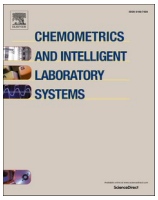
Identification of metabotypes in complex biological data using tensor decomposition

Chemometrics and Intelligent Laboratory Systems 233 (2023) 104733

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics



Identification of metabotypes in complex biological data using tensor decomposition

Viktor Skantze [a,b,*], Mikael Wallman [a], Ann-Sofie Sandberg [b], Rikard Landberg [b], Mats Jirstrand [a], Carl Brunius [b]

[a] Fraunhofer-Chalmers Research Centre for Industrial Mathematics, Gothenburg, Sweden
[b] Department of Biology and Biological Engineering, Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden

ARTICLE INFO

Keywords:
Personalized nutrition
Metabotyping
Tensor decomposition
Multiway analysis
Data mining

ABSTRACT

Differences in the physiological response to treatment, such as dietary intervention, has led to the development of precision approaches in nutrition and medicine to tailor treatment for improved benefits to the individual. One such approach is to identify metabotypes, i.e., groups of individuals with similar metabolic profiles and/or regulation. Metabotyping has previously been performed using e.g., principal component analysis (PCA) on matrix data. However, metabotyping methods suitable for more complex experimental designs such as repeated measures or cross-over studies are needed. We have developed a metabotyping method for tensor data, based on CANDECOMP/PARAFAC (CP) tensor decomposition. Metabotypes are inferred from CP scores using k-means clustering, and robustness is evaluated using bootstrapping of metabolites. As a proof-of-concept, we identified metabotypes from metabolomics data where 79 metabolites were analyzed in 8 time points postprandially in 17 overweight men that underwent a three-arm dietary crossover intervention. Two metabotypes were found, characterized by differences in amino acid metabolite concentration, that were differentially associated with baseline plasma creatinine ($p = 0.007$) and with the baseline metabolome ($p = 0.004$). These results suggest that CP decomposition provides a viable approach for metabotype identification directly from complex, high-dimensional data with improved biological interpretation compared to the more simplistic PCA approach. A simulation study together with results from measured data concluded that several preprocessing methods should be taken into consideration for CP-based metabotyping on complex tensor data.

1. Introduction

Non-communicable diseases (NCDs), such as cardiovascular diseases, cancers, diabetes, and chronic lung diseases, account for 71% of global deaths each year according to the World Health Organization [1]. Diet is one of the main modifiable risk factors for the prevention of NCDs [2]. Official dietary guidelines are issued by governmental agencies to promote healthy eating habits for the entire population, but there are large differences in inter-individual responses to the same diet. This calls for new strategies to improve the effectiveness of prevention. Providing foods or diets that are tailored to groups or individuals offers such a strategy [3].

Personalized nutrition (PN) can be defined as providing the right diet to the right person at the right time [4]. Different factors will determine how individuals or specific groups respond to a certain food or diet. Such factors include health status, metabolome, medication, gut microbiota, behavior, and the exposome [5]. To provide PN, the different factors affecting the response across individuals need to be considered. One approach towards PN is to fit diet to groups of individuals that have a similar metabolic profile, so-called metabotypes [6]. Although different definitions have been used [7], we define metabotyping as the grouping of individuals that have fundamental similarities in their metabolic phenotype, reflected by similar response to the same diet. The identification of metabotypes or metabolic phenotypes usually entails clustering individuals according to variables such as diet, lifestyle, anthropometric measures, clinical parameters, metabolic data, and gut microbiota [6]. The metabolic system can be exited in many ways and constitutive differences in metabolic profile is more probable to be captured using many dietary interventions on the same individuals rather than solely one. In this paper we use the metabolic response to three diets as an indicator of metabolic phenotype.

Metabolomics refers to the comprehensive measurement of small

* Corresponding author. Fraunhofer-Chalmers Research Centre for Industrial Mathematics, Gothenburg, Sweden.
E-mail address: viktor.skantze@fcc.chalmers.se (V. Skantze).

https://doi.org/10.1016/j.chemolab.2022.104733
Received 21 February 2022; Received in revised form 15 November 2022; Accepted 16 December 2022
Available online 18 December 2022

organic molecules in biological samples and has shown promise as a tool to capture metabotypes [8]. Metabolomics data may reflect dietary intake as well as dynamic metabolic responses to such intakes [9]. Metabolomics data thus provide a more intelligible identification of clusters of metabolic phenotypes than solely anthropometric and clinical data [10].

Different clustering methods have been proven useful to group individuals into distinct metabotypes [11,12]. Clustering is usually performed on two-mode data, i.e., a feature table where the first mode (typically the rows) is constituted by observations (e.g., individuals) and the second mode (typically the columns) by phenotypic variables (e.g., anthropometric measures and clinical parameters) [10,13]. Dimensionality reduction methods, like principal component analysis (PCA), non-negative matrix factorization, and singular value decomposition, are often used to compress the data and to identify clusters [14–16].

These dimensionality reduction methods operate by matrix decompositions and are thus well suited for such study designs that generate matrix data. However, several experimental designs, involving e.g., crossover or repeated measures, can generate datasets that can conveniently be thought of as multidimensional arrays, also known as tensors or multiway arrays, and can naturally be addressed using more than two indices (one for each mode or dimension). One way to analyze such data is to unfold higher dimensions in such a way that a matrix is obtained and then apply matrix analysis methods (Fig. 1).

However, unfolding higher dimensions into a matrix format inevitably decouples parts of the structure inherent in the data, which potentially leads to a loss of information. Thus, there is a need for methods specifically adapted to more complex study designs. One class of methods meeting this criterion is tensor analysis, which keeps the structure of the data intact and therefore holds potential to capture more information than the matrix analysis of corresponding unfolded data. Tensor decompositions were recently used for exploring the dynamics in simulated time-resolved metabolomics data [17], showing promising results. Here, we hypothesize that tensor analysis could be a useful tool to identify metabotypes from complex data consisting of more than two modes.

Tensor decomposition can be accomplished using e.g., Tucker decomposition [18] or CANDECOMP/PARAFAC (CP) [19,20]. While Tucker decomposition can provide a better fit to the data (due to orthogonal components and more parameters), it is normally also more challenging to interpret biologically since it does not necessarily provide an equal number of components between modes and also provides a so-called core tensor to address the interaction between components that does not intuit easily for our purposes. Furthermore, the orthogonal components can impair the interpretation of the biological data if the true underlying features are not orthogonal to each other. For these reasons, we chose to investigate CP for metabotyping.

CP can suffer from two-factor degeneracies which is an artifact of optimization, where components are mirrored and do not provide further information about the data [21]. Such degeneracies indicate that the best reconstruction of the data having r components does not exist and that the decomposition consequently should not be used for analysis [22,23]. However, these limitations can be addressed by imposing modeling constraints, such as non-negativity [24]. Moreover, the preprocessing of the tensor data prior to applying CP is not straightforward and should ideally be investigated on a study-specific basis [25].

Our aim was therefore to investigate the potential of CP to infer metabotypes, observed as groups of differential responders to treatment, with special emphasis on the effects of preprocessing data. We applied this methodology to time-resolved tensor data from an acute postprandial dietary intervention crossover study measuring 79 metabolites (from GC-MS metabolomics) at 8 time points (0–7 h) for 17 individuals, each undergoing the same three dietary interventions (pickled herring, baked herring, and baked beef), i.e., a dataset having four modes (individuals, time, metabolites, and diet) [26]. Obtained clusters were associated to anthropometric and clinical baseline data as a means to assess their metabolic relevance. The robustness of the clusters was further assessed using bootstrapping. The workflow was compared to using the *de facto* field standard PCA on the unfolded tensor. We also applied the method to simulated data, with a structure similar to the measured data, to investigate how the workflow and preprocessing methods would perform in an ideal case with known (*a priori*) ground truth. Code in Matlab to perform the proposed CP-based metabotyping workflow on generated synthetic data is freely available at https://github.com/FraunhoferChalmersCentre/MetabotypingUsingTensorDecomposition.

2. Materials and methods

The metabotyping workflow was subdivided into five steps as illustrated in Fig. 2:

(I) Time-resolved metabolomics data was obtained from a crossover acute postprandial dietary intervention study and organized as a tensor having four modes (individuals, time, metabolites and diets) (Fig. 2 (I), Sec. 2.2).
(II) The data was preprocessed to eliminate outliers that could affect the tensor decomposition, and to normalize the variables prior to decomposition (Fig. 2 (II), Sec. 2.3).
(III) The preprocessed data were decomposed using tensor decomposition to reduce the dimensionality of the data, using the "N-way toolbox for MATLAB" [27] (Fig. 2 (III), Sec. 2.4). A heuristic tensor rank assessment was performed on the preprocessed data (Fig. 2 (III), Sec. 2.5).
(IV) The scores (representing the individuals) of the decomposition were clustered while loadings were used for dynamical interpretation of the data (Fig. 2 (IV), Sec. 2.6).
(V) The clusters were investigated for associations with baseline clinical data to inform potential metabotypes. Additionally, cluster robustness was assessed by bootstrapping of metabolites and the metabolic relevance between cluster associations and

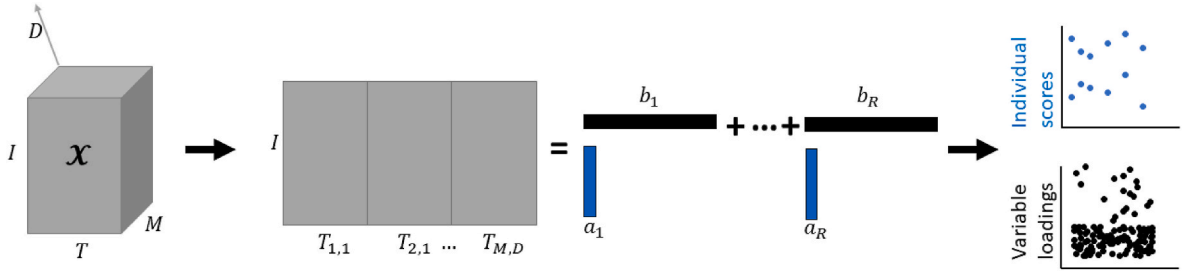


Fig. 1. PCA on the tensor $\mathcal{X}$ unfolded into a matrix, where the individuals are kept in one mode (rows, size I) and time, metabolites, and diets are concatenated into the same "variable" mode (columns size $T \times M \times D$). Scores and loadings of the PCA components can then be used for the identification of metabotypes as well as for visualization and interpretation.

