathbb{Y}$ na condição k, dada a expressão em 2.16.

Dado que a interação resulta de uma combinação de quatro sinais, podemos concluir que uma interação positiva verifica-se sempre quando temos um número par de sinais negativos ou positivos. Caso contrário a interação será negativa. Não esquecer que esta análise é obtida sobre coeficientes com valores absolutos consideráveis. O estudo da interação através da análise de variáveis com coeficientes com valores absolutos pequenos ou próximos de zero deve ser ignorada, uma vez que não são relevantes, dado que estão mais sujeitas ao ruído que o modelo poderá apresentar com a redução de dimensionalidade realizada.

Exemplo 2.5.1 *Consideremos um modelo com* 73% *de variância explicada e de dimensão* $(2 \times 2 \times 1)$ *com matrizes* $\boldsymbol{A}^X_{5\times2}$, $\boldsymbol{B}^Y_{3\times2}$ *e* $\boldsymbol{C}_{4\times1}$ *descritas na Tabela 2.1 de cada modo e o cubo* $\mathbb{G}_{2\times2\times1}$ *na Tabela 2.2.*

$\boldsymbol{A}^X$		
Variáveis	P_1	P_2
Ph	0.360	0.476
Condutividade	−0.551	0.417
Oxigénio	0.507	0.330
Nitratos	0.244	0.519
Fósfato	−0.499	0.471

$\boldsymbol{B}^Y$		
Espécies	Q_1	Q_2
Bsp	0.754	−0.653
Brh	0.654	0.756
Cae	0.045	−0.049

$\boldsymbol{C}$	
Períodos	R_1
Primavera	−0.214
Verão	0.283
Outono	0.880
Inverno	−0.313

Tabela 2.1: Matrizes exemplo $\mathbf{A}^X$, $\mathbf{B}^Y$ e $\mathbf{C}$ do 1º, 2º e 3º modo, respetivamente.

Cubo $\mathbb{G}$		
R_1		
Componentes	Q_1	Q_2
P_1	4.005	-0.151
P_2	0.001	-2.033

Tabela 2.2: Cubo exemplo $\mathbb{G}$.

Analisando a Tabela 2.2, o coeficiente com maior valor, em termos absolutos, situa-se na coordenada $(1 \times 1 \times 1)$ *e é positivo. A variabilidade que é explicada por esta combinação é calculada a partir da expressão 2.9:*

$$\mathfrak{V}_{111} = 0.73 \times \frac{4.005^2}{4.005^2 + (-0.151)^2 + 0.001^2 + (-2.033)^2} = 0.372$$

Vamos agora analisar as primeiras componentes de cada modo. Podemos visualizar que, no primeiro modo, o oxigénio apresenta o coeficiente mais positivo $(+0.507)$*. No segundo modo todos os coeficientes são positivos. As espécies Bsp e Brh apresentam os maiores valores* $(0.754$ *e* 0.654*, respetivamente). No último modo o outono apresenta o maior valor* (0.880)*. Utilizando a expressão em (2.16):*

$$sinal(I) = (+) \times (+) \times (+) \times (+) = (+)$$

Desta forma, concluímos que existe uma interação positiva, isto é, um aumento nos níveis de oxigénio contribuem para um aumento de todas as espécies (todas com sinais positivos), principalmente Bsp e Brh, na estação outono essencialmente (maior valor positivo). Por exemplo, em relação às variáveis condutividade e fosfato, a interação seria negativa, uma vez que a combinação resultante seria $(-) \times (+) \times (+) \times (+) = (-)$*.*

A ideia seria continuar a analisar as seguintes combinações de componentes, numa ordem decrescente em termos de valores absolutos dos coeficientes em $\mathbb{G}$*. Fica ao critério do leitor continuar esta análise.*

2.5.2 Interpretação gráfica

Como já referido anteriormente, para além da análise numérica estudada através das matrizes de cada modo, a interação entre variáveis numa via gráfica pode ser uma mais valia na clarificação dos resultados obtidos. De uma forma muito geral, a ideia que está por detrás deste procedimento é representar, usando biplots, uma projeção das componentes produzidas nos dois primeiros modos em cada componente do terceiro modo [1].

Designemos, a partir de agora, como plano frontal a matriz $\mathbb{G}_r$ $(P \times Q)$ obtida do cubo $\mathbb{G}$ na condição r. Desta forma, $\mathbb{G}$ é constituída por R planos frontais, correspondentes a cada componente do terceiro modo, onde cada um deles tem os pesos das $P \times Q$ combinações de componentes. É a partir destes planos frontais que, juntamente com as matrizes $\mathbf{A}^X$ e $\mathbf{B}^Y$ do primeiro e segundo modo, se constroem os biplots. Depois de ajustado o modelo Co-Tucker, esta construção vai partir de uma decomposição da matriz Δ [7], de dimensões $(I \times J)$, definida por:

$$\Delta = \mathbf{A}^X \mathbb{G}_r (\mathbf{B}^Y)^T \tag{2.17}$$

Neste sentido, cada plano frontal, vai sofrer uma decomposição em valores singulares, da forma $\mathbb{G}_r = \mathbf{U}_r \mathbf{\Lambda}_r \mathbf{V}_r^T$. Este procedimento pode ser interpretado como uma rotação de componentes

por uma matriz ortonormal, seguido de um alongamento ou encolhimento das mesmas [1]. Desta maneira, temos a expressão desenvolvida da seguinte forma:

$$\begin{aligned}\Delta &= \mathbf{A}^X \mathbb{G}_r (\mathbf{B}^Y)^T \\ &= \mathbf{A}^X \mathbf{U}_r \mathbf{\Lambda}_r \mathbf{V}_r^T (\mathbf{B}^Y)^T, \quad \text{SVD em } \mathbb{G}_r \\ &= \mathbf{A}^X \mathbf{U}_r \mathbf{\Lambda}_r^{1/2} (\mathbf{\Lambda}_r^{1/2})^T \mathbf{V}_r^T (\mathbf{B}^Y)^T \\ &= \underbrace{\left(\frac{I}{J}\right)^{1/4} \mathbf{A}^X \mathbf{U}_r \mathbf{\Lambda}_r^{1/2}}_{(\mathbf{A}^X)_r^*} \underbrace{\left[\left(\frac{J}{I}\right)^{1/4} \mathbf{B}^Y \mathbf{V}_r \mathbf{\Lambda}_r^{1/2}\right]^T}_{(\mathbf{B}^Y)_r^{*T}}\end{aligned} \tag{2.18}$$

Temos assim as coordenadas $(\mathbf{A}^X)_r^*$ das I variáveis de $\mathbb{X}$ e as coordenadas $(\mathbf{B}^Y)_r^*$ das J variáveis de $\mathbb{Y}$ projetadas no mesmo espaço. Os quocientes $(I/J)^{\frac{1}{4}}$ e $(J/I)^{\frac{1}{4}}$ permite-nos ter uma distância comparável [1]. O número de eixos que podemos construir depende do número de componentes escolhidas nos primeiros dois modos. Este número é $M = min(P, Q)$.

Em termos de projeções, direções e proximidades estes biplots são interpretados exatamente como os biplots clássicos [7]. Para a leitura e interpretação destes, temos de ter em conta os níveis da matriz $\mathbf{C}$ que representam as K condições. Suponhamos que as componentes dos dois primeiros modos são projetadas na r-ésima componente da matriz $\mathbf{C}$ que é dominada por um alto valor positivo no k-ésimo nível do terceiro modo. Suponhamos também que o i-ésimo nível do primeiro modo encontra-se próximo do j-ésimo nível do segundo modo no biplot em estudo. Significa isto que, a i-ésima variável de $\mathbb{X}$ e a j-ésima variável de $\mathbb{Y}$ apresentam uma interação positiva na k-ésima condição. No caso de $\mathbf{C}$ ser dominada por um valor negativamente alto no k-ésimo nível, a interação seria negativa, isto é, quanto maior a proximidade, mais negativa a interação é. A proximidade entre duas variáveis aqui mencionada é traduzida simplesmente pelo seu produto interno através das matrizes construídas $(\mathbf{A}^X)_r^*$ e $(\mathbf{B}^Y)_r^*$:

$$d_{ij}^r = \sum_{m=1}^{M} (a_{im}^X)^*{}_r (b_{jm}^Y)^*{}_r \tag{2.19}$$

Quanto maior esta distância, maior a proximidade entre a i-ésima variável de $\mathbb{X}$ e a j-ésima variável de $\mathbb{Y}$. Considerando a equação:

$$cos(\theta) = \frac{\langle u, v \rangle}{\|u\| \|v\|}, \quad \text{sendo } \theta \text{ o ângulo entre os vetores } u \text{ e } v.$$

Podemos retirar algumas conclusões acerca das proximidades patentes no biplot resultante da projeção na r-ésima componente do terceiro modo:

- **Caso 1**: Para $d_{ij}^r > 0$, temos $\langle \mathbf{A}_i^X, \mathbf{B}_j^Y \rangle > 0$, ou seja, o ângulo α formado pelos vetores das variáveis i e j é agudo porque $cos(\alpha) > 0$. Existe assim, à partida, uma interação positiva entre estas variáveis nos k-ésimos níveis positivos de $\mathbf{C}$;
- **Caso 2**: Para $d_{ij}^r < 0$, temos $\langle \mathbf{A}_i^X, \mathbf{B}_j^Y \rangle < 0$, ou seja, o ângulo β formado pelos vetores das variáveis i e j é obtuso porque $cos(\beta) < 0$. Existe assim, à partida, uma interação negativa entre estas variáveis nos k-ésimos níveis positivos de $\mathbf{C}$;
- **Caso 3**: Para $d_{ij}^r \approx 0$, temos $\langle \mathbf{A}_i^X, \mathbf{B}_j^Y \rangle \approx 0$, logo o ângulo γ formado pelos vetores das variáveis i e j é aproximadamente recto dado que $cos(\pi/2) = 0$. Não existe assim, à partida, interação entre estas variáveis.

A nível visual, com os biplots associados a cada componente do terceiro modo, estas proximidades analisam-se através da projeção das variáveis do segundo modo sobre a direção dos vetores das variáveis do primeiro modo. Consideremos as variáveis var_{j_1} e var_{j_2} pertencentes ao segundo modo (representadas por pontos) e a variável var_i do primeiro (representada por uma seta e esquematizada na Figura 2.4). Tal como a figura apresenta, os pontos P_1 e P_2, obtidos através da projeção dos pontos var_{j_1} e var_{j_2} sobre a direção do vetor var_i, mostram que var_{j_1} e var_{j_2} apresentam uma interação positiva com var_i, uma vez que encontram-se projetados no mesmo sentido da direção da seta, isto é, o ângulo formado entre var_{j_1} e var_{j_2} com var_i é agudo, inferior a 90 graus. A interação será maior entre var_{j_1} com var_i dado que P_1 encontra-se mais afastado da origem. Em suma, tudo o que foi explicado neste parágrafo está em conformidade com o caso 1 estabelecido anteriormente.

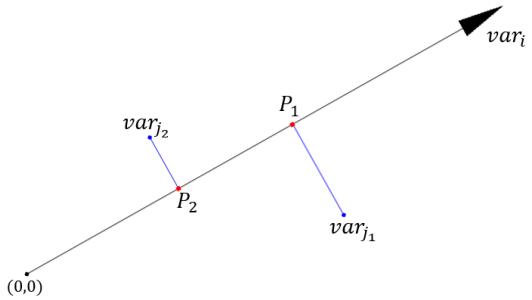

Figura 2.4: Interação positiva entre variáveis

Em relação ao segundo ponto, a interpretação é idêntica, com a exceção que neste caso P_1 e P_2 encontram-se projetados no sentido oposto à diretação da seta, isto é, o ângulo formado entre var_{j_1} e var_{j_2} com var_i é obtuso, superior a 90 graus. Neste sentido, a interação é negativa entre as variáveis em estudo.

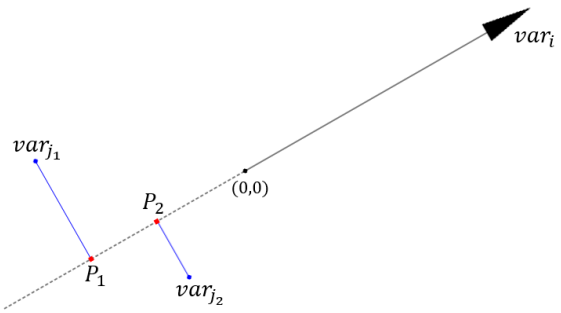

Figura 2.5: Interação negativa entre variáveis

No último ponto, os ângulos entre as variáveis são rectos, resultando em $P_1 = P_2 = (0, 0)$. Desta forma, não existe qualquer tipo de interação.

