"Comparison of PARAFASCA, AComDim, and AMOPLS approaches in the multivariate GLM modelling of multi-factorial designs"

Guisset, Séverine ; Martin, Manon ; Govaerts, Bernadette

Abstract

Data tables from experimental designs with multiple response variables and a large number of responses with respect to the number of observations are becoming more common, particularly in the omics and chemometrics fields. Traditional multivariate modelling approaches are not suitable for this type of data and more complex approaches like ANOVA Simultaneous Component Analysis (ASCA) and ANOVA Principal Component Analysis (APCA) have been developed to analyse them. They combine a matrix decomposition based on an ANOVA model with a multivariate step, Principal Component Analysis (PCA), which is applied to each decomposed matrix. This article compares ASCA and APCA to three other and less known techniques combining ANOVA and a multivariate step: PArallel Factor - Simultaneous Component Analysis (PARAFASCA), ANOVA Common Dimensions (AComDim), and ANOVA Multi-block Orthogonal Partial Least Squares (AMOPLS). The main advantages of these three methods are their ability to explore in a singl...

Document type : *Document de travail (Working Paper)*

UCL Université catholique de Louvain

INSTITUT DE STATISTIQUE
BIOSTATISTIQUE ET
SCIENCES ACTUARIELLES
(ISBA)

ISBA

DISCUSSION PAPER

2018/30

Comparison of PARAFASCA, AComDim, and AMOPLS approaches in the multivariate GLM modelling of multi-factorial designs

Guisset, S., Martin, M. and B. Govaerts

Comparison of PARAFASCA, AComDim, and AMOPLS approaches in the multivariate GLM modelling of multi-factorial designs

Séverine Guisset[a,*], Manon Martin[b], Bernadette Govaerts[b]

[a] *Statistical Methodology and Computing Support (SMCS), Institute for Multidisciplinary Research in Quantitative Modelling and Analysis (IMMAQ), Université catholique de Louvain (UCL), Voie du Roman Pays 20, bte L1.04.01, B-1348 Louvain-la-Neuve, Belgium*

[b] *Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), Université catholique de Louvain (UCL), Voie du Roman Pays 20, bte L1.04.01, B-1348 Louvain-la-Neuve, Belgium*

Abstract

Data tables from experimental designs with multiple response variables and a large number of responses with respect to the number of observations are becoming more common, particularly in the omics and chemometrics fields. Traditional multivariate modelling approaches are not suitable for this type of data and more complex approaches like ANOVA Simultaneous Component Analysis (ASCA) and ANOVA Principal Component Analysis (APCA) have been developed to analyse them. They combine a matrix decomposition based on an ANOVA model with a multivariate step, Principal Component Analysis (PCA), which is applied to each decomposed matrix.

This article compares ASCA and APCA to three other and less known techniques combining ANOVA and a multivariate step: PArallel Factor - Simultaneous Component Analysis (PARAFASCA), ANOVA Common Dimensions (AComDim), and ANOVA Multi-block Orthogonal Partial Least Squares (AMOPLS). The main advantages of these three methods are their ability to explore in a single and global procedure all (or a part) of the model effects, extract automatically the most important ones, interpret them visually and quantify their respective importance and significance. Theses approaches are presented in a common framework to make their comparison easier, some enhancements are introduced and they are also extended to unbalanced experimental designs thanks to a GLM version of the matrix decomposition.

The three methods are applied to a ^{1}H-NMR data set with a three-factor unbalanced experimental design. Quantitative and visual results are compared for the three techniques and show that they are all suitable for the analysis of a complex model, but differ in the type of outputs and the ease of interpretation of the results. This opens perspectives for the analysis of even more complex data, including models with random or quantitative effects.

Keywords: ASCA, APCA, multivariate, ANOVA, PARAFASCA, AComDim, AMOPLS, experimental design

1. Introduction

1.1. Background

Experimental designs consist in varying factor levels in an experiment in order to assess the impact of the factors on a response of interest. This type of data can be analysed using standard analysis techniques, such as Analysis of Variance (ANOVA), multiple linear regression or response surface analysis. ANOVA is traditionally applied when the factors of interest are all categorical. However, more complex approaches are

*Corresponding author

Email addresses: severine.guisset@uclouvain.be (Séverine Guisset), manon.martin@uclouvain.be (Manon Martin), bernadette.govaerts@uclouvain.be (Bernadette Govaerts)

needed to handle experimental data sets in the presence of multiple response variables, especially when the number of responses m is much larger than the number of observations N.
Such data sets will be referred to as Wide Multi-Response Experimental Design (WMRED) data in this article. They are commonly found in the omics field, which includes but is not limited to metabolomics, transcriptomics, proteomics and genomics. They are also commonly encountered in chemometrics when a set of spectral analytical data is obtained from a designed experiment.
Several approaches that can model factor effects while being applicable to WMRED data will be presented next.