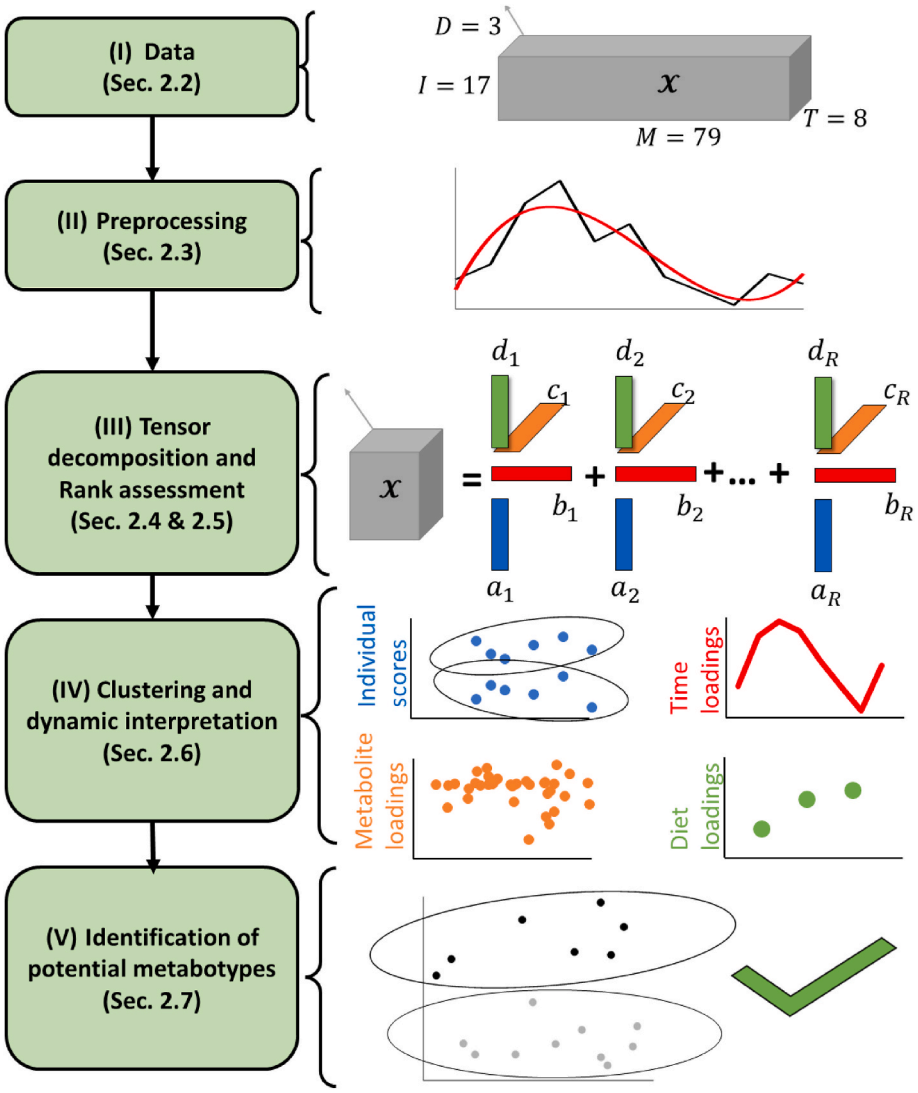


Fig. 2. Metabotyping workflow with indicated section numbers for further detail.

metabolite loadings were investigated to further justify the proposed metabotypes (Fig. 2 (V), Sec. 2.7).

The metabotyping workflow was applied to two data representations: **i)** the original metabolite time series data and; **ii)** baseline-subtracted data, i.e., where the first measurement value in each metabolite time series was subtracted from the measurement values of the rest of its data points, effectively making these preprocessed metabolite time series all start at zero. Parts (II)-(V) were also performed on a simulated dataset created with the tensor decomposition structure, to investigate how the preprocessing methods and workflow performed on ideal data. All calculations and analyses for metabotyping were performed in MATLAB version R2019a.

2.1. Dietary study design and data structure

2.1.1. Measured data

The metabolomics data used in this study originated from a randomized crossover intervention study examining acute postprandial metabolomic profiles and regulation associated with three different dietary meal exposures: Baked herring, pickled herring, and beef [26]. In the original study, targeted metabolites were measured from blood plasma using gas chromatography-mass spectrometry (GC-MS). The blood samples were taken from 17 middle-aged overweight men (BMI 25–30 kg/m^2, 41–67 years of age).

For each treatment, 79 metabolites were measured from baseline to 7 h after consumption at 1 h intervals, (Fig. 2). In addition, several plasma clinical and anthropometric measures were taken at baseline, including alanine aminotransferase (ALAT), aspartate transaminase (AST), gamma-glutamyl transferase (GGT), cholesterol (CHOL), low-density lipoprotein (LDL), creatinine(CR), thyroid stimulating hormone (TSH), and body mass index (BMI) [28].

The metabolomics data constituted a multiway array with four modes. Individuals (n = 17), time points (n = 8), metabolites (n = 79), and diets (n = 3) were stacked in a fourth order tensor $\mathcal{X} \in \mathbb{R}^{I \times T \times M \times D}$ where I, T, M, and D denote the number of individuals, time points, metabolites, and diets, respectively, (Fig. 3). The data is most easily viewed as three third order tensors where each tensor is 79 matrices of 17 rows and 8 columns (metabolite slabs) stacked to constitute a fourth order tensor of metabolite matrices. Three examples of the plotted slabs are seen above the tensor $\mathcal{X}$ (Fig. 3). The matrix slabs corresponding to different metabolites generally have different overall scales and offsets, as is common for GC-MS data.

2.1.2. Simulated data

Fourth order tensor data was generated as a tensor using the CP model and contained 17 individual time series with 8 time points per metabolite (n = 79) and 3 diets, to emulate the measured data.

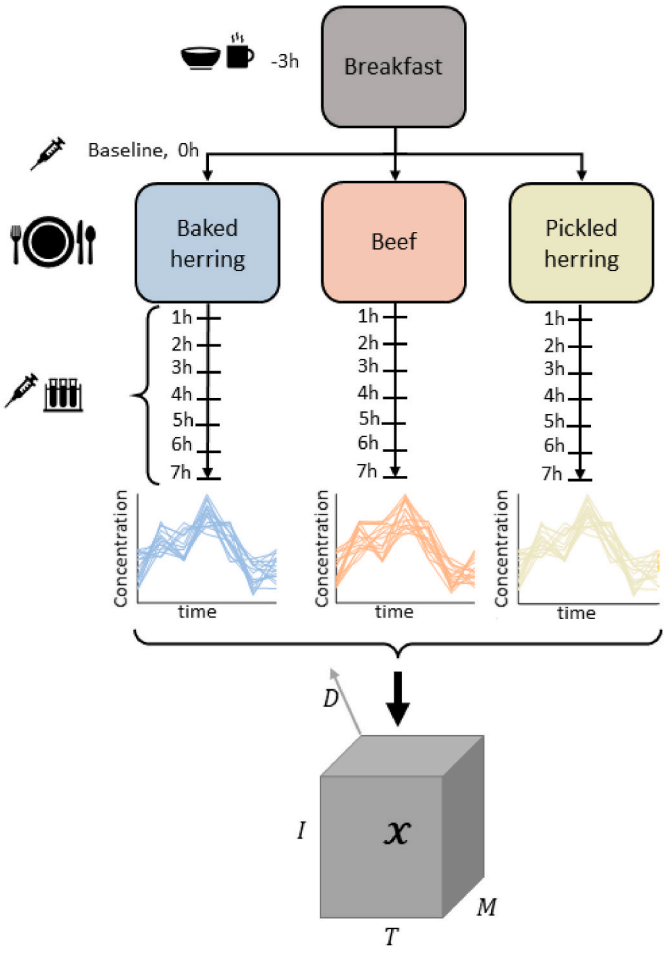


Fig. 3. Schematic of the experimental design. All subjects consumed all three test meals on different occasions in a crossover design. Breakfast was served 3 h before the test meals. Baseline blood samples were taken just before the test meals and then every hour for 7 h. One week of washout period was included between each meal. Collected data forms a fourth order tensor ($\mathcal{X}$) with individuals (I), metabolites (M), time points (T), and diets (D), i.e., four modes (axes).

To generate realistic data, metabolites were constructed in distinct clusters representing discrete metabolite patterns. Three CP components (three outer product terms) were used to construct the data as in Eq. (1) with the general definition in Eq. (11).

$$\mathcal{X}_s = \boldsymbol{a}_{fast} \circ \boldsymbol{b}_{fast} \circ \boldsymbol{c}_{fast} \circ \boldsymbol{d}_{all} + \boldsymbol{a}_{clusters} \circ \mathbb{1} \circ \boldsymbol{c}_{clusters} \circ \boldsymbol{d}_{first} + \boldsymbol{a}_{slow} \circ \boldsymbol{b}_{slow} \circ \mathbf{c}_{slow} \circ \boldsymbol{d}_{all} \tag{1}$$

Where $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ and $\boldsymbol{d}$ are design vectors representing individual time series amplitude, type of dynamic time series from a continuous linear dynamical system, scaling of slabs, and scaling of third order tensor representing one diet. The outer product operator is denoted $\circ$.