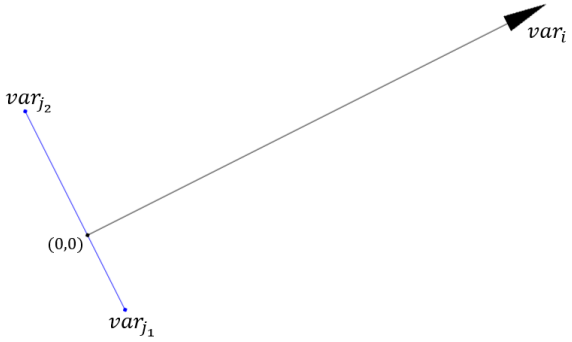

Figura 2.6: Interação nula entre variáveis

Exemplo 2.5.2 *Considerando o exemplo anterior (2.5.1), vamos analisar o biplot resultante, apresentado na Figura 2.7, tendo em conta os quatro sinais dos níveis da matriz* $\boldsymbol{C}$*:*

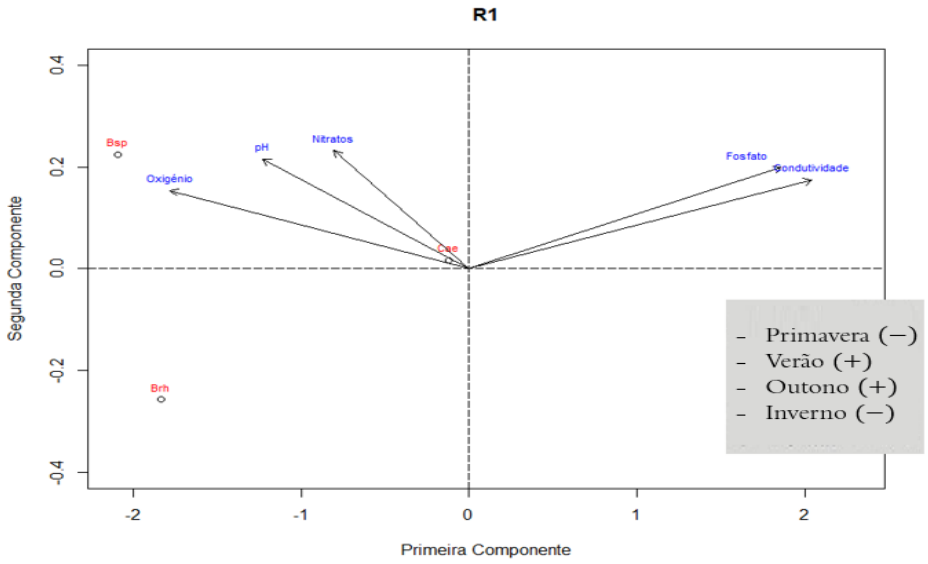

Figura 2.7: Biplot com a projeção das componentes do primeiro e segundo modo na primeira componente do terceiro modo.

Podemos observar que, cada nível do primeiro e segundo modo está projetado na primeira componente R_1 *do terceiro modo, onde os níveis do primeiro modo se encontram representadas por setas e os níveis do segundo por pontos. Importante relembrar que a representação escolhida é ao critério do investigador, no entanto todas estas variáveis são vistas como setas na medida de termos interpretações idênticas a um biplot clássico.*

As espécies Bsp e Brh encontram-se projetadas no mesmo sentido da orientação do oxigénio, Ph e Nitratos. Concluímos assim que existe uma interação positiva entre estas variáveis nas condições positivas em C (verão e outono) e interação negativa nas restantes condições (primavera e inverno). As mesmas espécies encontram-se no sentido oposto à orientação do sentido das setas de condutividade e fosfato. Desta forma, apresentam uma interação negativa nos níveis positivos de $\boldsymbol{C}$ *e uma interação positiva nos níveis negativos. A espécie Cae encontra-se situada relativamente perto da origem, pelo que se conclui que não apresenta*

interação.

Recorrendo à expressão 2.19 e à Tabela 2.2, os valores mais consideráveis são relativos às distâncias entre Bsp e oxigénio ($+3.759$)*, entre Bsp e condutividade* (-4.226) *e entre Bsp e fosfato* (-3.844)*, como já era esperado pela visualização do biplot.*

Capítulo 3

Aplicação de modelos Co-Tucker

3.1 Introdução

Tal como referido no Capítulo 1, iremos proceder à realização de dois modelos exploratórios sobre os dados utilizados: o primeiro, com o objetivo de estudar a interação entre os metabolitos da saliva e urina de grávidas, e o segundo modelo, no sentido de analisar esta mesma interação entre espaços temporais, possibilitando assim termos uma noção da trajetória do comportamento metabolóide dos dois biofluidos no tempo. Neste sentido, este capítulo responsabiliza-se por mostrar todos os resultados e interpretações obtidas. O tratamento de dados, a análise da co-inércia, a construção e o número de componentes escolhidas em cada modelo, o cálculo da variância explicada e as interpretações numéricas e gráficas estão em consonância com os procedimentos e definições descritos e detalhados no Capítulo 2.

3.2 Identificação de metabolitos

Como já referido no capítulo 1, os dados em estudo são representados em dois cubos de dados. O primeiro cubo $\mathbb{X}$, de dimensão $(7 \times 4 \times 3)$, e o segundo cubo $\mathbb{Y}$, de dimensão $(7 \times 24 \times 3)$. As 4 variáveis de $\mathbb{X}$ e 24 variáveis de $\mathbb{Y}$ correspondem a valores espectrais de cada metabolito. A Tabela 3.1 indica os metabolitos salivares. Alguns metabolitos, pelo seu extenso nome, foram abreviados.

Variáveis de $\mathbb{X}$	
Metabolitos	Identificação
NAG	Grupo N-acetilo das glicoproteínas.
Acetoína	
Ureia	
Etanol	

Tabela 3.1: Identificação das variáveis do cubo $\mathbb{X}$.

Também a escrita de alguns dos 24 metabolitos de $\mathbb{Y}$ foi abreviada.

Variáveis de $\mathbb{Y}$	
Metabolitos	Identificação
Pn3G	Metabolito da progesterona (provavelmente alopregnanolona e isómeros).
P3G	5β-pregnane-3α,20α-dio-3α-glucuronido.
X4.DEA	4-DEA, ácido 4-desoxieritrónico.
X3.HIBA	3-HIBA, 3-hidroxi-isobutirato.
Alanina	
X4.DTA	4-DTA, ácido 4-desoxitreónico.
Colina	
Carnitina	
Glicina	
Creatina	
Creatinina	
Glicose	
Hipurico	Ácido Hipúrico.
X1.6.anidro	1.6-anidroglucose.
Treonina	
Trigonelina	
Histidina	
Taurina	
X2.KG	2-KG, 2-cetoglutarato.
U3	2.95 ppm (singuleto).
U1	1.26 ppm (singuleto).
N5AC	N-acetilneuraminato.
U2	2.94 ppm (singuleto).

Tabela 3.2: Identificação das variáveis do cubo $\mathbb{X}$.

De forma a facilitar a leitura, a identificação dos metabolitos usada na primeira coluna das Tabelas 3.1 e 3.2, vai ser utilizada em diante.

3.3 Modelo Co-Tucker

3.3.1 Tratamento de dados

A interpretação dos resultados obtidos com a aplicacão deste modelo está dependente do tipo de dados que utilizamos. Dados centrados (Definição 3.3.1) conduzem a resultados que dados normalizados (Definição 3.3.2) não conseguem dar. Por outro lado, a normalização dos mesmos proporcionará outro tipo de informação que, de outra forma, não conseguiríamos obter. Resta assim estudar, de uma forma mais detalhada, a natureza destes dados, e, a partir daí, executar o tratamento dos mesmos com o intuíto de obter resultados mais interpretáveis.

Definição 3.3.1 *Seja $x_1, x_2, ..., x_n$ uma amostra de valores de uma variável aleatória X. Se a cada valor $x_1, x_2, ..., x_n$ subtrairmos o valor médio da amostra, obtemos os valores observados de uma variável Z centrada. A função no R que permite centrar os dados X é scale(X,center=TRUE,scale=FALSE).*

Definição 3.3.2 *Seja $x_1, x_2, ..., x_n$ uma amostra de valores de uma variável aleatória X. Se a cada valor $x_1, x_2, ..., x_n$ subtrairmos o valor médio da amostra e dividirmos pelo respetivo*

desvio padrão, obtemos a variável Z normalizada. A função no R é scale(X, center=TRUE, scale=TRUE).

Nesta dissertação apenas estes dois tipos de transformações de dados são executados. No entanto, muitos outros propostos existem na literatura.

A interação entre metabolitos da saliva e da urina é medida através da execução das matrizes cruzadas na análise da co-inércia (2.1). Dados centrados poderão levar a uma interação diferente de dados normalizados. Primeiramente, vamos considerar tanto os dados salivares do cubo $\mathbb{X}$ e urinários do cubo $\mathbb{Y}$ como centrados. Relembrar que, com esta metodologia, as matrizes cruzadas correspondem às matrizes de covariância entre as variáveis de $\mathbb{X}$ e $\mathbb{Y}$. A Figura 3.1 compara graficamente os valores da variância dos metabolitos salivares e urinários em estudo em cada um dos trimestres.

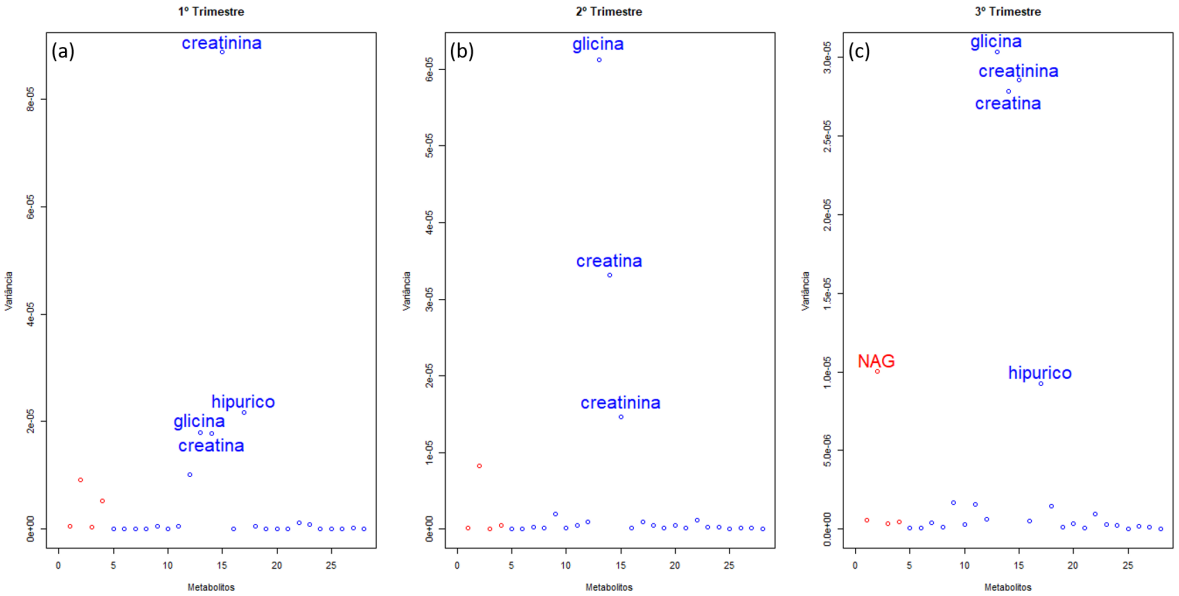

Figura 3.1: Variância dos metabolitos salivares, a vermelho, e dos metabolitos urinários, a azul, no primeiro (a), segundo (b) e terceiro (c) trimestres.

Embora os valores observados para as variâncias sejam baixos, em cada um dos trimestres, os metabolitos que mais se destacam, pela sua maior variância comparativamente aos restantes, pertencem à urina. A creatinina, a creatina e a glicina evidenciam variâncias maiores, principalmente no segundo (b) e terceiro (c) trimestres. O hipúrico também se sobressai no primeiro (a) e terceiro (c) trimestres. Do lado da saliva, o metabolito que se destaca é o NAG. Contudo, quando comparado aos anteriores mencionados, este apresenta uma variância inferior. Um possível problema de trabalhar com dados centrados é a grande evidência de variáveis que apresentam maiores variâncias, tal como ocorre na execução de uma ACP. Na análise de um modelo Co-Tucker, aquando da realização de um biplot, estas mesmas variáveis vão se distinguir, enquanto que as restantes localizar-se-ão próximas do centro. Em suma, a informação extraída pelo modelo será insuficiente, dado que se foca apenas em três ou quatro metabolitos. Desta forma, será recomendável homogeneizar os dados urinários através de uma normalização. Assim, as variáveis metabólicas pertencentes à urina vão sofrer uma normalização, enquanto que, no lado da saliva, as variáveis serão apenas centradas. Dado que temos

3 trimestres, os processos de centrar $\mathbb{X}$ e normalizar $\mathbb{Y}$ ocorrem três vezes. O primeiro sobre 3 matrizes (7×4) e o segundo processo sobre 3 matrizes (7×24).

3.3.2 Interações trimestrais com a análise da co-inércia

Depois de três matrizes (4×24) de covariâncias cruzadas (2.1) calculadas, é importante ter em consideração quais os trimestres que revelam uma maior interação entre os metabolitos salivares e urinários. A co-inércia foi calculada em cada trimestre a partir da expressão $CoIner_{X_kY_k}$ =traço$(X_kD_IX_k^TD_NY_kD_JY_k^TD_N)$, patente em 2.2.1, sendo D_I e D_J matrizes identidade de dimensão (4×4) e (24×24), respetivamente, e $k = 1, 2, 3$. A Figura 3.2 mostra os respetivos valores em cada trimestre.

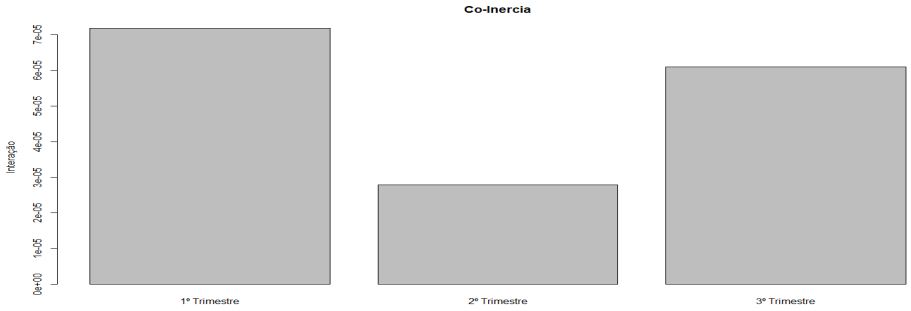

Figura 3.2: Co-inérica medida no 1º, 2º e 3º trimestres a partir de dados $\mathbb{X}$ centrados e $\mathbb{Y}$ normalizados.

No primeiro e terceiro trimestres os valores da co-inércia são superiores. Constata-se que, nestas condições temporais, as interações são mais fortes, tanto negativamente como positivamente (ver definição na secção 2.2.1). O contrário acontece no segundo trimestre, onde é apresentado um valor mais pequeno. Significa isto que, os valores presentes em X_2 centrado e Y_2 normalizado variam independentemente, ou, simplesmente, não variam. Tendo em conta estas informações, é de esperar que o modelo Co-Tucker dê mais relevância ao primeiro e terceiro trimestres.