1.2. Overview of techniques combining ANOVA modelling and multivariate data analysis when $m >> N$

Figure 1 presents the family of multivariate techniques suitable for the analysis of WMRED data that will be discussed in this paper. All techniques include a decomposition of the initial data matrix (or response matrix) into model effect matrices, followed by a multivariate dimension reduction step which allows exploring the relationship between experimental design factors and the observations or variables in the data set. Visualisations of observations and variables are available in all techniques, while some include the computation of importance measures for the factor effects and related significance tests.

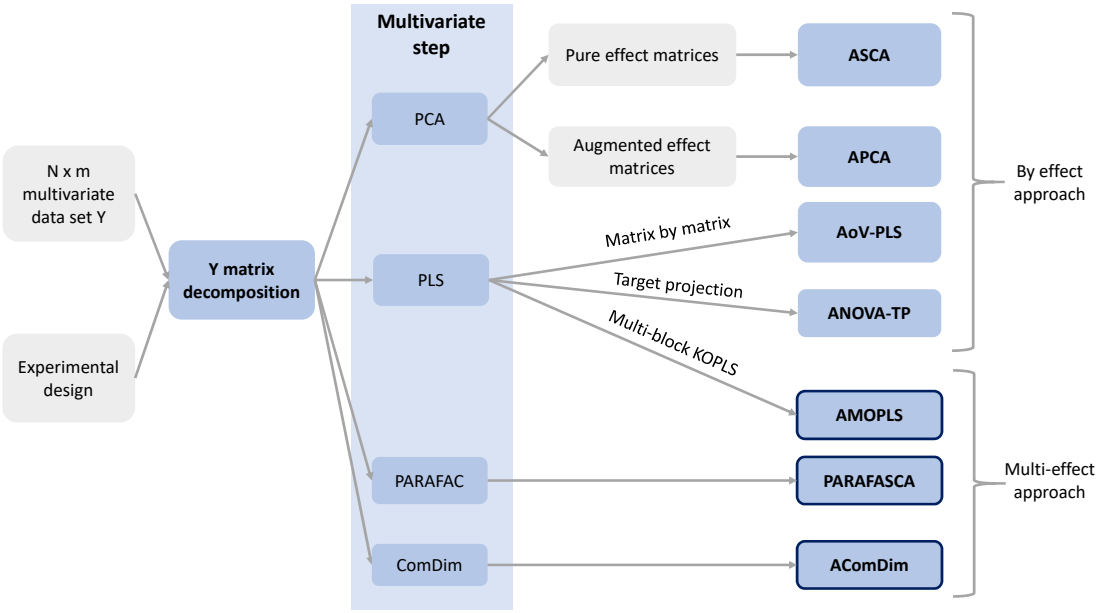

Figure 1: Overview of techniques combining ANOVA modelling and multivariate data analysis

ANOVA Simultaneous Component Analysis (ASCA) and ANOVA Principal Component Analysis (APCA) [1, 2] are the best known methods in this context and are widely used to analyse omics data. In these two approaches, a decomposition of the original data matrix in different effect matrices corresponding to each term of an ANOVA model is computed first, followed by a multivariate analysis of each effect matrix separately with Principal Components Analysis (PCA). The difference between the two techniques is that ASCA

applies the PCA to "pure" effect matrices (directly obtained in the decomposition) and APCA to "residual-augmented" effect matrices, i.e. the pure effect matrices added to the model residual matrix.

Several variants of ASCA and APCA have since been suggested in the literature. Analysis of Variance Partial Least Squares (AoV-PLS) applies PLS to the matrices resulting from the ANOVA decomposition [3]. In this approach, an individual PLS regression is implemented on each effect matrix individually. The PLS response matrix is the pure effect matrix and the predictor matrix is the residual-augmented effect matrix. If the PLS regression identifies the factor levels in the residual-augmented effect matrix, the effect of interest can be considered more important than the noise in the data.
ANOVA with target projection (ANOVA-TP) consists in applying the ANOVA decomposition followed by a PLS with target projection [4]. The PLS model is applied separately for each effect, with the corresponding residual-augmented effect matrix as the predictor matrix, and a matrix or vector of factor levels as the response. Once the model has been estimated, a target projection rotation is applied to ease interpretation. This consists in projecting the data matrix onto the regression vector(s) of the optimal PLS model in order to identify one (or several) latent variable(s), the target-projected component(s).
These four techniques allow analysing WMRED data, but are all based on an individual study of each model effect, which can be tedious when the model is complex. In contrast, this article will focus on three other approaches - PArallel Factor - Simultaneous Component Analysis (PARAFASCA), ANOVA Common Dimensions (AComDim), and ANOVA Multi-block Orthogonal Partial Least Squares (AMOPLS) - where the multivariate step allows to analyse and interpret all or several model effects with an integrated approach.

1.3. ASCA/APCA+, an extension of ASCA/APCA to unbalanced experimental designs

The ANOVA decomposition presented in the ASCA/APCA literature is not suitable for unbalanced designs as it is based on effect estimators that become biased when the design is not balanced. Thiel et al. [5] suggest an alternative approach based on the General Linear Model (GLM): ASCA/APCA+. This method replaces the ANOVA decomposition in ASCA/APCA by a GLM-based decomposition which provides identical results to an ANOVA decomposition in the balanced case, and unbiased estimators in the unbalanced case. This GLM framework also has the advantage (in comparison with the ANOVA approach) to be very general since it is not based any more on specific variance decomposition equations linked to each possible model. This allows for an easier computation of the effect matrices. GLM also offers a potential extension to models including quantitative or random effects.