The vectors $\boldsymbol{a}_{fast}, \boldsymbol{a}_{slow}$ have length 17 and were drawn from a uniform distribution between 0 and 1 and $\boldsymbol{a}_{clusters}$ is a vector of the same length, where the first ten elements were 3 and the rest 1, representing a difference in amplitude to create two clusters of individuals. The vectors $\boldsymbol{b}_{fast}$ and $\boldsymbol{b}_{slow}$ (lengths 8) were computed from the dynamical system (Eqs. (2) and (3)). Here, two dynamic profiles were used: a fast dynamic profile ($\boldsymbol{b}_{fast}$) with an absorption rate $k_a = 4\ t^{-1}$ and an elimination rate $k_e = 1\ t^{-1}$, and a slow dynamic profile ($\boldsymbol{b}_{slow}$) with $k_a = 0.5\ t^{-1}$ and $k_e = 0.01\ t^{-1}$. The dynamical system consisted of a linear compartment model containing two states (x_1, x_2) where the second represents metabolite concentration used to create dynamical profiles and the absorption and elimination rate are represented by k_a and k_e (Eqs. (2) and (3)).

$$\frac{dx_1}{dt} = -k_a x_1 \tag{2}$$

$$\frac{dx_2}{dt} = k_a x_1 - k_e x_2 \tag{3}$$

The initial state values of $x_1(0)$ and $x_2(0)$ were 10 C and 1 C, respectively, where C is the state unit. The second component (second outer product term in Eq. (1)) was modeled using a constant dynamic profile represented by $\mathbb{1}$. The scalings per slab are represented by $\boldsymbol{c}_{fast}$, $\boldsymbol{c}_{clusters}$ and $\boldsymbol{c}_{slow}$ which have length 79 and are encoded as zero vectors in all elements except for the first twenty, the first ten, and the last 59, respectively. The non-zero elements were drawn from a uniform distribution between 0 and 100 and represent the scale of each metabolite slab. The vectors $\boldsymbol{d}_{all}$ and $\boldsymbol{d}_{first}$ are the vectors [1 1 1] and [1 0 0], respectively.

Thus, the simulated data were made to contain two dynamic response patterns represented by two groups of mutually similar metabolites, one fast (containing ground truth clusters of individuals) and one slow dynamic pattern. Consequently, the first component models a group of highly correlated metabolites with fast dynamics (#1–20), with amplitudes drawn from the distribution $\boldsymbol{a}_{fast}$ in all three diets and uniformly random scaling per metabolite slab. The second component models two constant offsets only applied to the first ten fast metabolites in the first diet, creating two clusters of individuals (#1–10 and #11–17). The third component models a group of highly correlated metabolites with slow dynamics (#21–79), with amplitudes drawn from the distribution $\boldsymbol{a}_{slow}$ in all three diets and uniformly random scaling per metabolite slab. A graphical overview of the structure of the simulated data is found the supplementary material (Fig. S1).

Scalar offsets per metabolite slab $\mu_{i,j}$ were also added as:

$$\mathcal{X}_{s+}(:,:,i,j) = \mathcal{X}_s(:,:,i,j) + \mu_{i,j}\mathbb{1}\mathbb{1}^{\intercal} = \boldsymbol{A}\boldsymbol{P}_{i,j}\boldsymbol{B}^{\intercal} + \mu_{i,j}\mathbb{1}\mathbb{1}^{\intercal}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are the matrices created by the column-wise collected vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ and $\boldsymbol{P}_{i,j}$ is a diagonal matrix with the element-wise product of the $i:th$ and $j:th$ rows of matrices $\boldsymbol{C}$ and $\boldsymbol{D}$ as its diagonal, where $\boldsymbol{C}$ and $\boldsymbol{D}$ are the matrices of the column-wise collected vectors $\boldsymbol{c}$ and $\boldsymbol{d}$. The offsets per slab $\mu_{i,j}$ were drawn as integers from a uniform distribution between 1 and 100. Uniform random noise between 0 and $0.1\sigma_{i,j}$ was added to each time series at every time point, where $\sigma_{i,j}$ is the scalar overall standard deviation per slab before adding the offset and i and j are the metabolite and diet indices, respectively.

2.2. Preprocessing

Outliers in the measured data were identified as points or time series lying farther than three standard deviations from the population mean per metabolite. Outlier points were imputed and time series were scaled, to lie within three standard deviations from the mean, to reduce their impact on the tensor decomposition.

To address the amplitude and offset differences in metabolites when using CP decomposition, three preprocessing methods were investigated (named (P1), (P2) and (P3)). The first method (P1) consisted in scaling all data per metabolite (scaling within the metabolite mode) to unit variance as in Eqs. (4) and (5), where μ_m is the scalar overall average of the metabolite and $x_{i,t,m,d}$ denotes the tensor element-wise.

$$\widehat{x}_{i,t,m,d} = \frac{x_{i,t,m,d}}{s_m} \tag{4}$$

$$s_m = \sqrt{\left(\sum_{i=1}^{I} \sum_{t=1}^{T} \sum_{t=1}^{D} \frac{\left(x_{i,t,m,d} - \mu_m\right)^2}{I\,T\,D} \right)} \tag{5}$$

When preprocessing only by scaling within the metabolite mode we used the scaling to unit variance (Eq. (4)) instead of to unit mean square (Eq. (7)) as we observed offsets between metabolites which is assumed when scaling to unit variance [25].

The second method (P2) used centering across the individual mode (Eq. (6)) (effectively removing the average individual from the data) prior to scaling within the metabolite mode (Eq. (7) and (8)) as in the first preprocessing method but scaling to unit mean square [25].

$$\overline{x}_{i,t,m,d} = x_{i,t,m,d} - \sum_{i=1}^{I} \frac{x_{i,t,m,d}}{I} \tag{6}$$

$$\widehat{x}_{i,t,m,d} = \frac{\overline{x}_{i,t,m,d}}{s_m} \tag{7}$$

$$s_m = \sqrt{\left(\sum_{i=1}^{I} \sum_{t=1}^{T} \sum_{t=1}^{D} \frac{\overline{x}^2{}_{i,t,m,d}}{I\,T\,D} \right)} \tag{8}$$

The third preprocessing method (P3) consisted in centering each slab $\mathcal{X}(:,:,m,d)$ to have an overall zero mean (Eq. (9) and (10)) prior to scaling within the metabolite mode as in the second preprocessing method.

$$\overline{\mathcal{X}}(:,:,m,d) = \mathcal{X}(:,:,m,d) - \widehat{\mu}_{\mathrm{m,d}} \mathbb{1}\mathbb{1}^{\mathrm{T}} \tag{9}$$

$$\widehat{\mu}_{m,d} = \sum_{i=1}^{I} \sum_{t=1}^{T} \frac{x_{i,t,m,d}}{I\,T} \tag{10}$$

Notice that the centering of slabs in the third preprocessing method does not reduce the rank of a CP model but only replaces potentially large offsets with smaller ones and is not commonly used for preprocessing in tensor decomposition [25].

2.3. Tensor decomposition

CP decomposes a tensor $\mathcal{X}$ into a sum of outer products of vectors. In the case where the data has four modes $\mathcal{X} \in R^{I \times T \times M \times D}$, the decomposition is

$$\mathcal{X} \approx \sum_{r=1}^{R} \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r \circ \boldsymbol{d}_r. \tag{11}$$

Here the factor vectors of each component r are $\boldsymbol{a}_r \in \mathbb{R}^I, \boldsymbol{b}_r \in \mathbb{R}^T, \boldsymbol{c}_r \in \mathbb{R}^M$, $\boldsymbol{d}_r \in \mathbb{R}^D$ for $r = 1, \ldots, R$ and R is the number of components used to approximate the data $\mathcal{X}$. The factor vectors of a CP component are analogous to the score and loading vectors of a PCA component. Hence, we will refer to the factor vectors of the individual mode as scores while the ones corresponding to the remaining modes will be referred to as time, metabolite, and diet loadings. We also refer to a CP decomposition as a "model" interchangeably. Contrary to PCA, the components of an unconstrained CP model are not orthogonal which implies that the components in a 2-component model are not necessarily found in a 3-component model. Furthermore, since CP does not suffer from rotational freedom (contrary to matrix decompositions, unitary transformations change the fit of the model in CP), it can be assumed to be unique (apart from permutations of components and scalings of factor vectors) [19,20].

2.4. Rank assessment

We investigated methods for assessing the rank of the tensor, since this is not as straightforward as in the matrix case [29]. The rank of a tensor provides information about how many CP components are needed to reconstruct the tensor and we used four heuristic methods for this purpose.

The first method was the minimal cosine similarity between CP decompositions initialized from random starting points, which represents the stability of the decomposition, similar to the method described in Wiliams et. al, 2018 [30]. Low similarity suggests that the tensor decomposition is sensitive to initial values of the components, consequentially breaking the assumption of uniqueness and therefore should not be used for analysis (non-valid models). The second method was the core consistency introduced in Bro and Kiers, 2003 [31], which makes a similarity comparison with the Tucker decomposition. Valid models range from a core consistency of 50% (acceptable) to 100% (optimal). The third method assessed the reconstruction error between the decomposition and the data [32]. A stagnation in reconstruction error with an increasing number of components may indicate that noise is being modeled, similar to the use of Scree plots in PCA [30]. The fourth method consisted of examining the models for two-factor degeneracies. We used the indicator for degenerate models given by the parafac() function in the N-way toolbox while using an optimization convergence criterion of 10^{-8} relative change of fit [27].

Additionally, we compared CP time loading vectors to dynamic metabolite time series in the raw data visually [24]. High similarity in these time profiles could suggest that the decomposition captures the dynamic behavior of the data and that the model is valid.