Observação 3.3.3 *Para avaliar se a ordem dos trimestres esquematizada na Figura 3.2 se mantém, foram também calculados outros valores da co-inércia para outros tipos de tratamentos de dados. Neste sentido, a Figura 3.3 dispõe dessa mesma ordem.*

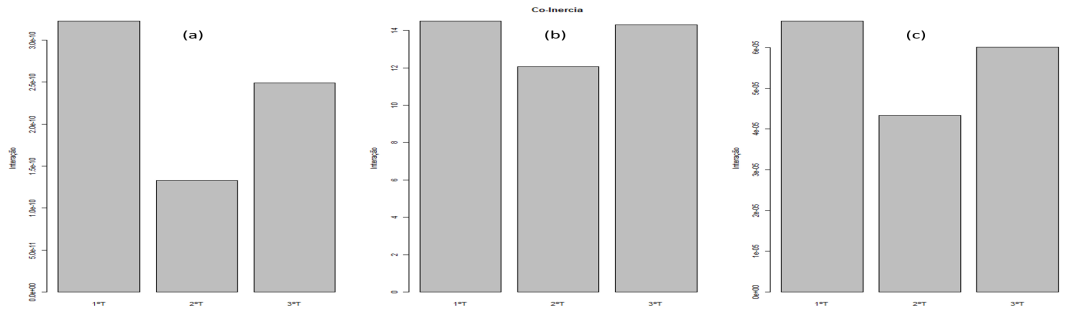

Figura 3.3: Co-inérica medida no 1º, 2º e 3º trimestres a partir de dados $\mathbb{X}$ e $\mathbb{Y}$ centrados (a), $\mathbb{X}$ e $\mathbb{Y}$ normalizados (b), e $\mathbb{X}$ normalizado e $\mathbb{Y}$ centrado (c).

Note-se que, com ambos os dados normalizados (b), cada barra corresponde ao quadrado da soma de correlações entre os 4 metabolitos salivares com os 24 metabolitos urinários nos

respetivos trimestres. Daí o motivo da escala vertical ter valores bastante superiores aos restantes. Em (a), ambos os dados são centrados, pelo que cada barra corresponde ao quadrado da soma das covariâncias. Os valores na escala são bastante reduzidos. Por último, em (c), a escala apresenta valores ligeiramente superiores aos anteriores devido à normalização do cubo $\mathbb{X}$*. Em suma, é claramente visível na Figura 3.3 que a ordem da interação entre trimestres é igual, independentemente do tipo de tratamento de dados que utilizamos. Este indicador revela que deveremos ter uma maior consideração naqueles dois trimestres.*

3.3.3 Escolha do modelo

Com a aplicação do método DifFit (2.4) obtemos a Tabela 3.3 com os detalhes de cada modelo. A programação deste método está presente em `https://github.com/Francisjcs1997/CT/blob/main/DifFit.R`.

N° Componentes	Complexidade	Variância	Diferencial	Rácio
3	1×1×1	0.5581	0.5581	2.402
5	1×2×2	0.7904	0.2323	3.971
6	2×2×2	0.8360	0.8360	< 1
7	2×3×2	0.8907	0.0547	< 1
8	2×3×3	0.9090	0.0183	< 1
9	2×4×3	0.9675	0.0585	6.359
10	3×4×3	0.9767	0.0092	1.195
11	3×5×3	0.9844	0.0077	1.351
12	3×6×3	0.9901	0.0057	1.727
13	3×7×3	0.9934	0.0033	1.179
14	4×7×3	0.9962	0.0028	1.167
15	4×8×3	0.9986	0.0024	3.429
16	4×9×3	0.9993	0.0007	1.400
17	4×10×3	0.9998	0.0005	2.500
18	4×11×3	1.000	0.0002	∞

Tabela 3.3: Critério DifFit.

Temos 15 modelos selecionados, onde dispomos, respetivamente, do número total de componentes de cada um deles, com $S = 3, 5, 6, ..., 18$, a sua complexidade, isto é, o número de componentes de cada modo, a sua proporção de variância explicada, $\mathfrak{V}_{\widehat{\mathbb{W}}_S}$, a diferença de proporção com o modelo anterior, e, por último, o rácio. O valor crítico $V_c = 1/16 = 0.0625$, pelo que excluímos a maioria dos modelos. Este fenómeno acontece uma vez que as componentes dos modelos mais simples captam a maioria da variância explicada, o que torna pequeno os valores da coluna diferencial da Tabela 3.3 a partir de $S = 6$ componentes. Note-se que as entradas na última coluna da Tabela 3.3 com "< 1"correspondem a modelos com rácio b_s inferior a 1, e também não serão à partida considerados.

Entre os modelos plausíveis de selecionar, com $dif_S > V_c$, temos o primeiro, com $S = 3$, constituído por uma componente em cada modo e com uma variância explicada de 55.81%. O segundo a ser considerado corresponde a um número de componentes $S = 5$, com complexidade $1 \times 2 \times 2$ e com uma variância explicada de 79.04%. Este último apresenta um maior rácio, com $b_5 = 3.971$. Um modelo com uma complexidade $1 \times 2 \times 2$ impossibilita a concretização de biplots, dado que temos apenas uma componente no primeiro modo, tornando impossível fazer uma representaccão bidimensional. Neste sentido, para contornar este problema e de forma

a enriquecer esta dissertação, o modelo escolhido corresponde ao modelo com um número de componentes $S = 6$, complexidade $2 \times 2 \times 2$, e com uma variância explicada de 83.60%. Assim, o modelo Co-Tucker é constituído por duas componentes no modo da saliva, P_1 e P_2, duas componentes no modo da urina, Q_1 e Q_2, e igualmente duas componentes no modo dos trimestres, R_1 e R_2.

3.3.4 Construção do modelo e as suas caraterísticas

Depois de determinar o número de componentes escolhidas para cada modo, a Tabela 3.4 apresenta os seus respetivos pesos.

Modo 1			Modo 2			Modo 3		
Saliva	P_1	P_2	Urina	Q_1	Q_2	Trim.	R_1	R_2
Etanol	−0.05	−0.34	Pn3G	−0.09	−0.16	1T	**0.64**	**-0.75**
NAG	**-0.99**	0.11	P3G	−0.13	−0.08	2T	−0.42	−0.13
Acetoina	−0.00	−0.10	X4.DEA	−0.21	0.14	3T	**0.65**	**0.65**
Ureia	0.09	**0.93**	X3.HIBA	0.14	0.18			
			Alanina	**0.35**	0.05			
			X4.DTA	−0.11	0.12			
			Colina	0.26	−0.17			
			Carnitina	0.14	0.14			
			Glicina	0.16	−0.09			
			Creatina	−0.01	−0.24			
			Creatinina	−0.26	0.06			
			Glicose	0.33	0.10			
			Hipurico	−0.07	0.26			
			GAA	0.03	**-0.49**			
			X1.6.anidro	−0.08	−0.17			
			Treonina	0.07	−0.35			
			Trigonelina	−0.17	−0.02			
			Histidina	0.17	0.07			
			Taurina	0.19	0.11			
			X2.KG	0.10	**0.50**			
			U3	0.28	−0.13			
			U1	−0.14	0.01			
			N5AC	**-0.31**	−0.14			
			U2	**0.41**	−0.12			

Tabela 3.4: Pesos das entradas das componentes do primeiro, segundo e terceiro modos.

Os pesos obtidos em cada componente de cado modo (2.3.2) através do método dos mínimos quadrados alternados foram conseguidos em apenas quatro iterações e estão expressos visualmente na Figura 3.4. A primeira inicialização com coeficientes das matrizes $\mathbf{B}^Y$ e $\mathbf{C}$ nulos foi a usada. As três diferentes inicializações para efetuar o método dos mínimos quadrados está patente em `https://github.com/Francisjcs1997/CT/blob/main/ALS.R`.

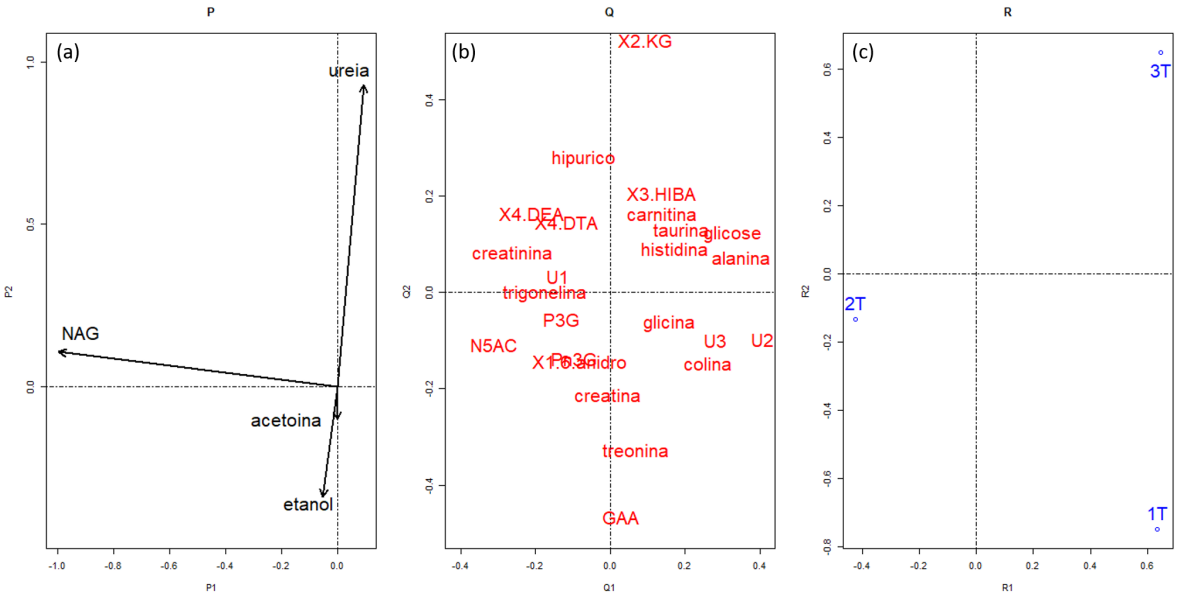

Figura 3.4: Coordenadas do primeiro modo (a), segundo modo (b) e terceiro modo (c).

No primeiro modo (a), relacionado com os metabolitos de saliva, podemos visualizar os seguintes resultados. NAG exibe um peso bastante negativo sob a componente P_1, enquanto que os restantes metabolitos apresentam valores próximos de zero. A segunda componente P_2 é claramente marcada pelo contraste entre a ureia, que apresenta um peso positivo, com o etanol e a acetoína, com cargas negativas mas de menor magnitude. A acetoína acaba por apresentar um valor bastante pequeno.

No segundo modo (b), relacionado com os metabolitos de urina, Q_1 revela uma oposição entre N5AC (seguido da creatinina e de X4.DEA) com U2 (seguido de alanina, glucose, colina e U3), estes últimos com cargas positivas. Quanto à componente Q_2, é também caracterizada por um contraste entre GAA (seguida de treonina) com X2.KG. Os restantes metabolitos encontram-se próximos da origem do gráfico.

No terceiro modo (c), a primeira componente R_1 apresenta um contraste entre o primeiro e terceiro trimestres (cargas positivas) com o segundo trimestre (carga negativa). Em R_2 o contraste acontece entre o primeiro e o terceiro trimestre, sendo que o segundo apresenta um peso muito próximo de nulo.

De forma a interpretar as relações entre os elementos dos diferentes modos, é-nos de seguida apresentado na Tabela 3.5 o cubo $\mathbb{G}_{2\times2\times2}$ (2.8) com as respetivas interações.

	$\mathbb{G}$			Proporção	
	P_1	P_2		P_1	P_2
			Componente R_1		
Q_1	-8.1e-08	3.4e-11		0.558	0.000
Q_2	7.4e-11	-1.7e-08		0.000	0.024
			Componente R_2		
Q_1	1.2e-10	1.8e-8		0.000	0.026
Q_2	5.2e-8	-3.8e-10		0.227	0.000

Tabela 3.5: Cubo $\mathbb{G}$

A partir da Tabela 3.5 conseguimos obter a Tabela 3.6 com a variância explicada de cada uma das componentes construída (2.10).

Modo		Soma	Proporção de Variância de cada componente	
1	Saliva ($P = 2$)	0.836	0.785	0.051
2	Urina ($Q = 2$)	0.836	0.585	0.251
3	Trimestres ($R = 2$)	0.836	0.583	0.253

Tabela 3.6: Proporção da variância explicada por cada uma das componentes construídas.

A projeção das componentes do modo da saliva e urina em cada componente do terceiro modo (R_1 e R_2) resulta nos gráficos da Figura 3.5.

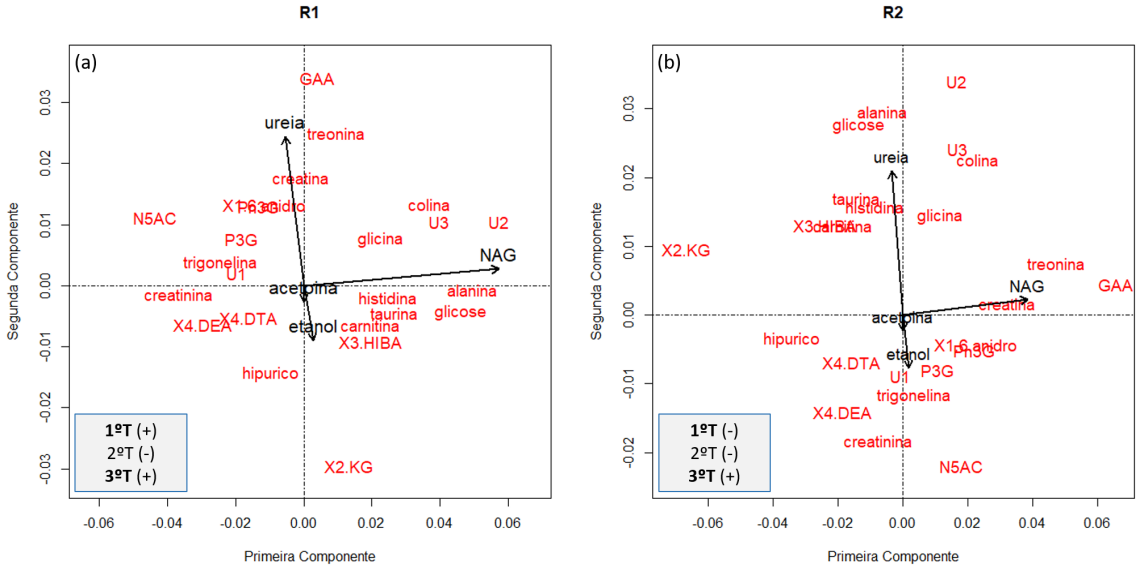

Figura 3.5: Projeção das componentes dos modos da saliva e urina em R_1 (a) e em R_2 (b).