1.4. Objectives

The overall aim of this paper is to present and compare possible ASCA/APCA-based techniques that are applicable to WMRED data with balanced or unbalanced designs and allow a multivariate analysis of all or several model effects in one global approach.
More precisely, this paper presents within one single framework and with common notations the three global techniques listed in Figure 1 : PArallel Factor - Simultaneous Component Analysis (PARAFASCA), ANOVA Common Dimensions (AComDim), and ANOVA Multi-block Orthogonal Partial Least Squares (AMOPLS) and combines them with the ASCA/APCA+ GLM approach in order to allow their extension to unbalanced designs. This article shows that each of these methods has specific valuable or problematic properties and offers several tools to order, interpret and represent graphically model effects results. The article also suggests some enhancements of the methods (e.g. testing procedures and new types of graphical outputs), clarifies some aspects that were not detailed in the original articles and suggests simplifications where possible.
The GLM extension of ASCA/APCA to unbalanced experimental designs and a review of ASCA/APCA are presented before outlining the three techniques of interest. These are then applied to a ^{1}H-NMR unbalanced data set, compared to each other and to ASCA/APCA approaches.

2. GLM version of ASCA/APCA

2.1. Overview

This section introduces a general framework for ASCA/APCA-like methods based on an effect matrix decomposition followed by a dimension reduction multivariate step.
Figure 2 shows the steps of ASCA/APCA approaches. In a first step, the original response data matrix Y is decomposed in effect matrices $\hat{M}_f$ according to a model linked to the experimental design. In the standard ASCA/APCA method, an ANOVA model is used, but this approach can be generalised using a General Linear Model (GLM) framework. The main difference between the ANOVA and GLM approaches is the parameter estimation method. The GLM extends linear regression and ANOVA to cases with several response variables, different types of independent variables and unbalanced designs. A GLM can be written $Y = X\Theta + E$ with Y the response matrix, X the model matrix, Θ the model parameter matrix and E the residual matrix.
In a second step, the decomposed effect matrices are analysed using a dimension reduction multivariate approach. PCA is used for ASCA and APCA, while AComDim, PARAFASCA and AMOPLS are based on other multivariate techniques that will be described later.

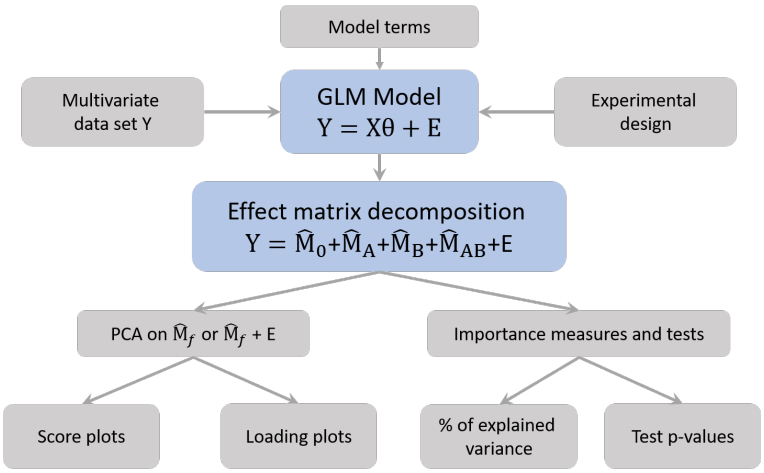

Figure 2: Overview of the GLM-based approach

2.2. Matrix decomposition methodology

The response matrix decomposition consists in decomposing the response matrix Y into matrices that each corresponds to the estimation of a given effect in a model adapted to the original experimental design. In Figure 2, the model is an ANOVA 2 model with two crossed factors A and B. Such an ANOVA 2 model with one response variable can be expressed using the "cell means model" [6] as: $y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}$ where y_{ijk} is the value of the response variable for the i-th level of factor A, the j-th level of factor B and the k-th observation, with i=1 to a, j=1 to b and k=1 to N. μ is the mean value of the response variable and ϵ_{ijk} is the error term usually supposed to be $\epsilon_{ijk} \sim iN(0, \sigma^2)$. α_i and β_j are the main effects for factors A and B and $\alpha\beta_{ij}$ is the interaction term.
In the standard ANOVA approach used in ASCA/APCA, the estimators for the model effects are based on mean differences: $y_{ijk} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + (\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}) + (y_{ijk} - \bar{y}_{ij}) = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\alpha\beta}_{ij} + e_{ijk}$
The following constraints are verified for these estimators when the design is balanced : $\sum_{i=1}^{a} \hat{\alpha}_i = 0$, $\sum_{j=1}^{b} \hat{\beta}_j = 0$, $\sum_{i=1}^{a} \hat{\alpha\beta}_{ij} = 0 \ \forall j$ and $\sum_{j=1}^{b} \hat{\alpha\beta}_{ij} = 0 \ \forall i$.
These estimators can then be gathered together column by column (one column per response) in effect matrices $\hat{M}_f$ and the response matrix decomposed as $Y = \hat{M}_0 + \hat{M}_A + \hat{M}_B + \hat{M}_{AB} + E$ where $\hat{M}_0$ contains the overall means, $\hat{M}_A$ corresponds to the main effects of factor A, $\hat{M}_B$ to the main effects of factor B, $\hat{M}_{AB}$