In addition to the best approximative model, lower rank models were also used for investigating metabotyping under the rationale that they represent the learning of overarching features of the data that may not be captured in higher rank approximations.

2.5. Clustering and dynamic interpretation

In every CP decomposition (model), one score (individuals) and three loading vectors (time, metabolites, and diets) were obtained per component. Potential metabotypes were identified from k-means clustering (50 repetitions) of individuals from all combinations of the score vectors (number of i-combinations for all i going from 1 to R, resulting in $2^R - 1$ combinations). As an example, a two-component model would be used to identify potential metabotypes from clustering of either components 1, component 2, or the combination of 1 and 2. Resulting in $2^2 - 1 = 3$ spaces in which k-means clustering is performed. The cluster count k ranged from 2 to 7 clusters for the k-means clustering algorithm.

The time loadings represent metabolite dynamics (therefore plotted as curves), whereas metabolite and diet loadings represent the contribution to the dynamics per metabolite and diet. As an example, a CP component representing a fast dynamic would have a time loading vector representing the fast dynamic, a metabolite loading vector showing larger values for metabolites with fast dynamics, and a diet loading vector indicating in which of the diets the fast dynamics are present. The weight of the CP scores (representing individuals), metabolite and diet loadings are interpreted as PCA scores and loadings.

2.6. Identification of potential metabotypes

Inference of potential metabotypes in the measured data was achieved using three different methods. In the first approach, we investigated associations of individual clusters with the baseline measured clinical parameters using one-way ANOVA. In the second approach, we assessed the biological plausibility of the observed associations between clinical parameters and metabolites (metabolite loadings) contributing most strongly to the clusters. Robustness of clustering was assessed by re-performing metabotyping for 1000 random subsets of the metabolites, selecting from 2 to 79 metabolites in each subset without repeats. To further investigate the predictability and metabolic relevance of the obtained clusters, we also investigated association between clusters and PCA scores of baseline metabolome using ANOVA. Finally, we compared the results obtained from CP decomposition with a PCA on the preprocessed unfolded tensor data in matrix format, i.e., with the individuals (n = 17) in the rows and all other modes combined in the

columns (Fig. 1).

3. Results and discussion

3.1. Preprocessing

3.1.1. Simulated data

From the simulation study, we observed that centering across the individual mode prior to scaling within the metabolite mode (P2) changed the dynamics of the data while the other preprocessing methods did not (Fig. 4). When scaling only within the metabolite mode (P1), static metabolite offsets still remained. Conversely, they were removed when using (P2) or when centering each metabolite slab prior to scaling within the metabolite mode (P3).

3.1.2. Measured data

Not attenuating outliers resulted in non-stable decompositions with components dominated by the outliers rather than the major discernible trends in the data. Attenuation of outliers was therefore deemed effective and consequently performed to improve tensor decomposition and subsequent metabotyping.

In this work, analysis was performed on baseline-subtracted data, which gave clearer metabotyping results in terms of associations to clinical parameters compared to baseline-included data. Baseline-subtraction places a clearer focus on the postprandial dynamics, although as a drawback, it might potentially propagate measurement errors throughout the time series. Results from baseline-included data were considered beyond the scope of this proof-of-concept study but should be explored in future studies.

The effects of using preprocessing methods P1–P3 on measured data were similar to those described for simulated data.

3.2. Rank assessment

3.2.1. Simulated data

After preprocessing of simulated data, three CP components were expected to recover the design vectors as the data was in fact created using three components. This assumption was confirmed for preprocessing methods (P1) and (P2) by visual inspection of the model components. But for preprocessing method (P3), four components were needed. However, it should also be noted that the number of modelling components is not unequivocally clear from the rank assessment plots (Fig. 5): Core consistency dropped markedly after one component using all preprocessing methods, making it dispensable for rank assessment. Although visual inspection indicated three components should be sufficient, using scaling only (P1), four components should be optimal when judging from the stagnation of decrease in the reconstruction error and decrease in stability (Fig. 5A). Using scaling after centering across individuals (P2) or metabolite slab centering (P3), three and four components were suggested, respectively when judging from the stagnation of decrease in the reconstruction error and decrease in stability (Fig. 5B and 5C)

3.2.2. Measured data

Models with 3-components gave the best approximation without degeneracy with 39% and 27% explained variance using preprocessing methods (P1) and (P3), respectively (Fig.6A,6C). A 2-component model was found to be the best approximate model without degeneracy using preprocessing method (P2) with 16% explained variance (Fig. 6B). All models with more components showed degeneracies.

3.3. Clustering and dynamic interpretation

3.3.1. Simulated data

Since all preprocessing methods aimed to normalize the metabolite

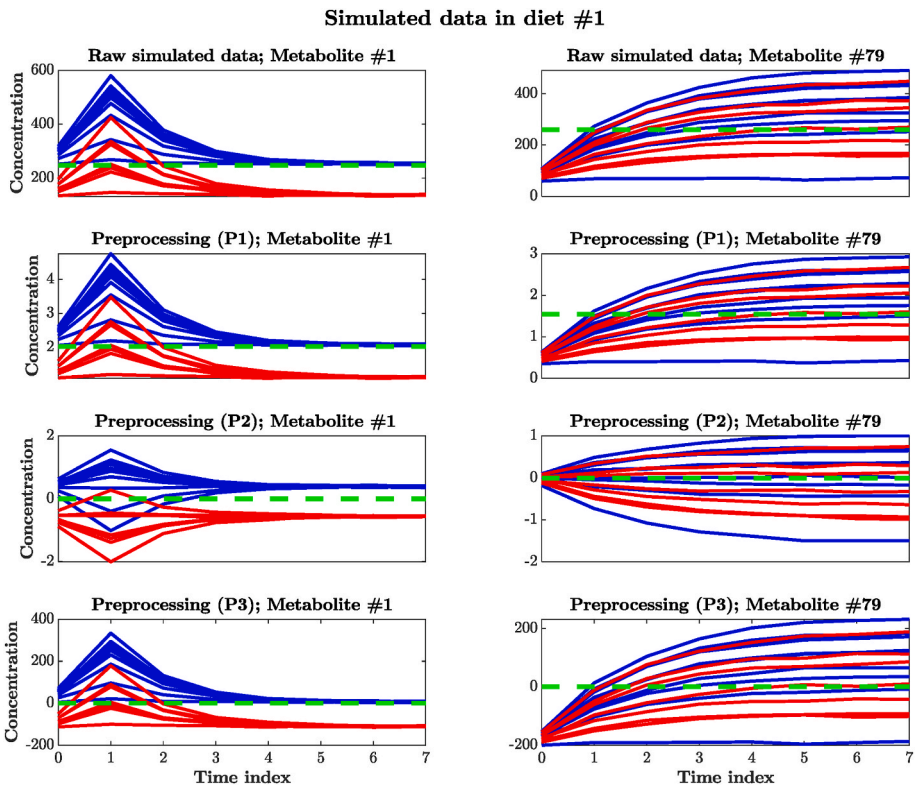


Fig. 4. Raw simulated data and the three preprocessing methods: scaling within the metabolite mode (P1), centering across the individual mode prior to scaling within the metabolite mode (P2) and centering each matrix slab prior to scaling within the metabolite mode (P3),exemplified for two metabolites (#1 and #79)) in the first diet. The red and blue time series represent the two inherent clusters. The overall mean of the matrix is shown as the green dashed line.

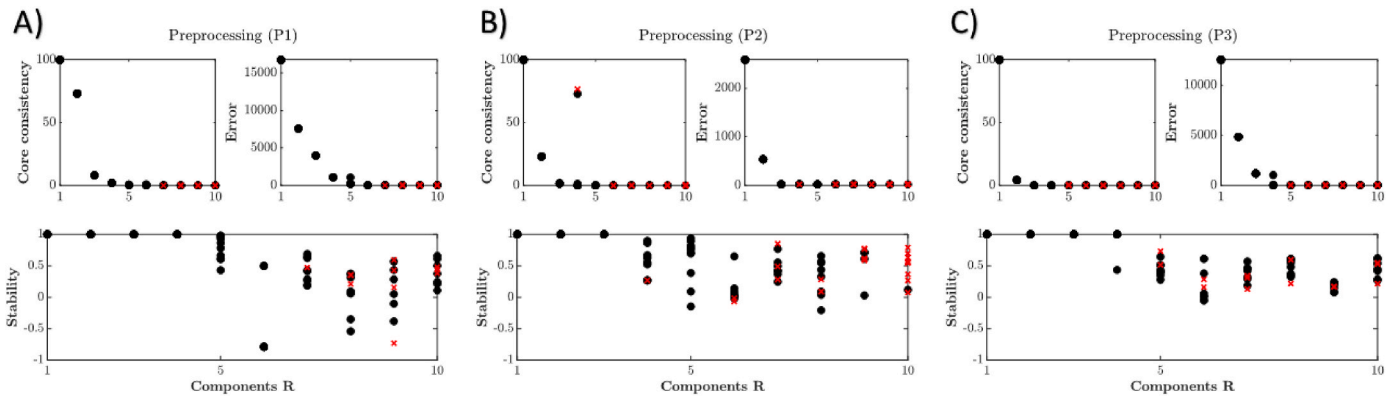


Fig. 5. Heuristic rank assessment results on the three preprocessing methods of the simulated data, i.e. A) scaling within the metabolite mode (P1), B) centering across the individual mode prior to scaling within the metabolite mode (P2) and C) centering each matrix slab prior to scaling within the metabolite mode (P3). For each preprocessing method, core consistency reconstruction error and model stability for CP models having 1–10 number of components are visualized. Black dots indicate valid models while red crosses indicate degenerate models.