Na Figura 3.5 (a) temos a primeira vs segunda componente do primeiro modo, com uma representação de 55.8% de variância explicada, e do segundo modo, contabilizando 2.4%. Em (b), temos também a primeira vs segunda componente do primeiro modo com 24.3% de variância explicada e do segundo modo com 1.1%.

Através da informação reunida na Tabela 3.4, no cubo $\mathbb{G}$ na Tabela 3.5, e nos gráficos anteriores, estamos em condições de enunciar as conclusões obtidas a partir do modelo Co-Tucker, de complexidade $2 \times 2 \times 2$, que são especificadas na próxima secção.

3.3.5 Interpretações

Através de $\mathbb{G}$, temos que a combinação de componentes que extrae uma maior variabilidade é referente às primeiras componentes de cada modo, com $g_{111} = (-)8.1e{-}08$ que explica 55.8% da variabilidade (Tabela 3.5). O metabolito da saliva tido em conta é apenas NAG. Enquanto que os pesos na componente P_1 dos restantes metabolitos são praticamente nulos,

este é considerável (-0.99). Desta forma, uma interação positiva ao longo de P_1, Q_1 e R_1 resulta das seguintes combinações (2.16):

- **Caso 1**: $P_1(-) \times Q_1(-) \times R_1(-)$;
- **Caso 2**: $P_1(-) \times Q_1(+) \times R_1(+)$.

Consequentemente, como resultado, observamos nas grávidas que os valores espectrais de NAG (que apresenta um alto número negativo em P_1) crescem quando o valor de N5AC (com alto número negativo em Q_1) cresce no segundo trimestre (valor negativo em R_1). Em sentido contrário, a interação destes mesmos metabolitos é negativa no primeiro e terceiro trimestre, dado que em R_1 apresentam valores positivos. Partindo agora da combinação 2, e observando as coordenadas positivas de Q_1, constatamos uma interação positiva de NAG com U2 e alanina no primeiro e terceiro trimestres e uma interação contrária no segundo. De uma forma geral, esta combinação de componentes responsabiliza-se por nos mostrar um contraste entre o segundo trimestre com o primeiro e terceiro através da interação entre NAG na saliva com N5AC, U2 e alanina na urina. Graficamente estas conclusões também estão patentes na Figura 3.5 (a), através da projeção em R_1, com uma variância explicada de 58.3%, onde temos a alanina, U2 e N5AC situados nos extremos do gráfico, tendo em consideração a direção do vetor NAG. Observa-se também que a glicose, colina e U3 posicionam-se relativamente perto da alanina e de U2, pelo que também terão o mesmo tipo de interação com NAG.

A próxima combinação a ser analisada corresponde ao valor $g_{122} = (+)5.2\text{e}{-08}$, com uma variância explicada de 22.7%, onde consideramos as componentes P_1, Q_2 e R_2. Novamente consideramos NAG como o metabolito representável de P_1. Assim, uma interação positiva é representada pelas seguintes combinações:

- **Caso 1**: $P_1(-) \times Q_2(-) \times R_2(+)$;
- **Caso 2**: $P_1(-) \times Q_2(+) \times R_2(-)$.

Por conseguinte, grávidas com altos valores de NAG (considerável peso negativo em P_1) apresentarão também elevados valores espectrais de GAA (peso mais negativo em Q_2) no terceiro trimestre (peso positivo em R_1). O contrário ocorre no primeiro trimestre com uma interação negativa, envolvendo estes mesmos metabolitos ($R_1(-) \times Q_2(-) \times R_2(-) \times (+) = (-)$). Ou seja, o aumento de NAG resulta numa diminuição dos valores de GAA. Considerando a combinação 2, teremos uma interação positiva de NAG com X2.KG no primeiro trimestre e negativa no terceiro ($P_1(-) \times Q_2(+) \times R_2(+) \times (+) = (-)$). Note-se que nesta combinação de componentes nunca evidenciamos o segundo trimestre. Deve-se isto ao facto de apresentar um peso próximo de zero na componente R_2 (-0.134). Na Figura 3.5 (b) facilmente se verifica o que acabou de ser exposto, através da projeção em R_2, com uma variabilidade explicada de 25.3%. GAA, seguida da treonina, apontam para uma proximidade com NAG, enquanto que X2.KG situa-se em posição oposta no biplot.

Temos, de seguida, a combinação de componentes P_2, Q_1 e R_2, com $g_{212} = (+)1.8\text{e}-08$ e representando uma variância explicada de 2.60%. Como já referido anteriormente, P_2 carateriza-se pelo contraste entre a ureia com o etanol e a acetoína. No entanto a ureia apresenta nesta componente um valor absoluto mais considerável (0.93), pelo que só iremos ter em conta este metabolito da saliva. Como efeito, as interações positivas decorrem da seguinte forma:

- **Caso 1**: $P_2(+) \times Q_1(+) \times R_2(+)$;

- **Caso 2**: $P_2(+) \times Q_1(-) \times R_2(-)$.

Deste modo, partindo da combinação 1, a ureia (peso positivo em P_2) apresenta uma interação positiva com U2 e alanina (peso positivo na componente Q_1) no terceiro trimestre (peso positivo em R_2). O contrário ocorrerá no primeiro (peso negativo em R_2), resultando numa interação negativa ($P_2(+) \times Q_1(+) \times R_2(-) \times (+) = (-)$). Tendo em conta a combinação 2, a ureia tem uma interação positiva com N5AC (valor negativo em Q_1) no primeiro trimeste, e negativa no terceiro ($P_2(+) \times Q_1(-) \times R_2(+) \times (+) = (-)$). No biplot (b) da Figura 3.5 observa-se também que a creatinina e X4.DEA posicionam-se relativamente próximos de N5AC, pelo que também apresentarão o mesmo tipo de interação com a ureia.

Por último, com $g_{221} = (-)1.7\text{e}{-08}$, representando uma variância explicada de 2.40%, temos a combinação de componentes P_2, Q_2 e R_1. Novamente, temos em consideração o metabolito ureia com um peso positivo em P_2. Com o coeficiente g_{221} negativo, as combinações com interação positiva vão ser da forma:

- **Caso 1**: $P_2(+) \times Q_2(+) \times R_1(-)$;
- **Caso 2**: $P_2(+) \times Q_2(-) \times R_1(+)$.

Assim sendo, grávidas com maiores valores de ureia apresentarão também maiores valores de X2.KG (valor positivo mais considerável em Q_2) no segundo trimestre. No primeiro e terceiro trimestres esta interação será negativa ($P_2(+) \times Q_2(+) \times R_1(+) \times (-) = (-)$). Na combinação 2, temos o GAA (coeficiente negativo em Q_2), que apresenta uma interação positiva com a ureia no primeiro e terceiro trimestres. No segundo trimestre, como esperado, o resultado será contrário ($P_2(+) \times Q_2(-) \times R_1(-) \times (-) = (-)$).

Nas restantes combinações de componentes a variância explicada por estas é irrelevante, dado que temos percentagens bastante reduzidas. Nesse sentido, não é apresentada mais nenhuma combinação.

Em suma, a projeção na primeira componente R_1, revela um contraste entre o primeiro e terceiro trimestres com o segundo. Este contraste dá-se através de duas interações. A primeira, que envolve, no lado da saliva, o metabolito NAG, e, do lado da urina, os metabolitos U2, alanina e N5AC e a segunda, a ureia com o GAA e X2.KG, metabolitos da urina. Também a projeção na segunda componente R_2 evidencia um contraste, agora entre o primeiro e o terceiro trimestre. Novamente os metabolitos em evidência são os mesmos de anteriormente. Informação relativa às interações que envolvem a acetoína e o etanol são mais escassas. Por esse motivo as setas relativas a estes metabolitos são também mais pequenas na Figura 3.5. Semelhante conclusão temos acerca dos metabolitos urinários não mencionados, dado que estão mais próximos do centro do biplot.

A construção deste modelo e todas as interpretações realizadas podem ser consultadas em `https://github.com/Francisjcs1997/CT/blob/main/CoTucker.R`.

Observação 3.3.4 *Posteriormente, foram também realizados mais 3 modelos Co-Tucker de complexidade* $2 \times 2 \times 2$ *com diferentes tratamento de dados. A Figura 3.6 apresenta a projeção em* R_1 *em cada um deles.*

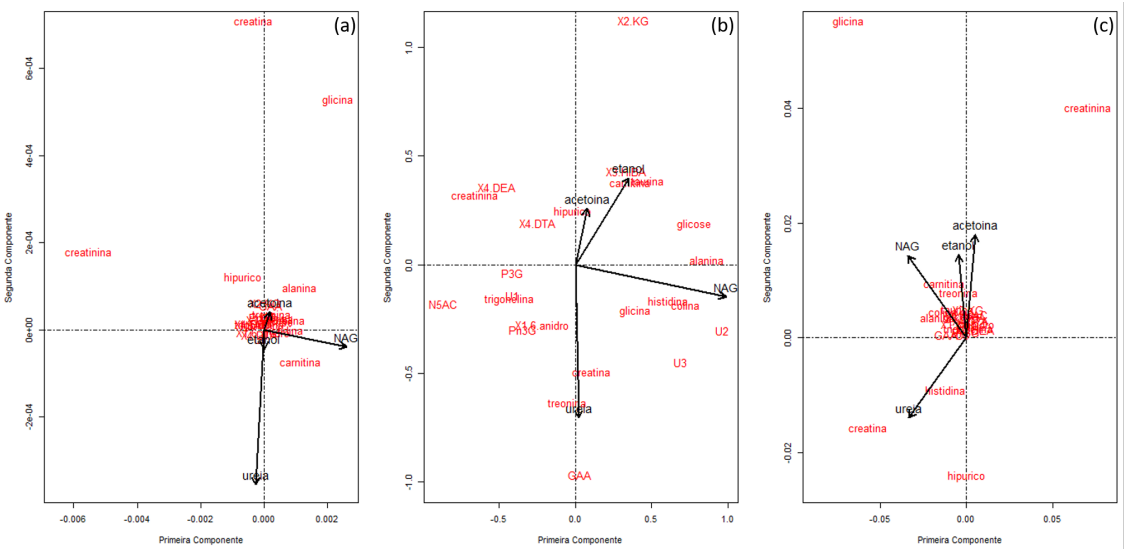

Figura 3.6: Projeção na primeira componente do terceiro modo R_1, com dados $\mathbb{X}$ e $\mathbb{Y}$ centrados (a), $\mathbb{X}$ e $\mathbb{Y}$ normalizados (b), e $\mathbb{X}$ normalizado e $\mathbb{Y}$ centrado (c).

Como esperado, a não homogeneização dos dados $\mathbb{Y}$ resulta numa evidência clara dos metabolitos que apresentam maior variância, sendo eles, a creatinina, creatina e glicina, tal como ilustrado na Figura 3.6 (a) e (c). As interpretações retiradas do modelo (b), com ambos os dados normalizados e com uma variância explicada de 52.9%, são extremamente semelhantes ao modelo principal, isto é, com $\mathbb{X}$ centrado e $\mathbb{Y}$ normalizado. As principais diferenças entre estes modelos situam-se no cubo $\mathbb{G}$. Enquanto que o modelo principal foca-se nas combinações mais importantes, o modelo (b) tenta dar relevo a todas, ficando mais sujeito a fenómenos de ruído. Importante também referir que o modelo principal apresenta uma variância explicada relativamente superior ao modelo em (b), uma vez que apenas $\mathbb{Y}$ sofre normalização.

3.3.6 Análise de correlações

Esta secção responsabiliza-se por nos exemplificar os constrastes anteriormente estudados, através da correlação de Spearman entre as variáveis salivares e urinárias, nos três diferentes espaços temporais. Este tipo de correlação foi efetuado no sentido de analisar as ordens espectrais de cada uma das grávidas em estudo. Assim, com o teste de correlação de Spearman e com um grau de confiança $\alpha = 0.05$, vamos efetuar o seguinte teste de hipóteses:

$$H_0 : \rho_{xy} = 0 \text{ vs } H_1 : \rho_{xy} \neq 0 \tag{3.1}$$

Onde x é uma variável salivar e y uma variável urinária. A função no R responsável por este teste é *cor.test(x,y,method="spearman")*.

As interações mais fortes relacionam-se com o metabolito salivar NAG. Determinadas através da distância (2.19), os metabolitos urinários que apresentam uma maior interação são, ordenadamente, U2, alanina, glucose, N5AC, U3 e a colina.