to the interactions between A and B, and E to the model residuals.
In the GLM approach used in ASCA/APCA+, the $N \times m$ response matrix Y is first written as the product of the $N \times p$ model matrix X and the $p \times m$ parameter matrix Θ, plus the $N \times m$ error matrix E [5]: $Y = X\Theta + E$. m is the number of variables in the response matrix, p is the number of parameters in the model and X may be built according to the model and design using a sum or deviation coding as explained in [5].

The model parameters are then simply estimated by ordinary least squares

$$\widehat{\Theta} = (X'X)^{-1}X'Y \tag{1}$$

and the effect matrices M_f can each be estimated by multiplying the block X_f of the model matrix X corresponding to the effect of interest by the corresponding block $\widehat{\Theta}_f$ of the $\hat{\Theta}$ matrix of estimated parameters:

$$\hat{M}_f = \left(\begin{array}{c|c|c} 0 & X_f & 0 \end{array} \right) \hat{\Theta} = X_f^* \widehat{\Theta}_f \tag{2}$$

The ANOVA and GLM approaches lead to the same parameter estimators for balanced designs and therefore perform equally well, but the GLM approach also allows estimating the model parameters without bias for unbalanced designs [5].

2.3. Multivariate dimension reduction step in ASCA/APCA+

Following the matrix decomposition, PCA is applied to each of the decomposed matrices. The major difference between ASCA and APCA is the matrix used for this step. In ASCA, the PCA is applied to the pure effect matrices $\hat{M}_f$ directly obtained from the ANOVA/GLM decomposition. In APCA, the PCA is applied to the residual-augmented effect matrices $\tilde{M}_f = \hat{M}_f + E$.
Due to this difference, ASCA scores plots only present mean scores for each factor, while APCA scores plots also include individual variations between observations not explained by the model. This is illustrated on a case study in Figure 3, which will be detailed in Section 5.

Note that in ASCA, PCA may be applied to the pure effect matrices either, very simply, by performing one separate PCA decomposition of each effect matrix or by applying a Simultaneous Components Analysis (SCA) on a vertical concatenation of all effect matrices (see [7, 8]). For simplicity purposes, the first approach is used in this article.

2.4. ASCA/APCA(+) model diagnostics and output interpretation

A range of outputs are available following an ASCA/APCA(+). They are explained in this subsection and will be implemented on the case study in Section 5.
The percentage of variance explained by each main effect or interaction in the model helps evaluating their relative importance. [5] explain how to derive these percentages in a general way on the basis of Type III sum of squares and Frobenius norms.
To test the significance of each effect, the following test statistic is defined for an effect f by extension of the ANOVA F-tests to the multivariate framework: $F_f^{pseudo} = \frac{\|E_{/f}\|^2 - \|E_{full}\|^2}{\|E_{full}\|^2}$ where E_{full} is the error matrix of the model with all effects included and $E_{/f}$ is the residuals matrix of the model estimated with all effects but effect f. The numerator of the statistic is a generalisation of Type III sum of squares in univariate GLM models. On the basis of this test statistic, a permutation approach allows computing p-values. Several permutation methods are described in the literature and must be adapted to the type of effect and model of interest (see [9, 10] and Section 5.4).
The loadings and scores obtained from the PCA decomposition of the pure or residual-augmented effect matrices provide rich information to interpret the result of the ANOVA/GLM modelling. Three types of scores plots can be obtained and are illustrated in Figure 3 with example plots on the effect of the three-level Hippurate factor in the case study of this article (further described in Section 5).
In ASCA(+), PCA is applied to pure effect matrices and the scores plots only display a few data points corresponding to the model parameters linked to the effect of interest and allow to see how each (combination

of) factor level(s) is situated with respect to the others. In the ASCA scores plot presented in Figure 3, the data points corresponding to the three levels of the Hippurate factor are mostly spread along the first PC, which explains 97.7% of the variation of the Hippurate effect matrix, while the second component only explains 2.3% of the variation. This means that the effect of Hippurate on the response matrix is mostly linear over these three levels, with a small second order effect (see details in Section 5).

In APCA(+), when PCA is applied to a residual-augmented effect matrix, the scores plots show how the factor levels and the N observations are situated with respect to each other when only the corresponding effect and the residuals are kept in the model and all the information related to the other model effects has been removed from the data. In Figure 3, the APCA scores plot shows that 88.7% of the variance is explained by the first component and separates the three Hippurate levels. The second PC explains 4.5% of the variance and is linked to model residuals. More residual information is included in further components. No other specific structure appears in these data.