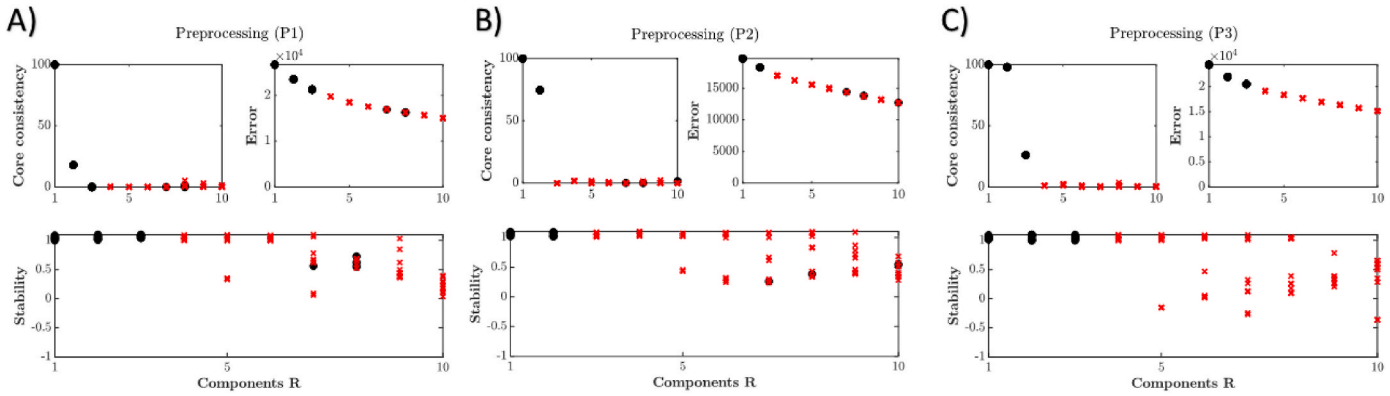


Fig. 6. Heuristic rank assessment results on the three preprocessing methods of the measured data, i.e. A) scaling within the metabolite mode (P1), B) centering across the individual mode prior to scaling within the metabolite mode (P2), and C) centering each matrix slab prior to scaling within the metabolite mode (P3). For each preprocessing method, core consistency reconstruction error and model stability for CP models having 1–10 number of components are visualized. Black dots indicate valid models while red crosses indicate degenerate models.

amplitude, optimally reconstructed design vectors $\boldsymbol{c}$ (Eq. (3)) should be the one-hot encoded vectors for the metabolite mode, representing in which metabolite each dynamic is present (1 if present and 0 if not present).

The 3-component model on the simulated data using preprocessing (P1) recovered the design vectors, although the metabolite loadings were noisy since no metabolite slab offsets were removed with this preprocessing method (Fig. 7A). Three components recovered the design vectors nearly perfectly for preprocessing (P2) (Fig. 7B), but not for preprocessing (P3). However, a 4-component model captured all the

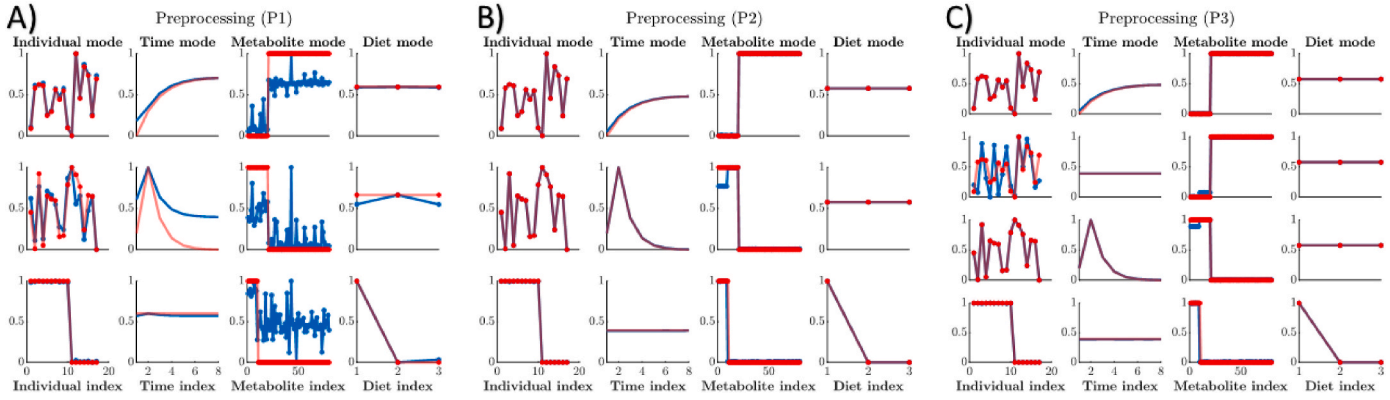


Fig. 7. Ground truth design vectors in transparent red and recovered scores and loadings from the CP models in blue. A) scaling within the metabolite mode (P1), B) centering across the individual mode prior to scaling within the metabolite mode (P2), and C) centering each matrix slab prior to scaling within the metabolite mode (P3). Design vectors are visually most easy to see in red in Fig. 7B where $\boldsymbol{a}_{slow}, \boldsymbol{b}_{slow}, \widehat{\boldsymbol{c}}_{slow}, \widehat{\boldsymbol{d}}_{all}$ are shown in components 1, $\boldsymbol{a}_{fast}, \boldsymbol{b}_{fast}, \widehat{\boldsymbol{c}}_{fast}, \widehat{\boldsymbol{d}}_{all}$ in component 2, and, $\boldsymbol{a}_{clusters}, \mathbb{1}, \widehat{\boldsymbol{c}}_{clusters}, \widehat{\boldsymbol{d}}_{first}$ in component 3, where vectors $\widehat{\boldsymbol{c}}$ and $\widehat{\boldsymbol{d}}$ are one-hot encoded versions of $\boldsymbol{c}$ and $\boldsymbol{d}$.

design vectors but also modeled a constant offset in the slow dynamic metabolites (Fig. 7C, component 2). Clustering using k-means could easily identify the inherent clusters in $\boldsymbol{a}_{clusters}$ from the scores of the models in component 3 using (P1), component 3 using (P2), and component 4 using (P3)

We hypothesized that the constant component (component 2, Fig. 7C) modeled the constant offsets remaining from the centering of each slow dynamic metabolite slab. To show this we considered an arbitrary slow metabolite slab with metabolite index i and diet index j.

$$\begin{aligned}\mathcal{X}_s(:,:,i,j) + \mu_{i,j}\mathbb{1}\mathbb{1}^\intercal &= \boldsymbol{A}\boldsymbol{P}_{ij}\boldsymbol{B}^\intercal + \mu_{i,j}\mathbb{1}\mathbb{1}^\intercal \\ &= \boldsymbol{a}_{fast}\boldsymbol{b}_{fast}^\intercal * \boldsymbol{c}(i)_{fast} * \boldsymbol{d}(j)_{all} + \boldsymbol{a}_{clusters}\mathbb{1}^\intercal * \boldsymbol{c}(i)_{clusters} * \boldsymbol{d}(j)_{first} \\ &\quad + \boldsymbol{a}_{slow}\boldsymbol{b}_{slow}^\intercal * \mathbf{c}(i)_{slow} * \boldsymbol{d}(j)_{all} + \mu_{i,j}\mathbb{1}\mathbb{1}^\intercal\end{aligned}$$

Since $\boldsymbol{c}(i)_{fast}$ and $\boldsymbol{c}(i)_{clusters}$ were both zero due to i being a slow metabolite index, we were left with the matrix (metabolite slab) $\boldsymbol{a}_{slow}\boldsymbol{b}_{slow}^\intercal * \boldsymbol{c}(i)_{slow} * \boldsymbol{d}(j)_{all} + \mu_{i,j}\mathbb{1}\mathbb{1}^\intercal$. We could rewrite the matrix using the centered design vectors where $\overline{\boldsymbol{a}}_{slow}$ and $\overline{\boldsymbol{b}}_{slow}$ defined as

$$\boldsymbol{b}_{slow} = \overline{\boldsymbol{b}}_{slow} + \mu_{b_{slow}}\mathbb{1} \text{ and } \boldsymbol{a}_{slow} = \overline{\boldsymbol{a}}_{slow} + \mu_{a_{slow}}\mathbb{1},$$

yielding

$$\boldsymbol{c}(i)_{slow}\boldsymbol{d}(j)_{all} * \left(\overline{\boldsymbol{a}}_{slow} + \mu_{a_{slow}}\mathbb{1}\right)\left(\overline{\boldsymbol{b}}_{slow} + \mu_{b_{slow}}\mathbb{1}\right)^\intercal + \mu_{i,j}\mathbb{1}\mathbb{1}^\intercal.$$

When we centered the matrix, we removed the overall average

$$\mathbf{M}_{overall} = \boldsymbol{c}(i)_{slow}\boldsymbol{d}(j)_{all}\mu_{a_{slow}}\mu_{b_{slow}}\mathbb{1}\mathbb{1}^\intercal + \mu_{i,j}\mathbb{1}\mathbb{1}^\intercal,$$

and we were left with

$$\mathcal{X}_s(:,:,i,j) + \mu_{i,j}\mathbb{1}\mathbb{1}^\intercal - \mathbf{M}_{overall} = \boldsymbol{c}(i)_{slow}\boldsymbol{d}(j)_{all} * \left(\overline{\boldsymbol{a}}_{slow}\overline{\boldsymbol{b}}_{slow}^\intercal + \mu_{b_{slow}}\overline{\boldsymbol{a}}_{slow}\mathbb{1}^\intercal + \mu_{a_{slow}}\mathbb{1}\overline{\boldsymbol{b}}_{slow}^\intercal\right)$$

This showed that centering the matrix removed the added offset $\mu_{i,j}\mathbb{1}\mathbb{1}^\intercal$ but got replaced by the offset of the design vectors

$$\boldsymbol{c}(i)_{slow}\boldsymbol{d}(j)_{all}\left(\mu_{b_{slow}}\overline{\boldsymbol{a}}_{slow}\mathbb{1}^\intercal + \mu_{a_{slow}}\mathbb{1}\overline{\boldsymbol{b}}_{slow}^\intercal\right)$$

which was then needed to be modeled even after scaling within the metabolite mode, although scaled differently. We consequently hypothesized that the constant component (component 2 Fig. 7C) was modeling the offsets of the design vectors of the slow metabolites. When using this preprocessing method, one thus needs to take into consideration that the number of components needed to model the data will **not** be reduced and that other offsets and noise can be modeled instead of slab offsets.