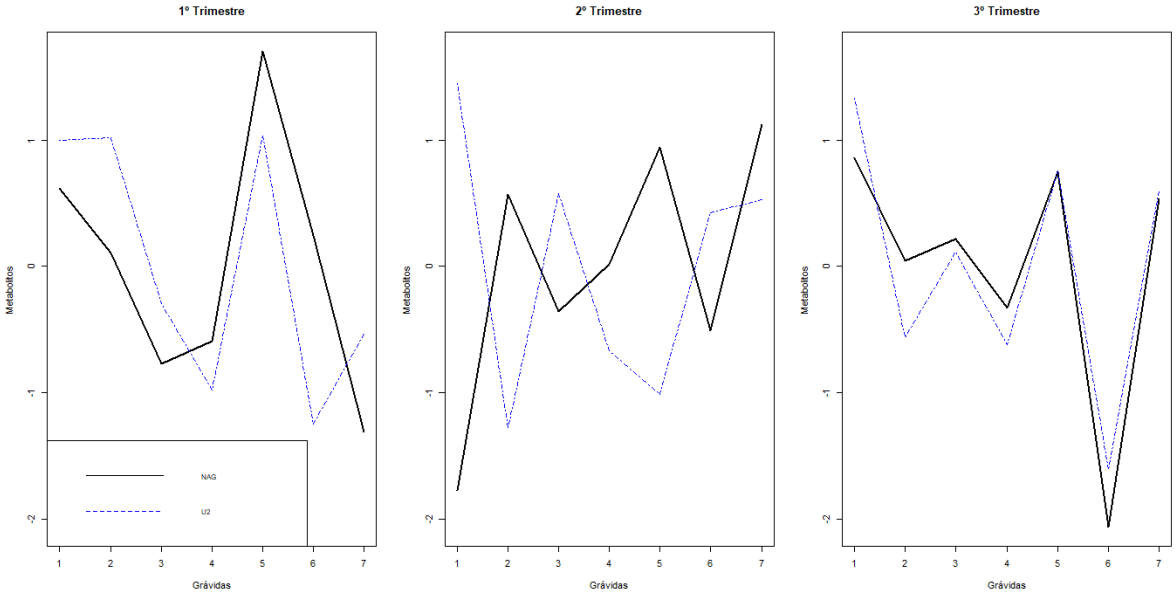

Figura 3.7: Espectro do NMR de NAG, a preto, e U2, a azul, em cada grávida e ao longo dos 3 trimestres

Por exemplo, na Figura 3.7, é claramente visível o contraste entre o primeiro e terceiro trimestres (interação positiva) com o segundo (interação negativa), envolvendo NAG e U2. Ao nível da projeção em R_2, representativa do contraste entre o primeiro e terceiro trimestre, GAA e X2.KG são os metabolitos pertencentes à urina que mais se destacam.

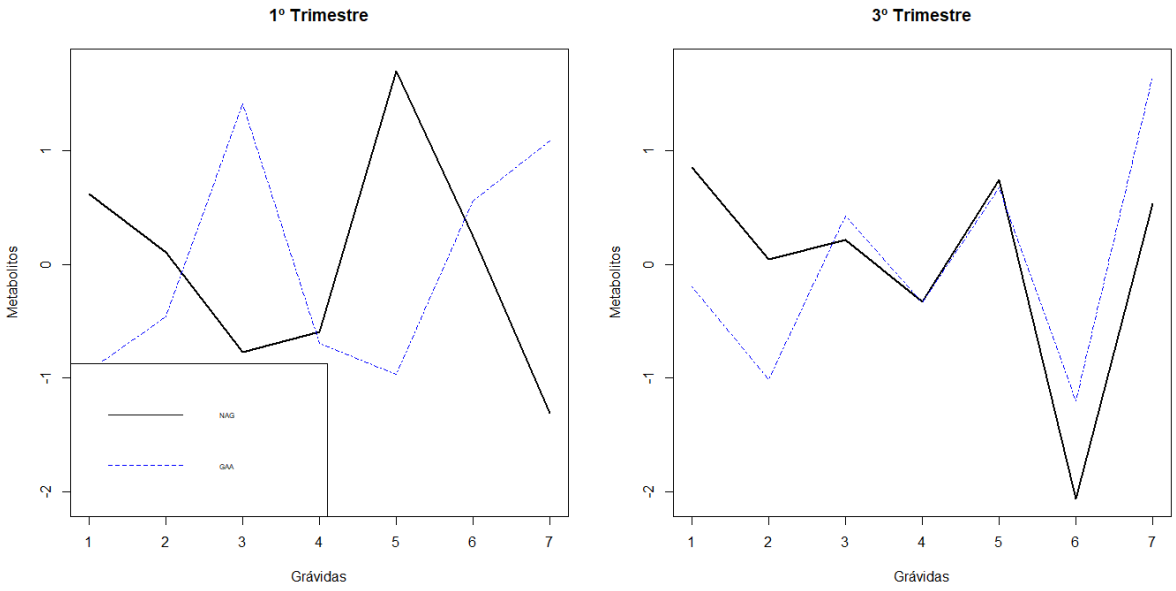

Figura 3.8: Espectro do NMR de NAG, a preto, e GAA, a azul, em cada grávida no primeiro e terceiro trimestres.

O contraste entre o primeiro e terceiro trimestres na Figura 3.8 é bastante visível na interação entre NAG com GAA. O segundo trimestre não é mostrado dado o reduzido peso na componente R_2.

De seguinda, a Tabela 3.7 apresenta os metabolitos cujo realização do teste (3.1) assinalou um valor-p $\alpha < 0.05$, isto é, testes onde a hipótese nula é rejeitada ($\rho_{xy} \neq 0$).

NAG					
1º Trimestre			3º Trimestre		
Urina	Correlação	Valor-p	Urina	Correlação	Valor-p
Glucose	0.8571	0.0238	U2	1.000	0.0000
X2.KG	0.8571	0.0238	Colina	0.9643	0.0028
GAA	-0.8214	0.0238	Alanina	0.8214	0.0341
			U3	0.7857	0.0480

Tabela 3.7: Correlação de Spearman de NAG com os metabolitos urinários

Em relação à primeira interpretação, podemos afirmar que, usando os níveis usuais de significância de $\alpha = 0.05$, a correlação de Spearman entre NAG e glucose, no primeiro trimestre é diferente de zero. O mesmo se concluí para NAG com U2, colina, alanina e U3, no terceiro trimestre. X2.KG e GAA, responsáveis pelo contraste entre o primeiro e terceiro trimestres, apresentam um valor-p inferior a 0.05 apenas no primeiro. Note-se que nenhum destes metabolitos aparecem em mais que um trimestre, apesar de serem os responsáveis pelas maiores interações. Importante referir também que no segundo trimestre nenhuma hipótese nula H_0 é rejeitada. Esta evidência sustenta o facto de na análise da co-inércia este trimestre apresentar menos interações (Figura 3.2), e também o facto de, tanto R_1 como R_2, apresentarem reduzidos coeficientes no segundo trimestre (Tabela 3.4).

Estudando agora as interações com a ureia, tal como a Tabela 3.5 indica, vamos ter uma proporção de variância explicada muito mais reduzida, pelo que as interações extraídas não são tão claras como as anteriores. A ureia apresenta uma maior interação com X2.KG (Figura 3.9) e GAA relativamente à projeção em R_1.

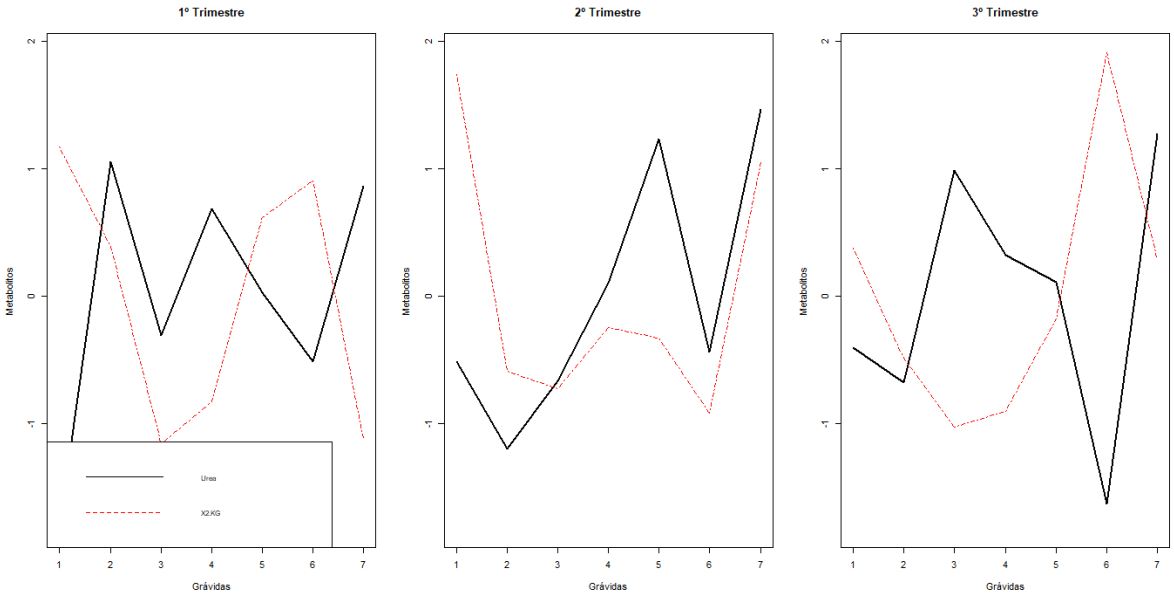

Figura 3.9: Espectro do NMR de ureia, a preto, e X2.KG, a vermelho, em cada grávida em cada um dos trimestres.

Quanto à projeção em R_2, o metabolito que apresenta maior interação com a ureia é o U2 (Figura 3.10).

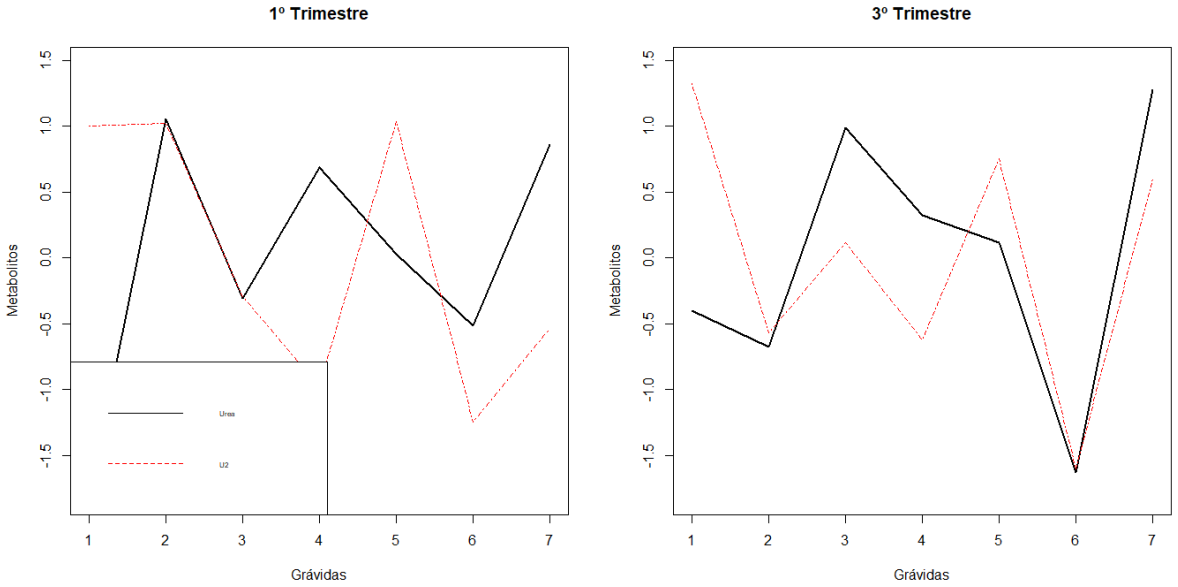

Figura 3.10: Espectro do NMR de ureia, a preto, e U2, a vermelho, em cada grávida no primeiro e terceiro trimestres.

Tal como esperado, este constraste não é tão nítido. A Tabela 3.8 mostra os metabolitos urinários cuja hipótese nula foi rejeitada.

Ureia		
3º Trimestre		
Urina	Correlação	Valor-p
GAA	0.8214	0.0341
Treonina	0.8214	0.0341

Tabela 3.8: Correlação de Spearman de ureia com os metabolitos urinários com um valor-p$< \alpha = 0.05$

GAA apresenta apenas relevância no terceiro trimestre. A treonina, apesar de não aparecer em nenhuma das anteriores interpretações, mostra-se aqui também com uma elevada correlação de Spearman neste mesmo trimestre.

3.3.7 Conclusões

O modelo Co-Tucker realizado focou-se essencialmente em interações que envolviam NAG e a ureia, havendo um maioritário foco para a primeira. Nos restantes metabolitos salivares, o etanol e a acetoína, o modelo não conseguiu exprimir informações, muito também devido à reduzida complexidade do mesmo. Claramente que, a maior informação a reter, prende-se com a estrutura entre NAG, da lado da saliva, com U2, alanina, glucose e N5AC, do lado da urina, responsáveis pelas interações opostas entre o primeiro e terceiro trimestres com o segundo. Também envolvendo o NAG, e outra estrutura a ter em consideração, é o contraste entre o primeiro e terceiro trimestre, com X2.KG e GAA como metabolitos urinários. As restantes

interpretações, apesar de também válidas, estão muito mais sujeitas ao ruído provocado pela redução de dimensionalidade.

Em relação ao metabolito salivar NAG, sabemos que as glicoproteínas são proteínas com glicanos ligados à cadeia lateral de aminoácidos. Isto poderia explicar a interação positiva com a glicose no 1º e 3º trimestres da gravidez. Por sua vez, a alanina e a histidina são aminoácidos gluconeogénico, que podem ser convertidos em glicose. Note-se que estes dois metabolitos, também no 1º e 3º trimestres, têm com uma interação positiva com NAG. Noutro sentido, o N5AC é o ácido siálico mais comum na urina, podendo estar relacionado com o NAG salivar. Focando na ureia, a forte interação com o GAA pode ser explicada por uma desregulação do ciclo da ureia ao longo da gravidez. O X2.KG é um intermediário do ciclo do Krebs e a interação com a ureia não conseguimos explicar neste momento. O modelo deu maior relevância ao primeiro e segundo trimestres, pelo que as interações basearam-se, principalmente, nestes dois eventos.