A third type of scores plot, suggested by [11] and referred to as "ASCA-E" in [5], allows including in the ASCA procedure a direct comparison of the factor levels with the residuals, as in APCA. This approach can be advantageous when the percentage of variation due to the residuals in the model is high because, in this case, the APCA scores plots may show a dispersion of the samples along PC1 due to noise while the separation of the factor levels is along PC2 or PC3.

The ASCA-E scores plot consists in first applying the ASCA approach by conducting a PCA on the pure effect matrices and then on projecting the residual-augmented effect matrices onto the principal components (PCs) obtained. Thanks to this technique, the components' interpretation is richer and the variation in the data is taken into account in the analysis. In Figure 3, the ASCA-E scores include individual variation on both axes as expected but caution must be taken in its interpretation: the horizontal variation observed on the first PCA component is of interest since this component explains 97.7% of the model effect variation. However, the second axis is of limited interest since it is linked to a very small second order effect of Hippurate and does not provide the best representation of the variation of the model residuals.

The loading plots extracted from the PCA decomposition of the pure or the residual-augmented effect matrices are important to identify which of the m original variables are linked to each effect in the model. In Figure 3, the loading plots corresponding to the first PC in ASCA and APCA clearly display the four expected Hippurate peaks (see Section 4 for more details). In general, ASCA loadings should be preferred to APCA ones because they are less affected by residuals and will show a purer image of the model effect for each of the m variables.

In Section 5, loadings plots will be calculated with the ASCA approach and scores plots with the ASCA-E approach.

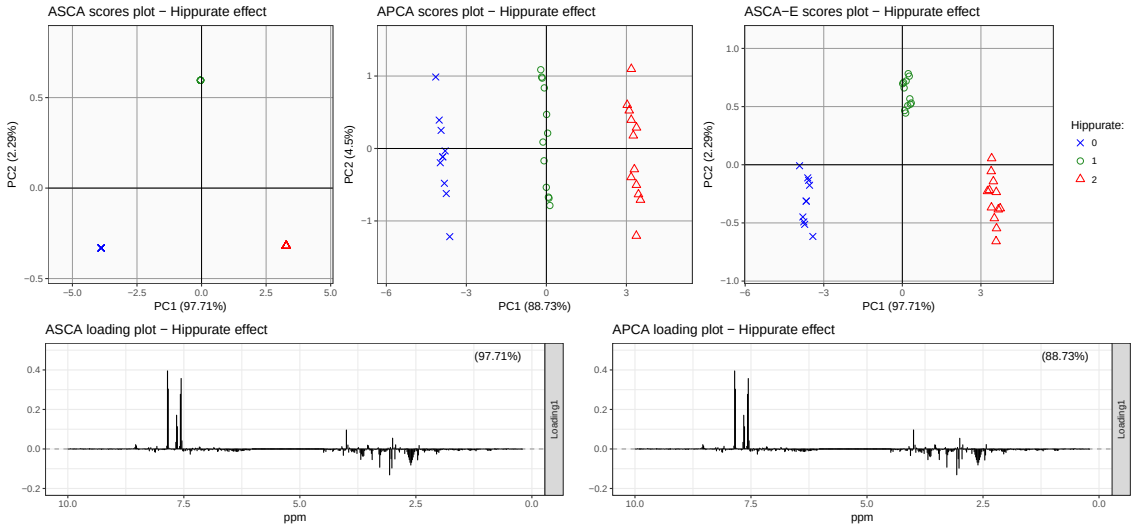

Figure 3: Exemples of scores and loadings for ASCA, APCA and ASCA-E for the effect of the three-level Hippurate factor in the case study of this article (see Section 5)

3. PARAFASCA, AComDim, and AMOPLS

The three methods are presented in chronological order of their publication in the literature.

3.1. PARAFASCA

PARAFASCA is described in Jansen et al. [12]. In this article, the technique presented consists in implementing a PARAFAC following the ANOVA decomposition instead of applying one PCA to each effect matrix as done in ASCA/APCA. PARAFASCA can be seen as complementary to ASCA: it can help visualise the complex model effects in a more interpretable and parsimonious way.

The original article does not provide the detailed steps of a general technique as PARAFAC was applied to a three-factor repeated measure design where only the main effect of one factor and the interaction effect of this factor with another are of interest in the full three-way ANOVA model. The corresponding model coefficients are folded into a three-way array with the three dimensions (or modes) corresponding to the two factors of interest and the vector of m response variables. PARAFAC is then applied to this array in order to provide a suitable way to visualise and interpret the set of effects of interest.

The approach presented in the current paper is more generic: it shows how PARAFASCA can be applied to a full multi-factor crossed ANOVA model using the GLM approach in keeping the spirit of [12]. It also illustrate, on the case study, the interest of applying PARAFASCA on a partial model.