When the data is constructed using a CP model with additional slab offsets, only scaling the data within the metabolite mode (preprocessing (P1)) can lead to noise in the recovered design vectors, while centering the data across individuals prior to scaling within the metabolite mode (preprocessing method (P2)) can lead to nearly perfectly recovered design vectors. Centering each matrix slab prior to scaling within the metabolite mode (preprocessing (P3)) can lead to a nearly perfect reconstruction of the design vectors while also modeling undesired offsets.

3.3.2. Measured data

Two dynamic profiles in the time loadings were predominantly observed among the measured metabolites: Metabolites with either fast dynamics (primarily amino acids and sugars) or slow dynamics (primarily fatty acids). The fast and slow metabolite dynamics could be identified in the first two component time loadings of the 3-component model on data using preprocessing method (P1) (fast dynamics in the first component and slow dynamics in the second component) (Fig. 8A). Similar dynamics were identified in the first and third components when using preprocessing method (P3) (Fig. 8C). A third dynamic profile was identified in the third, first and second components using preprocessing methods (P1), (P2), and (P3), respectively (Fig. 8A, 8B, 8C). We hypothesize that these components approximate a mixture of two features in the data, one being the fast dynamical profile observed in the time loadings which is present in all the diets, and the other being a time-varying separation between individuals observed in the beef diet. The second component time loadings from preprocessing method (P2) showed a dynamical profile that has not been observed in the data (Fig. 8B).

Centering across the individual mode as in preprocessing (P2) should remove offsets present in each metabolite slab and not affect the loadings, but in theory only center the scores, as was shown in the simulated case (Sec. 3.3.1) (mathematically shown in Supplementary Section S.1) [25]. However, this is only true when the data conforms closely to the CP structure (Eq. (11)), which we believe not to be the case for the

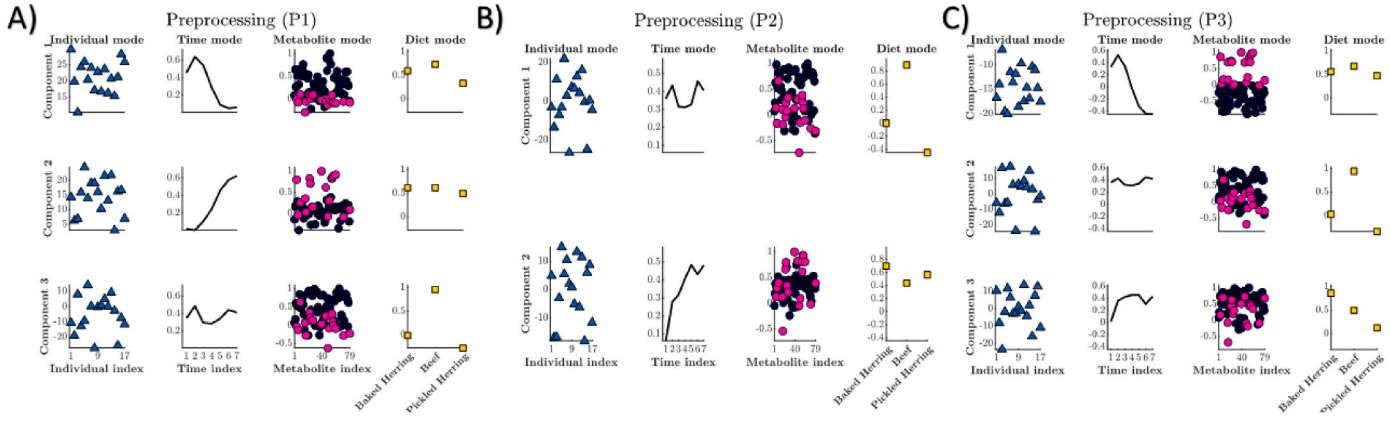


Fig. 8. (A) 3-component model using preprocessing method (P1) on measured data. (B) 2-component model using preprocessing method (P2) on measured data. (C) 3-component model using preprocessing method (P3) on measured data. The pink dots represent the "slow" dynamic metabolites, the black represent "fast" dynamic metabolites.

Table 1
Total number of significant ($p < 0.05$) ANOVA associations between clusterings and clinical variables for all the non-degenerate models of the three preprocessing.

Preprocessing	ALAT	AST	GGT	CHOL	CR	LDL	TSH	BMI
(P1)	0	1	1	0	2	0	0	0
(P2)	0	0	0	0	1	0	0	0
(P3)	0	0	0	0	2	0	0	0

measured data. Thus, removing the average individual from the data arguably emphasizes differences between individuals, which may be beneficial for metabotyping. However, it may also distort the time loadings such that overarching dynamics cannot be recovered (such as the fast and slow dynamic profiles in Fig. 8A and C)

3.4. Clustering of individuals into potential metabotypes from measured data

Clustering of score vectors was performed on all non-degenerate CP models, using all combinations of available factors, thus entailing models with 1–3, 1–2, and 1–3 components using preprocessing methods (P1), (P2), and (P3), respectively. Associations of all clusterings with clinical measures were found using ANOVA and indicated that only aspartate transaminase (AST), gamma-glutamyl transferase (GGT), and creatinine (CR) were significantly linked to the scores of the different models (Table 1). No biological relevance was found between clusters relating to AST or GGT when inspecting the metabolite loadings of the models. However, resulting clusters associated most often and most strongly to creatinine (Table 1), with the strongest associations observed for a clustering obtained both from a 1-component model on data from preprocessing method (P2) (Fig. 9A) and a 2-component model from preprocessing method (P3) (second component Fig. 9B). For clarity, we refer to this clustering as $\boldsymbol{C_c}$.

The two clusters in $\boldsymbol{C_c}$ differed in creatinine (89.2 ± 6.9 (mean $\pm$ sd) vs 78.5 ± 6.4 μmol/L; $p = 0.007$). Furthermore, investigation of loadings indicated that the clustering $\boldsymbol{C_c}$ related predominantly to amino acids in plasma after consuming the beef diet. The cluster with higher levels of creatinine (Fig. 9A and 9B) also had higher levels of several amino acids, especially taurine, phenylalanine, L-allothreonine, threonine, proline, tyrosine (Fig. 9C). Since creatine is derived from amino acids (glycine, L-arginine, and S-adenosyl-*l*-methionine) and is metabolized to creatinine, we speculate that the clustering could be driven by differing absorption and/or metabolism of the amino acid in the assay. This hypothesis was strengthened by the association to distribution in fasting creatinine [33]. Bootstrapping of the metabolites in the measured data prior to clustering was performed to investigate if clustering was due either to an artifact of optimization or possibly due to high correlation between metabolites used in the CP decomposition. Results showed that clustering was remarkably similar using preprocessing methods (P2) and (P3) down to approximately randomly selected 10 variables (Fig. S2), indicating that established underlying patterns in the data corresponding to potential metabotypes were stable. This clustering stability further suggest that the obtained clusters are not likely to reflect artifacts from either algorithm optimization or variable clusters. In addition, preprocessing (P1) notably did not result in clusters with biologically significant associations to clinical data, even though clusterings similar to $\boldsymbol{C_c}$ were found. This result suggests that not taking metabolite offsets into account may impede the potential to infer metabotypes using a CP-based workflow.

We also further investigated the metabolic relevance of dynamic clusters ($\boldsymbol{C_c}$) by assessing whether they could be predicted from the baseline metabolome, represented by PC scores from the metabolome at time point 0. Interestingly, we observed an association in the PC1 score vector in both the meat ($p = 0.0042$) and the pickled herring diet ($p = 0.0458$) where PC1 loadings resembled the CP loadings corresponding to $\boldsymbol{C_c}$, representing the separation between fast and slow dynamic metabolites (Figs. S3 and S4). That the strong association in the meat diet was also reflected in the baseline state from another diet indicates partial robustness of the associations. However, the fact that the dynamic response did not associate to baseline conditions in all diets could reflect that the metabolome baseline is not stable over time or that the clustering $\boldsymbol{C_c}$ was artifactually related to propagation of measurement errors from baseline subtraction. The latter option cannot be fully ruled out, although we deem it more likely that the baseline metabolome was not stable, which was supported by our data (not shown), and also that the association to $\boldsymbol{C_c}$ was observed both in the baseline metabolome measurement of two diets as well as the fasted creatinine levels measured at screening. The results thus indicate that dynamic metabotypes could potentially be predicted from baseline measurements, which could be a cost-efficient and practical tool for *a priori* determination of metabotypes e.g. for personalized health strategies. Albeit with the caveat that baseline metabolome stability over time represents a potential issue.