3.4 Modelo Co-Tucker diferencial

3.4.1 Tratamento de dados

Enquanto que anteriormente, no modelo Co-Tucker executado, as três condições do terceiro modo são relativas aos três trimestres, no modelo Co-Tucker que agora propomos analisar e que designamos por "diferencial", partiremos de cubos de dados com a diferença espectral de metabolitos entre os mesmos trimestres. Neste sentido, os cubos de dados $\mathbb{X}^\star$ e $\mathbb{Y}^\star$ a serem estudados são calculados da seguinte forma:

- **Condição 1** - $X_1^\star = X_2 - X_1$ e $Y_1^\star = Y_2 - Y_1$, que representam, respetivamente, a diferença dos espectros de metabolitos salivares e urinários, do primeiro para o segundo trimestre;
- **Condição 2** - $X_2^\star = X_3 - X_1$ e $Y_2^\star = Y_3 - Y_1$, que representam, respetivamente, a diferença dos espectros de metabolitos salivares e urinários, do primeiro para o terceiro trimestre;
- **Condição 3** - $X_3^\star = X_3 - X_2$ e $Y_3^\star = Y_3 - Y_2$, que representam, respetivamente, a diferença dos espectros de metabolitos salivares e urinários, do segundo para o terceiro trimestre;

Assim, na matriz $\mathbf{C}$ do terceiro modo, cada nível corresponde a um espectral diferencial entre trimestres.

Os cubos $\mathbb{X}^\star$ e $\mathbb{Y}^\star$ foram centrados. Tal como a Figura 3.11 indica, a creatinina, a creatina e a glicina, pelas suas elevadas variâncias, são as que mais se destacam do lado da urina. O hipúrico também se destaca em (b). Neste sentido, e tal como anteriormente, é fundamental uniformizar $\mathbb{Y}^\star$ através de uma normalização.

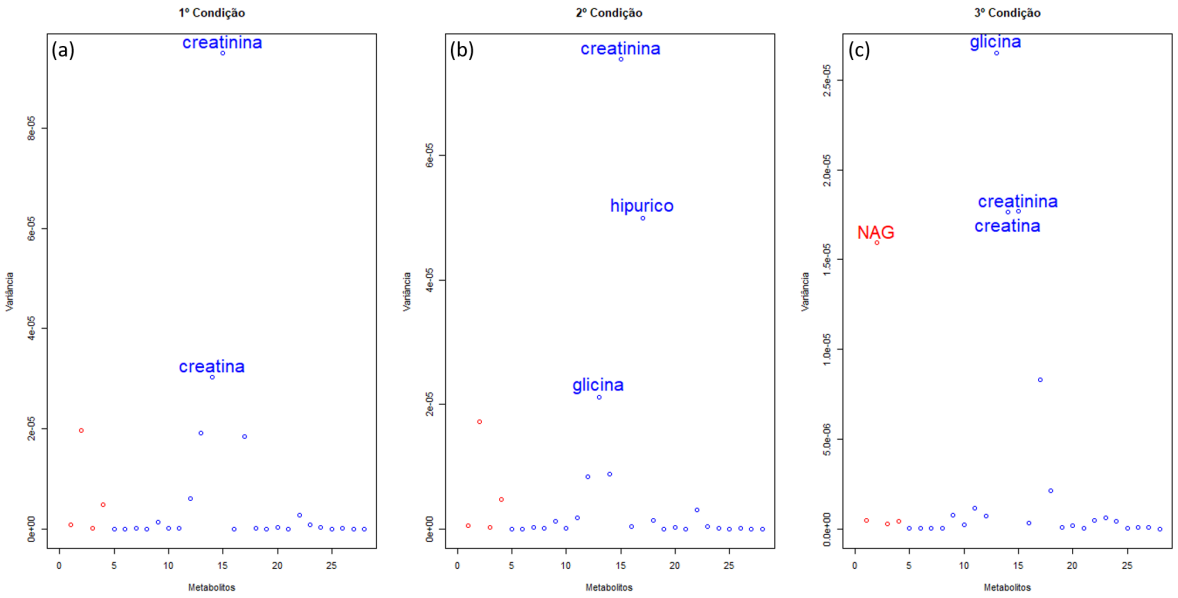

Figura 3.11: Variância dos metabolitos salivares, a vermelho, e dos metabolitos urinários, a azul, na primeira (a), segunda (b) e terceira (c) condição.

Enquanto que, no modelo anterior, centrar ou normalizar $\mathbb{X}$ era irrelevante, no modelo Co-Tucker diferencial o mesmo não acontece. Com $\mathbb{X}^\star$ centrado, apenas os metabolitos NAG e ureia se destacam. Com $\mathbb{X}^\star$ também uniformizado, conseguimos obter destaque em relação ao etanol e à acetoína. Uma vez que anteriormente não conseguimos retirar informações destes metabolitos, torna-se apreciável trabalhar com ambos os cubos $\mathbb{X}^\star$ e $\mathbb{Y}^\star$ normalizados, dado o propósito de extraír mais informações. Novamente, estes dois processos ocorrem três vezes. O primeiro, sobre três matrizes de dimensão (7×4), e o segundo, sobre três matrizes de dimensão (7×24).

3.4.2 Interações entre trimestres com a análise da co-inércia

A Figura 3.12 ilustra os valores da co-inércia em cada uma das condições em estudo.

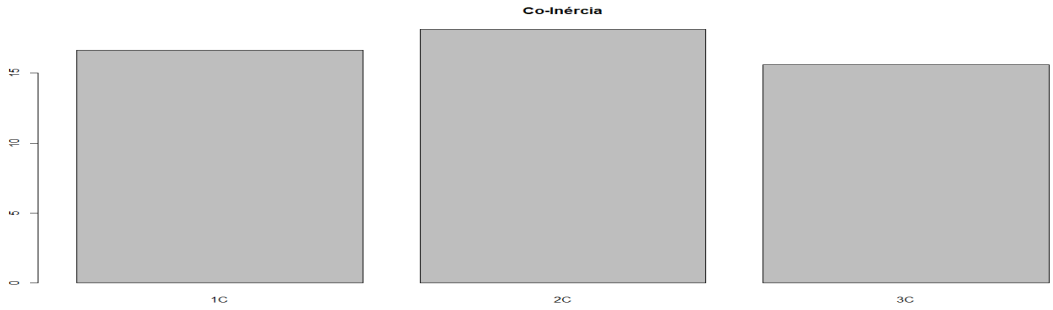

Figura 3.12: Co-inérica medida na 1º, 2º e 3º condição a partir de dados $\mathbb{X}^\star$ e $\mathbb{Y}^\star$ estandardizados.

Uma vez que os dados encontram-se estandardizados, cada barra da Figura 3.12 representa a soma do quadrado das correlações. A segunda condição apresenta um valor superior, o que revela que existem maiores interações na passagem do primeiro para o terceiro trimestre. No entanto, todas as condições apresentam valores de co-inércia semelhantes.

Observação 3.4.1 *Foi calculada a co-inércia para outros tipos de tratamento de dados. A Figura 3.13 apresenta esses valores em cada uma das condições em estudo.*

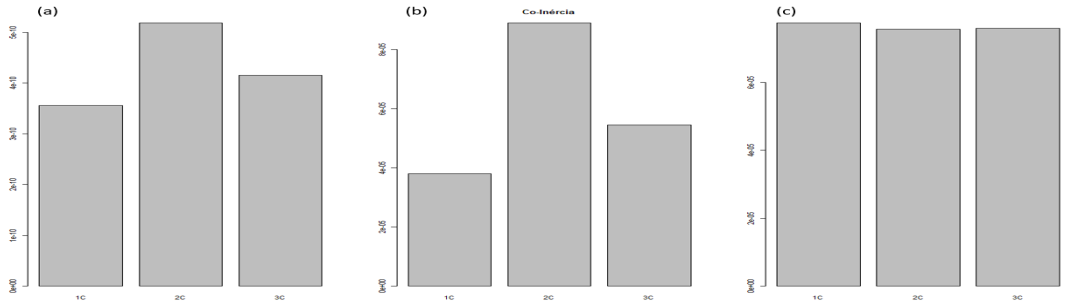

Figura 3.13: Co-inérica medida na 1º, 2º e 3º condições a partir de dados $\mathbb{X}^\star$ e $\mathbb{Y}^\star$ centrados (a), $\mathbb{X}^\star$ normalizado e $\mathbb{Y}^\star$ centrado (b), e $\mathbb{X}^\star$ centrado e $\mathbb{Y}^\star$ normalizado (c).

Apenas no gráfico (c) a segunda condição não se evidencia das restantes. A normalização de $\mathbb{Y}^\star$ é responsável por esse fenómeno, pois abafa os elevados valores presentes em $Y_2^\star$ centrado. Tal como presente na Figura 3.12, o modelo Co-Tucker diferencial e os gráficos (a) e (b) mostram maior relevância na segunda condição. No entanto, estes últimos apresentam maior descrepância com a primeira e terceira, devido à não homogeneização de $\mathbb{X}^\star$ e/ou $\mathbb{Y}^\star$

3.4.3 Escolha do modelo diferencial

Mais uma vez, o método DifFit (2.4) foi responsável pelo escolha da complexidade do modelo Co-Tucker. A Tabela 3.9 apresenta as informações de cada modelo.

N° Componentes	Complexidade	Variância	Diferencial	Rácio
3	1×1×1	0.2860	0.2860	2.058
5	2×2×1	0.4253	0.1393	1.275
6	2×2×2	0.4630	0.0377	< 1
7	2×3×2	0.5718	0.1088	1.216
8	3×3×2	0.6590	0.0872	< 1
9	3×4×2	0.7424	0.0834	< 1
10	3×4×3	0.8319	0.0895	1.823
11	3×5×3	0.8810	0.0491	1.256
12	4×5×3	0.9201	0.0391	1.164
13	4×6×3	0.9537	0.0336	1.577
14	4×7×3	0.9750	0.0213	1.420
15	4×8×3	0.9900	0.0150	1.875
16	4×9×3	0.9980	0.0080	5.714
17	4×10×3	0.9994	0.0014	2.800
18	4×11×3	0.9999	0.0005	5.000
19	4×12×3	1.000	0.0001	< 1

Tabela 3.9: Critério DifFit aplicado ao modelo Co-Tucker diferencial.

Tal como anteriormente, a escolha recai sobre modelos com um número reduzido de componentes. Entre os modelos possíveis de selecionar, com $Dif_s > V_c = 1/16 = 0.0625$, é escolhido o primeiro, com $S = 3$ componentes e com um rácio de 2.058. O problema mantém-se, uma vez que não existe possibilidade de fazer uma interpretação gráfica, ao nível de biplots, dado

o reduzido número de componentes em cada modo. Assim, foi escolhido o modelo com cinco componentes, complexidade $2 \times 2 \times 1$, e com uma variância explicada de 42.53%. O modelo Co-Tucker diferencial é constituído por duas componentes P_1 e P_2 do modo salivar, duas componentes Q_1 e Q_2 do modo urinário e por uma componente R_1 no modo das condições.

3.4.4 Construção do modelo diferencial

A Tabela 3.10 expõe as matrizes de cada modo com os pesos das componentes, destacando a negrito os pesos com maiores magnitudes.

Modo 1			Modo 2			Modo 3	
Saliva	P_1	P_2	Urina	Q_1	Q_2	Cond.	R_1
Etanol	−0.24	−0.36	Pn3G	−0.01	−0.20	1C	**0.58**
NAG	−0.09	**-0.91**	P3G	0.06	−0.27	2C	**0.73**
Acetoina	**-0.68**	0.19	X4.DEA	0.06	**-0.35**	3C	0.36
Ureia	**0.69**	−0.06	X3.HIBA	−0.18	−0.20		
			Alanina	−0.20	−0.02		
			X4.DTA	**0.45**	−0.03		
			Colina	0.03	0.28		
			Carnitina	0.28	−0.01		
			Glicina	0.27	0.06		
			Creatina	0.02	0.07		
			Creatinina	0.04	**-0.35**		
			Glicose	−0.07	0.14		
			Hipurico	0.05	−0.23		
			GAA	−0.01	0.04		
			X1.6.anidro	**0.32**	0.28		
			Treonina	−0.20	0.16		
			Trigonelina	0.06	0.14		
			Histidina	−0.13	0.12		
			Taurina	**0.33**	0.11		
			X2.KG	0.07	−0.02		
			U3	**-0.32**	0.19		
			U1	**0.41**	0.00		
			N5AC	−0.07	**-0.40**		
			U2	−0.11	**0.31**		

Tabela 3.10: Pesos das entradas das componentes do primeiro, segundo e terceiro modo do modelo Co-Tucker diferencial.

Todos estes pesos foram obtidos através do método dos mínimos quadrados alternados (2.3.2). Foram executadas 46 iterações, tendo, este modelo, uma proporção de variância explicada de 0.4253 e uma norma de Frobenius (2.11) de 28.94. A Figura 3.14 mostra, numa perspetiva gráfica, os pesos das componentes do primeiro e segundo modo. Obviamente que, no terceiro modo, esta representação bidimensional é impossível visto que temos apenas uma componente, R_1.