3.1.1. Approach

PARAFAC is a multivariate technique to analyse D-way arrays. First suggested in [13], the technique was again put forward by [14] under the name Canonical Decomposition (also called CANDECOMP) and [15] under the name PARAFAC. The technique is sometimes referred to as CP, which stands for CANDECOMP/ PARAFAC [16].

PARAFAC is conceptually a generalisation of PCA on a $N \times m$ matrix with 2 "modes" (observations and variables) to arrays with D modes. PARAFAC provides a multilinear decomposition of the array into R sets of D vectors, where R is the number of components extracted. The vectors can be interpreted as scores and loadings as in PCA but since they are treated the same numerically, the word used to name them is unimportant and will depend on the context [17]. The corresponding loadings/scores matrices can be computed by Alternating Least Squares (ALS) or other algorithms which are discussed in more details in [18] and [19].

PARAFAC is used in a range of scientific fields among which chemistry, biology, medicine, environmental and computer sciences. In particular, it is widely used in chemometrics to analyse higher dimension data obtained by spectroscopy, chromatography or NMR [20, 18].

PARAFASCA has been applied to metabolomic data obtained by NMR spectroscopy in the original article [12], while approaches close to PARAFASCA and based on ANOVA and PARAFAC have been used to analyse multi-way fluorescent probe spectroscopy anaesthetics data in [21] and three-way descriptive sensory analysis data in [22].

3.1.2. Model

The first step of PARAFASCA is based on the same principle as the matrix decomposition in ASCA(+). It consists in estimating the ANOVA or GLM model parameters for each of the m response variables but without building the complete effect matrices M_f. The next step of PARAFASCA applies a PARAFAC decomposition to a selected combination of model terms according to the questions of interest in the study and the structure of the design (crossed, nested etc...). Such decomposition is especially suited to models containing complex interaction terms. The existing articles related to PARAFASCA do not present a generic methodology but illustrate the principles on case studies. On this basis, we present below a generic way to apply PARAFASCA to the ANOVA 2 model discussed in Section 2 when no prior knowledge is available on the way to combine model effects.

After the GLM model estimation, the idea is to combine and fold in a three way array Y_G all the parameters estimated in the ANOVA/GLM model except for the constant term (see Figure 4). Y_G is of size $a \times b \times m$ and the element y_{ijh} of Y_G is calculated as follows:

$$y_{ijh} = \hat{\alpha}_{ih} + \hat{\beta}_{jh} + \hat{\alpha\beta}_{ijh}$$

where $i = 1, ...a$, $j = 1, ...b$ are the factor levels and h is the response of interest with $h = 1, ...m$.

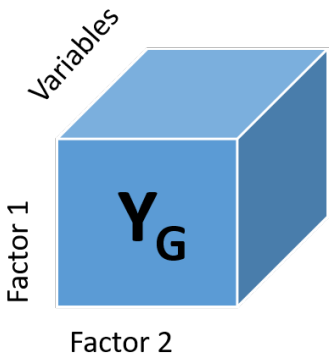

Figure 4: Example of a PARAFASCA array

The PARAFAC decomposition can then be written as [17, 23]:

$$y_{ijh} = \sum_{r=1}^{R} u_{ir} v_{jr} w_{hr} + e_{ijh} \tag{3}$$

or

$$Y_G = \sum_{r=1}^{R} u_r \Delta v_r \Delta w_r + E_G \tag{4}$$

where R is the number of computed PARAFAC components with r=1 to R. U, V and W are the three loadings/scores matrices of the PARAFAC decomposition with elements u_{ir}, v_{jr} and w_{hr}, and dimensions $a \times R$, $b \times R$ and $m \times R$ respectively. u_r, v_r and w_r are vectors and the r-th columns of U, V and W respectively. E_G, with elements e_{ijh}, is the residual array of the PARAFAC model, of the same dimension as Y_G. Δ denotes a tensor product.
Figure 5 presents a schematic visualisation of the three-way PARAFAC decomposition of matrix Y_G with R components and a residual array E_G.

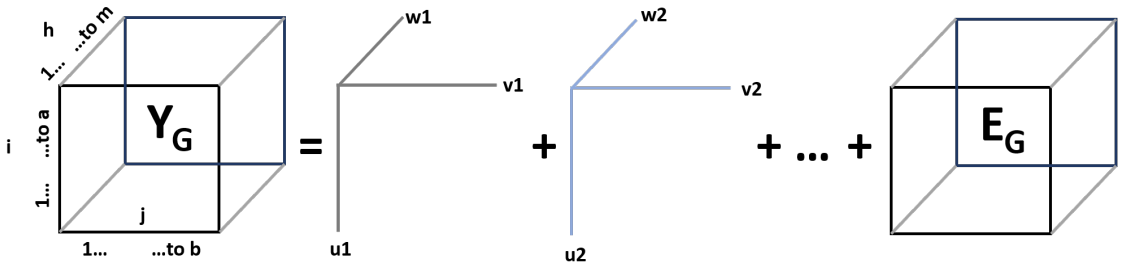

Figure 5: Visualisation of a three-way PARAFAC decomposition of matrix Y_G with R components and a residual array E_G