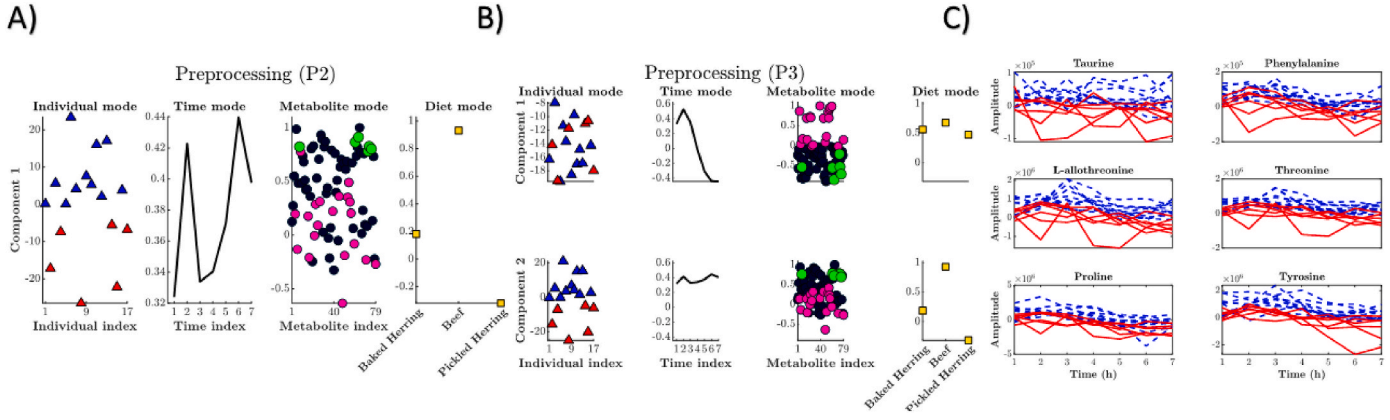


Fig. 9. A) 1-component model using preprocessing method (P2) on measured data. B) 2-component model using preprocessing method (P3) on measured data that resulted in the clustering $\boldsymbol{C_c}$ with the strongest association to clinical data not involved in the C/P and clustering procedure. The clustering $\boldsymbol{C_c}$ was further strengthened from bootstrapping. The pink dots represent the "slow" dynamic metabolites, the black and green dots represent "fast" dynamic metabolites and the green dots represent amino acids. Clusters of scores marked as red and blue triangles associate to baseline creatinine. C) Amino acids time series color-coded by clustering indices from clustering of scores (red and blue) (no preprocessing method applied).

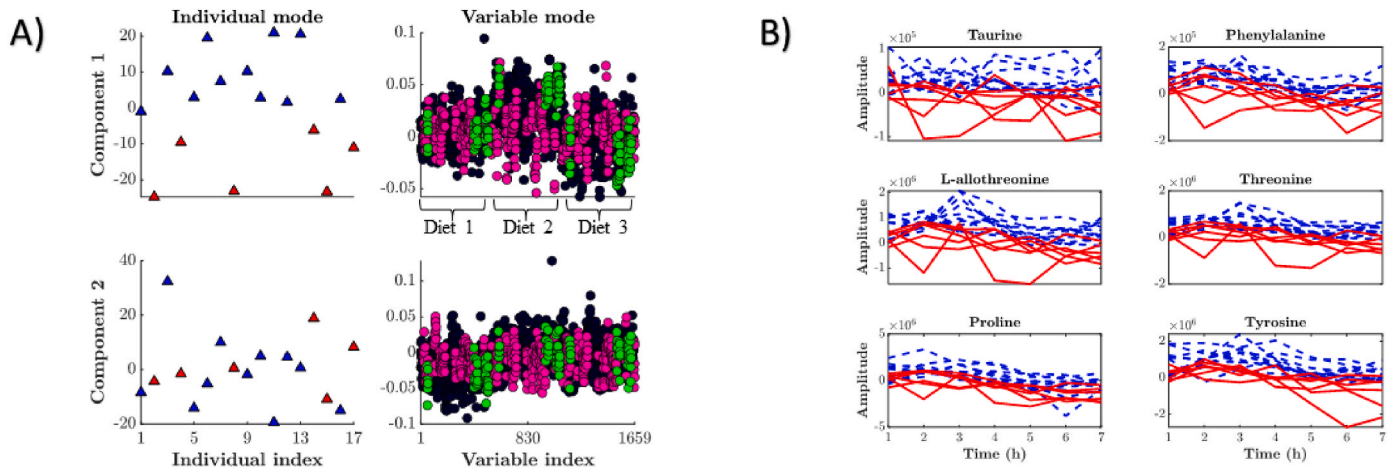


Fig. 10. A) A two-component PCA model on the unfolded baseline-subtracted tensor as in Fig. 1. The scores are color-coded by metabotype 1 (red triangles) and metabotype 2 (blue triangles). Loadings (T x M x D) are color-coded by fast (black), slow (pink) dynamic metabolites and the green dots represent amino acids. B) Amino acids time series color-coded by clustering indices from clustering of scores (red and blue).

3.5. Comparison to PCA

Other methods have been developed to analyze metabolomic time series data to cluster dynamic profiles, like the timeOmics and MetaboClust frameworks [34,35]. However, so far, no framework has reached general acceptance as a field standard, and most methods, including timeOmics and MetaboClust, actually employ PCA for dimension reduction. We therefore compared our results to PCA on the unfolded tensor, partly to show the common model-interpretability between PCA and CP but also the advantage of using CP on tensor data compared to PCA.

As a consequence of the unfolding procedure, PCA imposes several limitations: The modes are not kept intact, which hampers the discovery of overarching patterns, such as the relationship between clusters, diets, and metabolites or the existence of intelligible dynamical profiles inherent in the data. This is a consequence of that each time point of a metabolite is interpreted as a new variable while they ideally should be analyzed as one. Additionally, visualization and interpretability of the loadings become challenging since the loading vector per component contains $(T \times M \times D)$ elements instead of the loading vectors with $(T + M + D)$ total elements in the CP case per component. The loading plots of the PCA is interpreted as the weight of the variable with the absolute value (as in the metabolite and diet loadings of the CP) but the time loadings using CP is far more interpretable for dynamic interpretation since they are separated from the metabolite and diet loadings.

Interestingly, the exact same clustering $\boldsymbol{C_c}$ (as the one found in Fig. 9) was also apparent when analyzing the preprocessed unfolded tensor data using PCA (Fig. 10 A). However, it is difficult to interpret dynamic profiles from the loading plot as time, metabolites and diets are mixed. For instances, it is clearly more challenging to interpret that the cluster $\boldsymbol{C_c}$ is prominent in the amino acids from inspecting the loading plot of the unfolded PCA (Fig. 10 A) compared to CP (Fig. 9). However, the drawbacks with PCA seem to relate predominantly to interpretability, at least in this measured data, since the same clustering $\boldsymbol{C_c}$ was identified from associations to clinical data.

It should be acknowledged that identifying potentially viable metabotypes using unsupervised analysis is a challenging task, especially when data on hard endpoints (such as disease conditions or time to diagnosis) are lacking. As validation-by-proxy, we tested associations between a range of clustering solutions with clinical and anthropometric variables not used for the actual clustering. By doing so, we have introduced the possibility of finding associations due to chance which might not reflect actual metabotypes. However, in this proof-of-concept study, the evidence from associations of clusters to clinical data, biological interpretability and robustness from bootstrapping suggest clustering based on tensor decomposition as a viable approach to infer potential metabotypes from data. We believe that also Tucker decompositions could be relevant for metabotyping in this context, although interpretation becomes more challenging. Since CP is a very restricted yet interpretable model, it may not give a good fit to all tensor data and Tucker decompositions or unfolded PCA could be better alternatives if a better fit is desired.

4. Conclusions

We have investigated CP tensor decomposition as an unsupervised tool to infer metabotypes, i.e., clusters of individuals with similar metabolic profiles and/or regulation, in complex biological data that may arise from Omics studies with e.g., time-resolved data or crossover designs. Although more complex than PCA, CP decomposition uses far fewer parameters than PCA and preserves the data structure better, which greatly enhances the visualization and interpretability of results. We further showed that data preprocessing and tensor rank assessment are critical for CP decomposition. Using both synthetic and measured data, we showed that scaling within the metabolites (P1) or combining scaling either with centering across individuals (P2) or with centering each metabolite matrix (P3) highlight different aspects of the data. Whereas all these preprocessing methods could be considered for CP-based metabotype identification, centering (P2 and P3) presents advantages if the measured metabolites have different offsets in concentration, like e.g. in metabolomics. Both these preprocessing methods can be used to investigate the data for metabotypes. Furthermore, P3 preserves dynamics, thus facilitating interpretation. On the other hand, while distorting this dynamic, P2 can be better suited for focusing on differences between individuals. Subtracting baseline (T0) values from time series data uncovered dynamic profiles better compared to non-subtracted data, thus facilitating potential metabotype discovery. However, this comes at an increased risk of propagating measurement errors. Finally, we found associations between baseline metabolome and clusters representing dynamics, showing the utility of our method to predict dynamic clusters from baseline values. The presented CP-based workflow for metabotyping was demonstrated for fourth order tensor data. However, the workflow is generalizable to tensor data of different

order and resolution, corresponding to complex experimental designs including time-resolved data and crossover interventions.

Author statement

Viktor Skantze: Writing- Original draft preparation, Writing-Reviewing and Editing, Formal analysis, Validation, Investigation, Conceptualization, Methodology, Software, Visualization, Investigation Ann-Sofie Sandberg: Data curation, Carl Brunius: Supervision, Writing-Reviewing and Editing, Conceptualization, Methodology, Mikael Wallman: Supervision, Writing- Reviewing and Editing, Methodology, Mats Jirstrand: Supervision, Writing- Reviewing and Editing, Rikard Landberg: Supervision, Writing- Reviewing and Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

A weblink to the software and data is provided.

Acknowledgments

This work has been supported by the Swedish Foundation for Strategic Research (FID17-0020) and Formas (2016-00314), which are gratefully acknowledged. We thank Rasmus Bro for valuable discussions regarding tensor decomposition.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2022.104733.