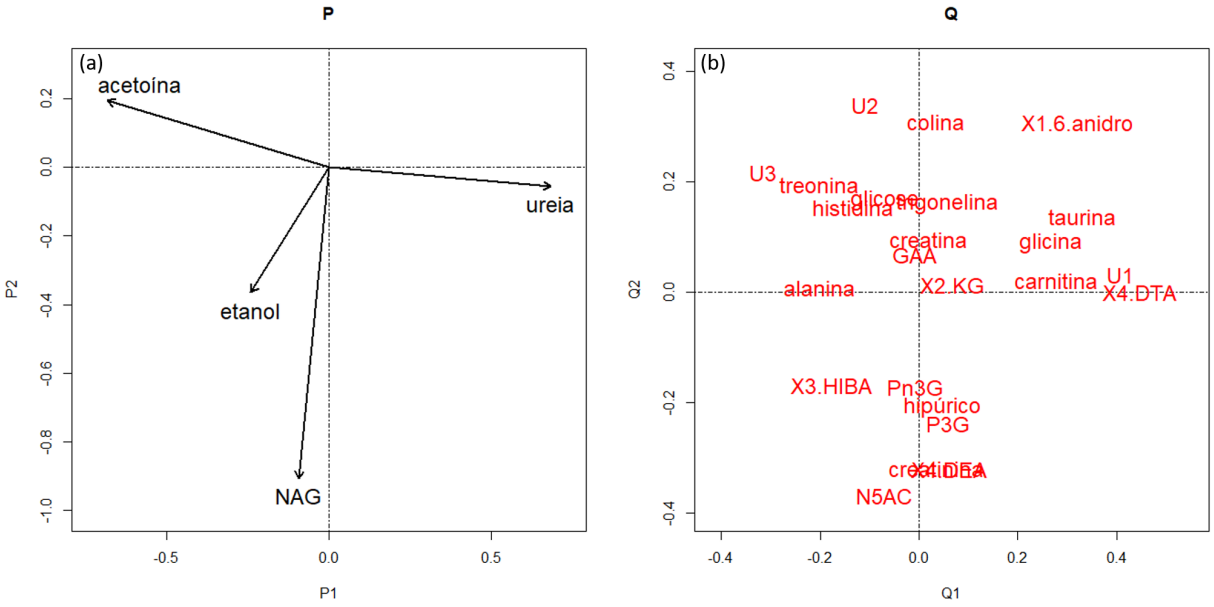

Figura 3.14: Coordenadas do primeiro (a) e segundo (b) modo.

Da Figura 3.14, em relação ao modo salivar (a), a primeira componente P_1 é caraterizada pelo contraste entre a acetoína e a ureia. A acetoína dispõe de uma coordenada negativa e a ureia de uma coordenada positiva. NAG apresenta um valor praticamente nulo. Por último, o etanol tem um peso negativo, mas mais pequeno quando comparado à acetoína ou ureia. Quanto a P_2, esta é marcada por um peso negativo maior, ao nível de NAG. Também com peso negativo, o etanol, mas não tão grande. A acetoína e a ureia têm valores bastante reduzidos nesta componente.

Quanto ao modo urinário (Figura 3.14 b), e a nível da primeira componente Q_1, os metabolitos que mais se destacam são o X4.DTA e U1, com pesos positivos, em oposição a U3, com um peso negativo. Em relação a Q_2, N5AC é o metabolito que mais se destaca, seguido de X3.DEA e da creatinina, com um maior número negativo. Em sentido contrário, U2, com uma carga positiva nesta componente.

Quando ao último modo, todas as condições em estudo apresentam valores positivos (Tabela 3.10), sendo, a segunda condição, aquela que mais se destaca, seguida da primeira. Dados estes valores, concluí-se que as interações que iremos estudar relacionar-se-ão, principalmente, com a passagem de tempo do primeiro para a terceiro trimestre (2C) e com a passagem do primeiro para o segundo trimestre (1C).

De forma a termos em consideração as combinações que transmitem um maior número de informação, a Tabela 3.11 apresenta o cubo $\mathbb{G}^{\star}{}_{2\times2\times1}$ (2.9).

	$\mathbb{G}^{\star}$			Proporção	
	P_1	P_2		P_1	P_2
			Componente R_1		
Q_1	-3.769	0.000		0.282	0.000
Q_2	0.000	-2.686		0.000	0.143

Tabela 3.11: Cubo $\mathbb{G}^{\star}$ do modelo diferencial.

A partir da Tabela 3.11, determinamos a proporção de variância que é explicada por cada componente construída. A Tabela 3.12 mostra esses valores, tendo em consideração a expressão (2.10).

Modo		Soma	Proporção de Variância de cada componente	
1	Saliva ($P = 2$)	0.425	0.282	0.143
2	Urina ($Q = 2$)	0.425	0.282	0.143
3	Trimestres ($R = 1$)	0.425	0.425	

Tabela 3.12: Proporção da variância explicada por cada uma das componentes construídas no modelo diferencial.

A projeção de cada uma das componentes P_1, P_2, Q_1 e Q_2 do primeiro e segundo modo em R_1, é observada na Figura 3.15, onde temos a primeira vs segunda componente do primeiro modo, com uma representação de 28.2% de variância explicada, e do segundo modo, contando com 14.2%.

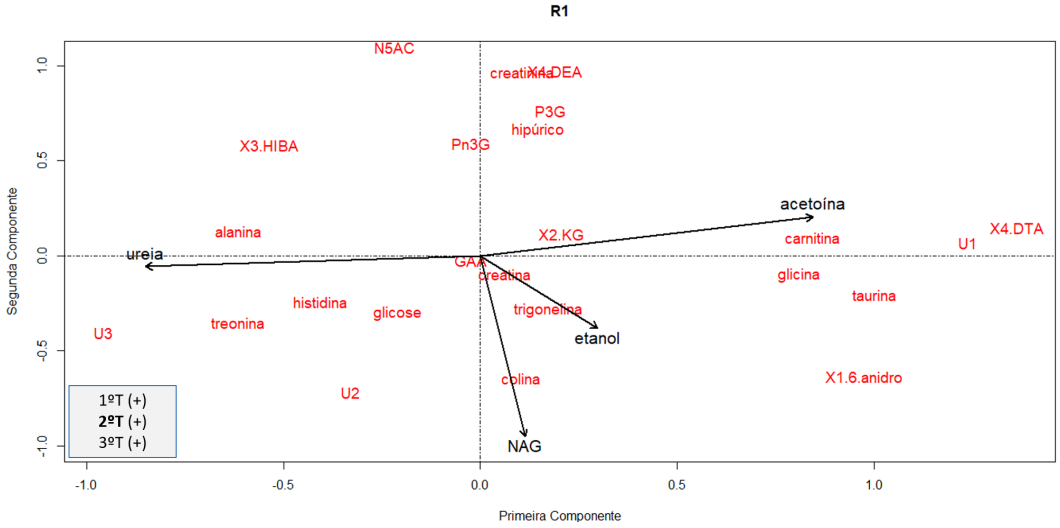

Figura 3.15: Projeção das componentes do primeiro e segundo modo em R_1 do modelo diferencial.

Tendo em conta a informação contida na Tabela 3.10, no cubo $\mathbb{G}^\star$ e no biplot da Figura 3.15, estamos em condições de interpretar os resultados que o modelo Co-Tucker diferencial oferece.

3.4.5 Interpretações do modelo diferencial

Primeiramente, começamos por estudar a combinação de componentes que mais variância explicada fornece. Dessa combinação, fazem parte as componentes P_1, Q_1 e R_1, com $g^\star_{111} = (-)3.769$ a representar 28.2% de variabilidade explicada (Tabela 3.11). Dado que a componente R_1 é constituída apenas por pesos positivos (Tabela 3.10), uma interação positiva resulta das seguintes combinações de componentes:

- **Caso 1**: $P_1(+) \times Q_1(-) \times R_1(+)$;

- **Caso 2**: $P_1(-) \times Q_1(+) \times R_1(+)$.

A ureia apresenta um coeficiente positivo em P_1, e o maior em termos absolutos. Tendo em conta o caso 1, em Q_1 destaca-se U3 com o peso mais negativo. Desta forma, existe uma interação positiva entre a ureia e U3 entre trimestres, na primeira, segunda e terceira condições, mas principalmente na segunda (maior coeficiente absoluto em R_1). Resulta disto que, o crescimento/decrescimento dos valores metabólicos espectrais da ureia e de U3, do 1º para o 3º (2C), do 1º para o 2º (1C) e do 2º para o 3º trimestres (3C), estão associados no mesmo sentido (por exemplo, se há um aumento de ureia na mudança de um trimestre para outro, será também expectável que o valor de U3 aumente). Em sentido contrário, considerando o metabolito acetoína, com coeficiente negativo e também elevado, em termos absolutos, em P_1, a interação com U3 é negativa nas três condições ($P_1(-) \times Q_1(-) \times R_1(+) \times (-) = (-)$). O crescimento/decrescimento dos valores metabólicos espectrais da ureia e de U3 do 1º para o 3º (2C), do 1º para o 2º (1C) e do 2º para o 3º (3C) trimestres foram contrários. Também com interação negativa, e com pesos positivos em Q_1, destacam-se os metabolitos X4.DTA e U1 com um comportamento espectral oposto ao da ureia ($R_1(+) \times Q_1(+) \times R_1(+) \times (-) = (-)$). Tendo agora em conta o caso 2, a acetoína (alto valor negativo em P_1) apresenta uma interação positiva com os metabolitos salivares X4.DTA e U1 (maiores pesos positivos em P_1) em todas as condições em estudo. Esta combinação de componentes mostra, claramente, um grande contraste de interações, e já descritas, ao nível da ureia com a acetoína, metabolitos salivares, e de U3 com X4.DTA e U1, metabolitos urinários. Graficamente, e em relação à projeção em R_1 (Figura 3.15), observa-se claramente o contraste entre estes metabolitos, dado o ângulo raso que se encontram uns dos outros, tendo em conta a direção dos vetores ureia e acetoína.

A segunda, e última, combinação de componentes a ser estudada, envolve P_2, Q_2 e R_1, com $g^{\star}_{221} = (-)2.686$ e com uma variância explicada de 14.3%. Tendo em conta que R_1 é apenas constituída por pesos positivos, e que os maiores pesos, em termos absolutos, em P_2, são negativos, a única combinação de componentes que resulta numa interação positiva é da forma:

1. $P_2(-) \times Q_2(+) \times R_1(+)$.

Assim, o NAG (maior valor negativo em P_2) apresenta uma interação positiva com U2 em todas as condições. Uma interação negativa resulta com os metabolitos urinários que têm os pesos mais negativos em Q_2. São eles: N5AC, X4.DEA e a creatinina. A nível gráfico observam-se as mesmas conclusões, dado o ângulo raso entre o vetor de NAG com N5AC, X4.DEA e creatinina. U2 encontra-se próximo de NAG, dada a projeção ortogonal com o vetor. Semelhantes interações poderíamos observar com o etanol. No entanto, o seu coeficiente em P_2 é pequeno, tal como o comprimento do vetor no biplot, pelo que está mais suscetível a sofrer fenómenos de ruído, dada a redução de dimensionalidade provocada pelo modelo.

Uma vez que os coeficientes $g^{\star}_{212}$ e $g^{\star}_{121}$ são, aproximadamente, nulos, mais nenhuma combinação de componentes é interpretada. A construção deste modelo e todas as interpretações realizadas podem ser consultadas em `https://github.com/Francisjcs1997/CT/blob/main/CoTucker_Diferencial.R`.

3.4.6 Análise de correlações no método diferencial

Mais uma vez, através do teste de hipóteses mencionado em 3.1, analisamos se a correlação de Spearman é ou não significativa. Envolvendo os metabolitos da primeira interpretação

analisada, os que apresentam maior interação são, do lado salivar, a acetoína e a ureia, com X4.DTA, da lado urinário.

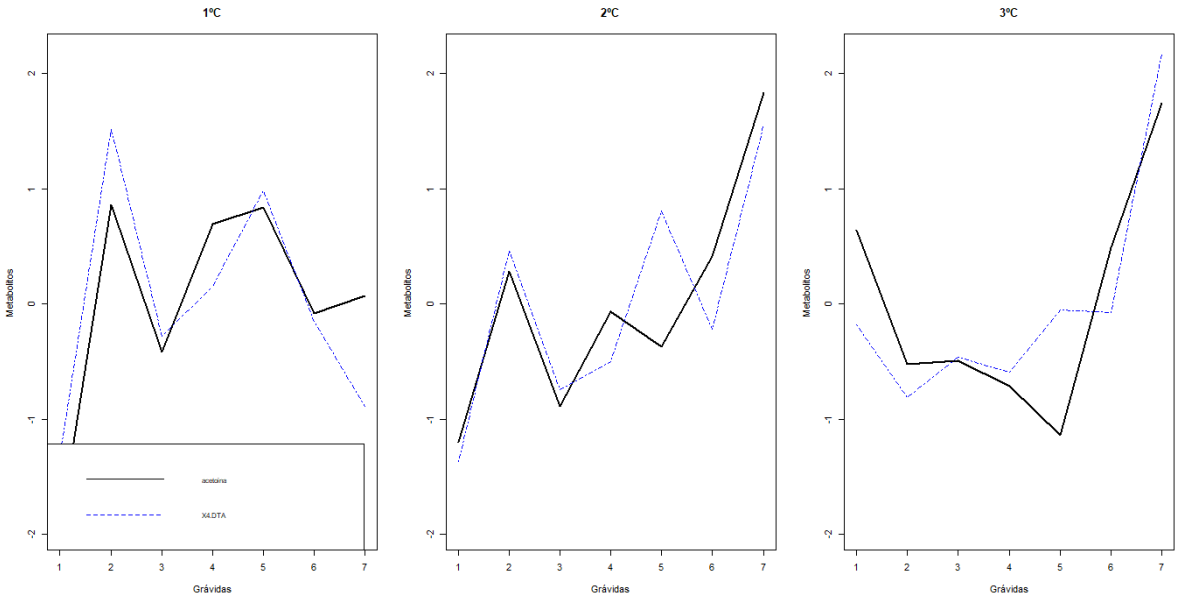

Figura 3.16: Espectro do NMR de acetoína, a preto, e X4.DTA, a azul, em cada uma das condições.

Em cada uma das condições, a Figura 3.16 mostra claramente uma interação positiva, com os crescimentos/decrescimentos espectrais semelhantes entre a acetoína e X4.DTA. Em sentido contrário, e recorrendo à Figura 3.17, a ureia apresenta uma interação negativa em cada uma das condições, principalmente a primeira e segunda.