The details of an Alternating Least Squares (ALS) algorithm which can be used to compute PARAFAC components are presented in Appendix 7.1.
Different techniques are available to identify the optimal number of components to extract in a PARAFAC analysis and are discussed in more details in [17], [24] and [25]. The split-half technique cannot be applied in this context as none of the array dimensions corresponds to the observations. The number of components can however be selected based on a scree plot and confirmed by the interpretation of the results. Please note that a high number of components may lead to unstable algorithm results, with different effects corresponding to different components in each run of the same model.
Some ALS algorithms also allow to force the orthogonality of the components on one or several modes. In the PARAFASCA context, a reasonable approach is to impose orthogonality between the loadings of the Dth mode. This prevents model terms mixing and makes their interpretation easier.

3.1.3. Output

The following outputs can be derived from a PARAFASCA analysis. For a chosen number of components R, a (marginal) percentage of explained variance can be computed for each component r of the PARAFAC decomposition. If a given component can be linked to one of the model effects, this percentage provides information on the relative size of the effect in the ANOVA model. Please note that if components are not constrained to be orthogonal in the ALS algorithm, these marginal percentages of variance explained may be difficult to interpret because they do not sum to the total variance explained by the decomposition.

$$\mathrm{PC}_r = \frac{\|u_r \Delta v_r \Delta w_r\|^2}{\|Y_G\|^2} \times 100$$

A scree plot presenting the total percentage of variance explained by models with different numbers of components R can also be used to select the number of components to be included in the model. This approach will be applied in Section 5.
The scores/loadings matrices U, V and W can then be represented graphically in order to interpret the parameters of the m ANOVA/GLM models. Mean effect scores available in matrices U and V allow to interpret the main and interaction effect of factors on the response and the contribution of the model terms to each component. The loading matrix W of size $m \times R$ may be represented column by column to visualise the link between the response variables and each PARAFAC component. Section 5 will show that it is

convenient to represent the scores either by mode, to illustrate which mode or combination of modes is linked to each component, and then, to represent for each component the tensor product of scores in parallel to the corresponding loading vector (mode D) to understand which response changes are linked to the possible combinations of factor levels.

3.2. AComDim

AComDim is described in Jouan-Rimbaud Bouveresse et al. [26].

3.2.1. Approach

AComDim applies the ComDim approach to the matrices obtained from the ANOVA/GLM matrix decomposition. AComDim is based on a Singular Value Decomposition (SVD) of **sample** "variance-covariance matrices" calculated on the residual-augmented effect matrices $\tilde{M}_f$. It aims at comparing the variation of the response due to the factor levels of the experimental design with the residual variability [26] and therefore provides information on the importance of specific factor effects and interactions.

ComDim finds its origins in the Common Components and Specific Weight Analysis (CCSWA) approach which was first used to analyse sensory and chemometric data linked to food [27, 28, 29]. The method was then applied to other contexts and renamed Common Dimensions [30].

ComDim aims at analysing simultaneously several data matrices of size $N \times m_i$ by finding a common space of representation, with each matrix having a specific weight (or salience) λ_f^r in each of the orthogonal common dimensions or common components (CC) of this space [29]. The technique relies on the weighted sum of **sample** variance-covariance matrices or association matrices. This is different from many multivariate approaches where the **variable** variance-covariance matrix is used.

AComDim has been applied to the analysis of duplicate apple reflectance spectra, NMR relaxation curves of starch-lignin, fluorescence and Mid-InfraRed spectroscopies of wine in the original article [26]. The technique has also been used on near infrared sensors spectral data [31], irradiated resins spectral data [32] and metabolic indicators of olive oil [33].

3.2.2. Model

The first step of AComDim is the ANOVA/GLM matrix decomposition. In a second step, the residual-augmented effect matrices $\tilde{M}_f$ $(f = 1, ..., F-1)$ and the residual matrix E (written $\tilde{M}_F$ below) are column-centred - which is already the case for a balanced design - and normalised at the matrix level through a division by their Frobenius norm. This normalisation ensures that the matrices are of similar magnitude and avoids the dominance of one matrix over the others [26]. ComDim is then simultaneously applied to this set of residual-augmented effect matrices $\tilde{M}_f$ (including the residual matrix E) to compute common components and related effect saliences.

To do this, these centred and normalised matrices $\tilde{M}_f$ $(f = 1, ...F)$ are used to build related $N \times N$ **sample** variance-covariance matrices Ψ_f defined as $\tilde{M}_f \tilde{M}_f'$. This approach ensures that the resulting matrices $\Psi_f = \tilde{M}_f \tilde{M}_f'$ all have the same $(N \times N)$ dimensions and can be summed, even if the matrices $\tilde{M}_f$ include a different number of variables m_i (although this is not the case with effect matrices).

ComDim decomposition is then computed dimension by dimension in order to minimise the following final criteria:

$$\sum_{f=1}^{F} ||\Psi_f - \lambda_f^r q_r q_r'||^2 \tag{5}$$

where each q_r is obtained from an SVD and $\lambda_f^r = q_r \Psi_f q_r'$.