References

[1] World Health Statistics. https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/world-health-statistics, 2020.
[2] F. Branca, A. Lartey, S. Oenema, V. Aguayo, G.A. Stordalen, R. Richardson, M. Arvelo, A. Afshin, Transforming the food system to fight non-communicable diseases, BMJ (Online) 364 (2019), https://doi.org/10.1136/bmj.l296.
[3] M. Verma, R. Hontecillas, N. Tubau-Juni, V. Abedi, J. Bassaganya-Riera, Challenges in personalized nutrition and health, in: Frontiers in Nutrition, vol. 5, Frontiers Media S.A, 2018, https://doi.org/10.3389/fnut.2018.00117.
[4] C.L. Bush, J.B. Blumberg, A. El-Sohemy, D.M. Minich, J.M. Ordovás, D.G. Reed, V. A.Y. Behm, Toward the definition of personalized nutrition: a proposal by the American nutrition association, J. Am. Coll. Nutr. 39 (1) (2020) 5–15, https://doi.org/10.1080/07315724.2019.1685332.
[5] J. de Toro-Martín, B.J. Arsenault, J.P. Després, M.C. Vohl, Precision nutrition: a review of personalized nutritional approaches for the prevention and management of metabolic syndrome, in: Nutrients, vol. 9, MDPI AG, 2017, https://doi.org/10.3390/nu9080913.
[6] M. Palmnäs, C. Brunius, L. Shi, A. Rostgaard-Hansen, N.E. Torres, R. González-Domínguez, R. Zamora-Ros, Y.L. Ye, J. Halkjær, A. Tjønneland, G. Riccardi, R. Giacco, G. Costabile, C. Vetrani, J. Nielsen, C. Andres-Lacueva, R. Landberg, Perspective: metabotyping—A potential personalized nutrition strategy for precision prevention of cardiometabolic disease, Adv. Nutr. (2019), https://doi.org/10.1093/advances/nmz121.
[7] A. Riedl, C. Gieger, H. Hauner, H. Daniel, J. Linseisen, Metabotyping and its application in targeted nutrition: an overview. https://doi.org/10.1017/S0007114517001611, 2017.
[8] A. Tebani, S. Bekri, Paving the way to precision nutrition through metabolomics, in: Frontiers in Nutrition, vol. 6, Frontiers Media S.A, 2019, https://doi.org/10.3389/fnut.2019.00041.
[9] E. Ryan, A. Heuberger, C. Broeckling, E. Borresen, C. Tillotson, J. Prenni, Advances in nutritional metabolomics, Current Metabolomics 1 (2) (2013) 109–120, https://doi.org/10.2174/2213235x11301020001.
[10] A. Riedl, N. Wawro, C. Gieger, C. Meisinger, A. Peters, M. Roden, F. Kronenberg, C. Herder, W. Rathmann, H. Völzke, M. Reincke, W. Koenig, H. Wallaschofski, H. Hauner, H. Daniel, J. Linseisen, Identification of comprehensive metabotypes associated with cardiometabolic diseases in the population-based KORA study, Mol. Nutr. Food Res. 62 (16) (2018), https://doi.org/10.1002/mnfr.201800117.
[11] C.B. O'Donovan, M.C. Walsh, C. Woolhead, H. Forster, C. Celis-Morales, R. Fallaize, A.L. Macready, C.F.M. Marsaux, S. Navas-Carretero, S. Rodrigo San-Cristobal, S. Kolossa, L. Tsirigoti, C. Mvrogianni, C.P. Lambrinou, G. Moschonis, M. Godlewska, A. Surwillo, I. Traczyk, C.A. Drevon, L. Brennan, Metabotyping for the development of tailored dietary advice solutions in a European population: the Food4Me study, Br. J. Nutr. 118 (8) (2017) 561–569, https://doi.org/10.1017/S0007114517002069.
[12] T.T.Y. Wang, A.J. Edwards, B.A. Clevidence, Strong and weak plasma response to dietary carotenoids identified by cluster analysis and linked to beta-carotene 15,15'-monooxygenase 1 single nucleotide polymorphisms, JNB (J. Nutr. Biochem.) 24 (8) (2013) 1538–1546, https://doi.org/10.1016/j.jnutbio.2013.01.001.
[13] C. Morris, C. O'Grada, M. Ryan, H.M. Roche, M.J. Gibney, E.R. Gibney, L. Brennan, Identification of differential responses to an oral glucose tolerance test in healthy adults, PLoS One 8 (8) (2013), https://doi.org/10.1371/journal.pone.0072890.
[14] P. Drineas, A. Frieze, R. Kannan, S. Vempala, V. Vinay, Clustering large graphs via the singular value decomposition, Mach. Learn. 56 (1) (2004) 9–33, https://doi.org/10.1023/B:MACH.0000033113.59016.96.
[15] Q. Gu, K. Veselkov, Bi-clustering of metabolic data using matrix factorization tools, Methods (San Diego, Calif.) 151 (2018) 12–20, https://doi.org/10.1016/j.ymeth.2018.02.004.
[16] B. Worley, R. Powers, Multivariate analysis in metabolomics, Current Metabolomics 1 (1) (2013) 92–107, https://doi.org/10.2174/2213235X11301010092.
[17] L. Li, H. Hoefsloot, A.A. Graaf, E. Acar, A.K. Smilde, Exploring dynamic metabolomics data with multiway data analysis: a simulation study. https://doi.org/10.21203/rs.3.rs-526282/v1, 2021.
[18] L.R. Tucker, Some mathematical notes on three-mode factor analysis, Psychometrika 31 (3) (1966) 279–311, https://doi.org/10.1007/BF02289464.
[19] J.D. Carroll, J.J. Chang, Analysis of individual differences in multidimensional scaling via an n-way generalization of 'Eckart-Young' decomposition, Psychometrika 35 (3) (1970) 283–319, https://doi.org/10.1007/BF02310791.
[20] R.A. Harshman, Foundations of the PARAFAC Procedure: Models and Conditions for an 'explanatory' Multi-Modal Factor Analysis, University of California at Los Angeles, 1970.
[21] W.S. Rayens, B.C. Mitchell, Two-factor degeneracies and a stabilization of PARAFAC, Chemometr. Intell. Lab. Syst. 38 (2) (1997) 173–181, https://doi.org/10.1016/S0169-7439(97)00033-6.
[22] V. de Silva, L.-H. Lim, Tensor rank and the ill-posedness of the best low-rank approximation problem, SIAM J. Matrix Anal. Appl. 30 (3) (2006) 1084–1127.
[23] A. Stegeman, Degeneracy in candecomp/parafac explained for $p \times p \times 2$ arrays of rank $p + 1$ or higher, Psychometrika 71 (3) (2006) 483–501, https://doi.org/10.1007/s11336-004-1266-6.
[24] R. Bro, Chemometrics and intelligent laboratory systems Tutorial PARAFAC. Tutorial and applications, in: Chemomemcs and Intelligent Laboratory Systems 38, 1997, pp. 149–171.
[25] R. Bro, A.K. Smilde, Centering and scaling in component analysis, J. Chemometr. 17 (1) (2003) 16–33, https://doi.org/10.1002/cem.773.
[26] A.B. Ross, C. Svelander, I. Undeland, R. Pinto, A.-S. Sandberg, Herring and beef meals lead to differences in plasma 2-aminoadipic acid, β-alanine, 4-Hydroxyproline, cetoleic acid, and docosahexaenoic acid concentrations in overweight men, J. Nutr. 145 (11) (2015) 2456–2463, https://doi.org/10.3945/jn.115.214262.
[27] C.A. Andersson, R. Bro, The N-way toolbox for MATLAB, Chemometr. Intell. Lab. Syst. 52 (1999) 2000–2001. www.elsevier.comrlocaterchemometrics.
[28] C. Svelander, B.G. Gabrielsson, A. Almgren, J. Gottfries, J. Olsson, I. Undeland, A. S. Sandberg, Postprandial lipid and insulin responses among healthy, overweight men to mixed meals served with baked herring, pickled herring or baked, minced beef, Eur. J. Nutr. 54 (6) (2015) 945–958, https://doi.org/10.1007/s00394-014-0771-3.
[29] T.G. Kolda, B.W. Bader, SANDIA REPORT tensor decompositions and applications. http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online, 2007.
[30] A.H. Williams, T.H. Kim, F. Wang, M. Schnitzer, T.G. Kolda, N. Neuroresource, S. Vyas, S.I. Ryu, K.V. Shenoy, S. Ganguli, Unsupervised discovery of demixed, low-dimensional neural dynamics across multiple timescales through tensor component analysis-dimensional neural dynamics across multiple timescales through tensor component analysis, Neuron 98 (2018) 1–17, https://doi.org/10.1016/j.neuron.2018.05.015.
[31] R. Bro, H.A.L. Kiers, A new efficient method for determining the number of components in PARAFAC models, J. Chemometr. 17 (5) (2003) 274–286, https://doi.org/10.1002/cem.801.
[32] P.M. Kroonenberg, Model selection procedures in three-mode component models, Studies in Classification, Data Analysis, and Knowledge Organization (2005) 167–172, https://doi.org/10.1007/3-540-27373-5_20, 0(211289).
[33] H. Taegtmeyer, J.S. Ingwall, Creatine-A dispensable metabolite?, in: Circulation Research, vol. 112, 2013, https://doi.org/10.1161/CIRCRESAHA.113.300974. NIH Public Access.
[34] A. Bodein, O. Chapleur, A. Droit, K.-A. Lê Cao, A generic multivariate framework for the integration of microbiome longitudinal studies with other data types, Front. Genet. 10 (2019). https://www.frontiersin.org/article/10.3389/fgene.2019.00963.
[35] M.J. Rusilowicz, M. Dickinson, A.J. Charlton, S. O'Keefe, J. Wilson, MetaboClust: using interactive time-series cluster analysis to relate metabolomic data with perturbed pathways, PLoS One 13 (10) (2018), e0205968, https://doi.org/10.1371/journal.pone.0205968.