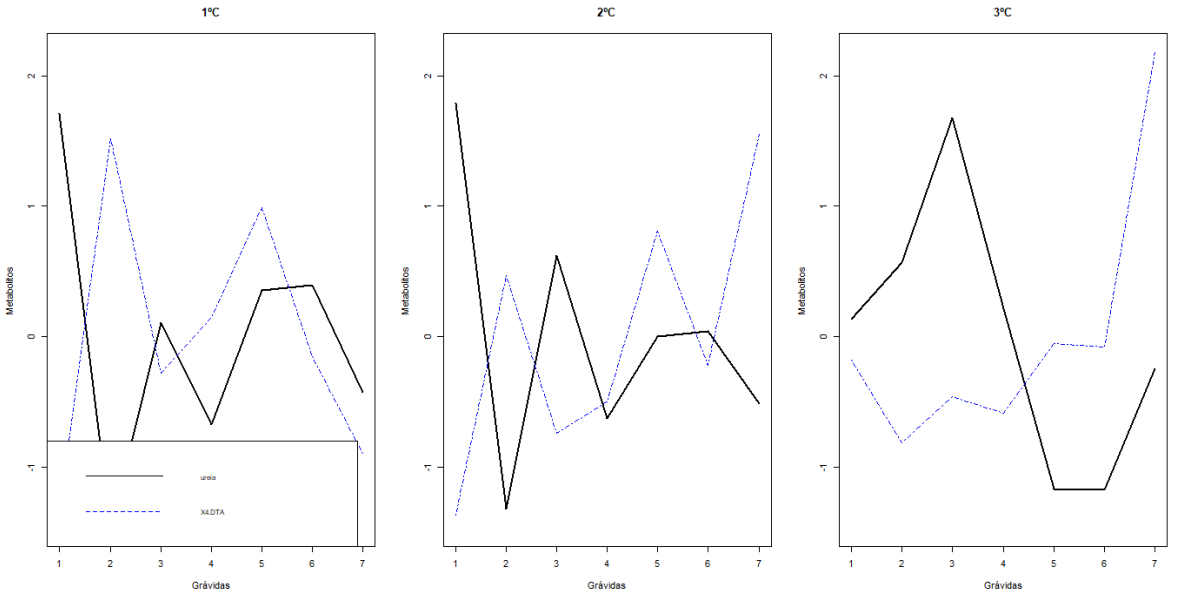

Figura 3.17: Espectro do NMR de ureia, a preto, e X4.DTA, a azul, em cada uma das condições.

Envolvendo NAG, a Figura 3.18 ilustra a interação negativa com N5AC em todas as condições.

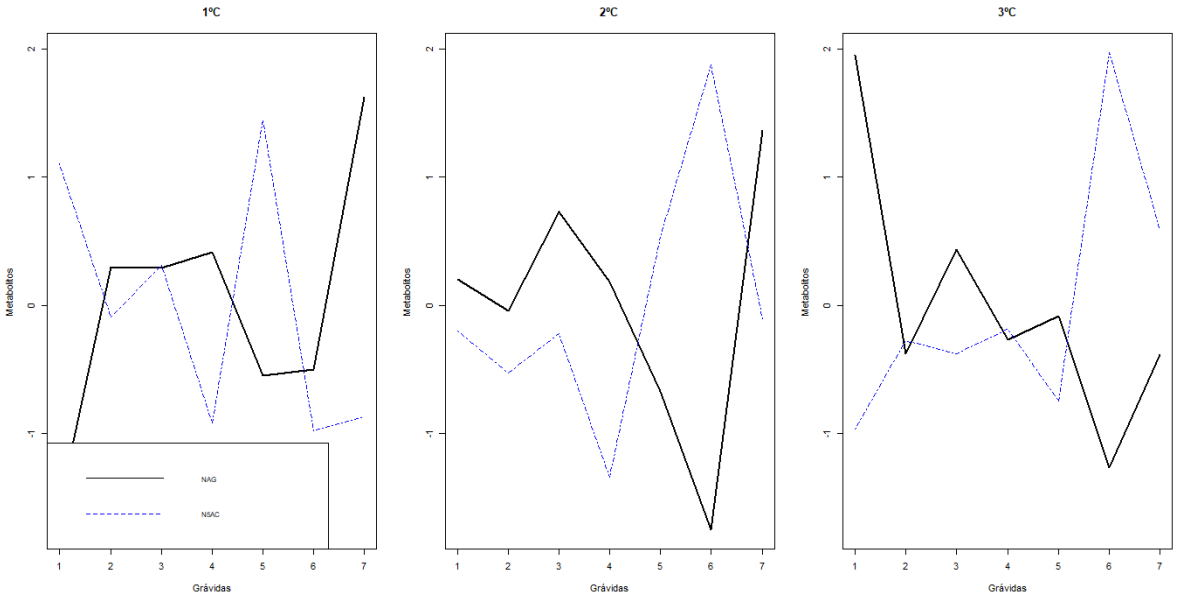

Figura 3.18: Espectro do NMR de NAG, a preto, e N5AC, a azul, em cada uma das condições.

Realizado o teste de hipóteses ao nível de significância $\alpha = 0.05$, a Tabela 3.13 mostra, em cada condição, os metabolitos cuja hipótese nula do correlação nula foi rejeitada.

	1C			2C		
Saliva	Urina	Correlação	Valor-p	Urina	Correlação	Valor-p
Acetoína	X4.DTA	0.8214	0.0341	carnitina	0.8571	0.0238
	U3	-0.8571	0.0238			
Ureia				X1.6.anidro	-0.7857	0.048
				U1	-0.7857	0.048
NAG	Trigonelina	0.8571	0.0238	Treonina	0.8571	0.0238
				U2	0.8571	0.0238
				U3	0.7857	0.048
	3C					
Ureia	Glicina	-0.8214	0.0341			
NAG	N5AC	-0.9286	0.0067			
	Glicina	-0.8571	0.0238			
	Creatinina	-0.8571	0.0238			
	U3	0.8571	0.0238			

Tabela 3.13: Correlação de Spearman da acetoína, ureia e NAG com os metabolitos urinários.

Usando os níveis usuais de significância $\alpha = 0.05$, existem evidência estatística de que a correlação de Spearman entre a acetoína com X4.DTA e U3 na primeira condição, e com carnitina na segunda condição, é diferente de zero. O mesmo podemos afirmar para a interação entre a ureia com X1.6.anidro e U1, na segunda condição, e com a glicina, na terceira condição. Considerando NAG, a hipótese nula é rejeitada para U2 na segunda condição e para N5AC

e a creatinina na terceira condição. De facto, estas interações já foram analisadas nestas passagens de tempo. Em especial, na passagem do primeiro para o terceiro trimestre (2C) e do primeiro para o segundo trimestre (1C), dado que contêm maior coeficiente em R_1. Também a Tabela 3.13 corrobora com o contraste observado entre a acetoína com a ureia e entre X4.DTA e U1 com U3.

3.4.7 Conclusões do modelo diferencial

O modelo Co-Tucker diferencial começa por exprimir o forte contraste envolvendo os metabolitos salivares, acetoína e ureia, e os metabolitos urinários, U1, X4.DTA e U3 entre mudanças de trimestres. As interações entre estes metabolitos são as mais fortes, principalmente ao nível da passagem do primeiro para o terceiro trimestre. Cerca de 2/3 da variância explicada do modelo concentra-se neste contraste de interações. A restante proporção é focada na interação entre NAG com N5AC, creatinina e X4.DEA. No entanto, esta é mais fraca. O facto do etanol não apresentar um peso consideravelmente grande nas componentes P_1 e P_2, e o reduzido comprimento do vetor na Figura 3.13, impossibilitam o modelo de transmitir informações não sujeitas a ruído acerca deste metabolito.

A ureia correlacionou-se negativamente com a carga bacteriana, o que indica a utilização de metabolitos derivados do hospedeiro pela microbiota oral. De facto, a ureia está implicada no aumento do pH oral por meio da conversão em amoníaco. Portanto, o consumo (diminuição) da ureia microbiana pode ser um mecanismo de sobrevivência na presença do aumento do crescimento de batérias sacarolíticas acidogénicas [14]. As interações reveladas pelo modelo deram-se em todos os momentos de tempo, no entanto o modelo diferencial deu uma ligeira relevância à passagem do primeiro para o terceiro trimestre e do primeiro para o segundo trimestre.

Capítulo 4

Considerações finais

Nesta dissertação foi abordada uma metodologia para a análise de dados tridimensionais emparelhados. Ela apresenta pontos fortes mas também algumas desvantagens. Como desvantagem podemos apontar, na estimação dos coeficientes das matrizes $\mathbf{A}^X$, $\mathbf{B}^Y$ e $\mathbf{C}$, pelo método dos mínimos quadrados alternados, este poder apresentar um elevado tempo de execução quando aplicado sobre dados com muitas variáveis. Esta desvantagem não foi observada neste trabalho, uma vez que os dados espetrais em estudo da saliva e urina considerados, são de dimensões não elevadas (4 variáveis salivares vs 24 variáveis urinárias). Porém, com um elevado número de elementos numa qualquer das três dimensões do cubo, o Co-Tucker pode ser difícil de implementar, dado o elevado tempo para processamento.

Outra questão sensível nesta metodologia é a dependência do pré-processamento dos dados, pois esta pode influênciar os resultados finais obtidos. Também, uma escolha menos adequada para a transformação a considerar nos dados, por exemplo centrar ou normalizar, pode conduzir a obtenção de um modelo menos eficaz, no sentido de não transmitir as informações mais relevantes. A metodologia Co-Tucker conduz a uma redução da dimensão, levando certamente a uma perda de informação. O que se pretende é que a transformação a considerar não oculte informação mais relevante.

Em contrapartida, a aplicação da metodologia Co-Tucker sobre diferentes transformações nos dados, pode torna-se interessante dado que possibilita explorar diferentes perspetivas na interação entre variáveis e/ou eventos em estudo. Por conseguinte, podemos dizer que o modelo Co-Tucker proporciona flexibilidade ao nível da interpretabilidade.

A indicação da fonte de variabilidade é outro ponto a favor do método Co-Tucker. As combinações que revelam um maior número de informação indicam onde deve ser dado maior enfâse ao nível da interpretação de resultados. Ainda no âmbito da interpretabilidade, a possível escolha de um diferente número de componentes, em cada modo, permite extrair, do modelo reduzido, informações que de outro modo poderiam não ser capturadas e ainda sob diferentes perspetivas.

Outra vantagem desta metodologia que importa salientar, é a componente gráfica que esta envolve para além da numérica. A capacidade de permitir visualizar os dados é uma mais valia desta metodologia. A análise do modelo sobre as matrizes $\mathbf{A}^X$, $\mathbf{B}^Y$ e $\mathbf{C}$ acompanhada de uma análise de biplots, oferece ao investigador uma maior facilidade de compreensão do que está a ser estudado.

Finalmente é importante destacar que o método Co-Tucker trabalha com variáveis quantitativas, como é o caso das variáveis espectrais que foram estudadas no Capítulo 3. Seria interessante, como trabalho futuro, avaliar a possibilidade de estender o modelo a cubos emparelhados envolvendo dados qualitativos. Tal implica que a interação medida através do cálculo de matrizes de covariâncias cruzadas teria de ser adaptada, como por exemplo utilizando medidas de associação. E adicionalmente estudar a interpretabilidade desta adaptação ao nível dos biplots.

Bibliografia

[1] Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. Psychometrika, 31(3), 279-311.

[2] Ceulemans, E., & Kiers, H. A. L. (2006). Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method. British Journal of Mathematical and Statistical Psychology, 59(1), 133-150.

[3] Kiers, H. A. L., & Der Kinderen, A. (2003). A fast method for choosing the numbers of components in Tucker3 analysis. British Journal of Mathematical and Statistical Psychology, 56(1), 119-125.

[4] Dray, S., Chessel, D., & Thioulouse, J. (2003). Co-inertia analysis and the linking of ecological data tables. Ecology, 84(11), 3078-3089.

[5] Timmerman, M. E., & Kiers, H. A. L. (2000). Three-mode principal components analysis: Choosing the numbers of components and sensitivity to local optima. British Journal of Mathematical and Statistical Psychology, 53(1), 1-16.

[6] Acar, E., & Yener, B. (2009). Unsupervised multiway data analysis: A literature survey. IEEE Transactions on Knowledge and Data Engineering, 21(1), 6-20.

[7] Mendes, S., Fernández-Gómez, M. J., Marques, S. C., Pardal, M. Â., Azeiteiro, U. M., & Galindo-Villardón, M. P. (2017). CO-tucker: a new method for the simultaneous analysis of a sequence of paired tables. Journal of Applied Statistics, 44(15), 2729-2755.

[8] Martin-Barreiro, C., Ramirez-Figueroa, J. A., Nieto-Librero, A. B., Leiva, V., Martin-Casado, A., & Purificación Galindo-Villardón, M. (2021). A new algorithm for computing disjoint orthogonal components in the three-way tucker model. Mathematics, 9(3), 1-22.

[9] Hadden, D. R., & McLaughlin, C. (2009). Normal and abnormal maternal metabolism during pregnancy. Seminars in Fetal and Neonatal Medicine, 14(2), 66-71.

[10] Kroonenberg, P. M. (2007). Applied Multiway Data Analysis. Applied Multiway Data Analysis (pp. 1-589). Wiley Blackwell.

[11] Nicholson, J. K., Lindon, J. C., & Holmes, E. (1999). "Metabonomics": Understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica. Taylor and Francis Ltd.

[12] Vignoli, A., Ghini, V., Meoni, G., Licari, C., Takis, P. G., Tenori, L.,... Luchinat, C. (2019, January 21). High-Throughput Metabolomics by 1D NMR. Angewandte Chemie - International Edition. Wiley-VCH Verlag.

[13] Trygg, J., Holmes, E., & Lundstedt, T. (2007, February). Chemometrics in metabonomics. Journal of Proteome Research.

[14] Gardner, A., Parkes, H. G., So, P. W., & Carpenter, G. H. (2019). Determining bacterial and host contributions to the human salivary metabolome. Journal of Oral Microbiology, 11(1).