It is then possible to summarise the final results in the following global formulae [34, 26]. Each sample variance covariance matrix is firstly decomposed as

$$\Psi_f = \tilde{M}_f \tilde{M}_f' = \sum_{r=1}^{R} \lambda_f^r q_r q_r' + E_f = \Psi_f^* + E_f \tag{6}$$

and, written globally, as

$$\Psi_G = \sum_{f=1}^{F} \Psi_f^* + E_G = Q\Gamma Q' + E_G \tag{7}$$

where Q is the $N \times R$ matrix $(q_1, q_2, ..., q_R)$. Γ is a diagonal matrix with $\Gamma_{rr} = \sum_{f=1}^{F} \lambda_f^r$ and $E_G = \sum_{f=1}^{F} E_f$.

The details of the algorithm used to compute common components and saliences are presented in Appendix 7.2.

3.2.3. Output

The percentages of explained variance for each common dimension r show their relative importance. They can be calculated as [27, 28]:

$$\frac{\mu_r}{\sum_{f=1}^{F} ||\Psi_f||^2} \text{ where } \mu_r = \sum_{f=1}^{F} (\lambda_r^f)^2 \tag{8}$$

A scree plot presenting the cumulative percentage of variance explained by the common dimensions can be used to select the number of common dimensions to visualise. This approach will be applied in Section 5.

For each common dimension r, the saliences λ_r^f help to interpret the link between the dimension and the model effects. A high salience for an effect matrix on a given common dimension indicates that this dimension contains a large amount of information linked to the corresponding model effect. The sum of saliences for a given matrix over all dimensions tends to 1, which means that saliences can be interpreted as proportions. The first common dimension is linked to the largest common source of variance in the data. This is usually the residuals as they were added to each effect matrix to produce the residual-augmented effect matrices. The residual matrix is thus expected to have a high salience on the first common component.

The scores and loadings computed in AComDim can be interpreted in the same way as for other techniques in the same family. The scores for each dimension are simply defined as the singular vectors q_r. They link common dimensions with design factor levels and allow visualising single observations.
Loadings link common components with the original variables. The loadings are not precisely defined in the original AComDim article and different options are available to calculate and visualise them. In [26], the authors suggest in their application to compute loading vectors for dimension r as the product $p_r = \tilde{M}_f' q_r$ where f is the model effect with maximum salience λ_f^r for dimension r. In this paper, we suggest using a more general definition which takes into account effect matrices for all model terms weighted by their respective saliences: $p_r = \sum_{f=1}^{F} \sqrt{\lambda_f^r} \tilde{M}_f' q_r$. These loadings will provide profiles similar to the previous ones when one salience is really dominant. Case studies show that this is often the case for components which are not linked to noise.
[26] also suggest an F-test to test the significance of the main effects and interactions in the ANOVA model based on the fact that saliences can be considered proportions of variance and that the first common dimension is linked to the residuals. A ratio of the saliences of an effect matrix $\tilde{M}_f$ and the residual matrix on the first common dimension can therefore be computed to assess how much noise the effect matrix contains:

$$\frac{\lambda_1^{residuals}}{\lambda_1^f} = \frac{q_1' E E' q_1}{q_1' \Psi_f q_1} \tag{9}$$

This ratio is by definition always greater than or equal to one and, according to [26], a F distribution with $((N-1),(N-1))$ degrees of freedom can be used to provide a p-value to test the significance of the corresponding model effect. The null hypothesis of the test is that the model term has no effect on the response and the p-value can be calculated as $P(F_{N-1;N-1} < \lambda_1^{residuals}/\lambda_1^f)$. If the p-value is small, the term is considered significant, which indicates that some of its related parameters are not null.
It is however important to highlight that no mathematical justification is given in [26] to justify the fact that the test statistics follows a $F_{N-1;N-1}$ distribution under H_0. In Section 5 of this article, a permutation test will be suggested to compute an alternative p-value.

3.3. AMOPLS

AMOPLS is described in Boccard and Rudaz [35].

3.3.1. Approach

AMOPLS consists in applying a kernel-based multi-block OPLS approach on the pure and residual-augmented matrices issued from the ANOVA decomposition of the design response matrix and provides, as output, a global view of important effects in the model which allows their interpretation.

PLS is considered the "standard method in supervised linear modelling in the field of chemometrics" [36]. It models the relationship between a predictor matrix X and a response matrix Z and is suited to WMRED data sets. OPLS is an extension of PLS which uses an orthogonal signal correction filter in order to distinguish Z-predictive (PrC) and Z-orthogonal (OC) components in the analysis. The resulting predictive components therefore only include variation related to Z, which simplifies the interpretation of the results [37, 36].

AMOPLS is a supervised method where, the response matrix Z is defined in order to guide the choice of OPLS predictive components that are directly linked to significant model effects. Orthogonal components gather the data variability due to the residuals and non significant model effects.

AMOPLS relies on a kernel-based version of OPLS, KOPLS, suggested in [36]. In KOPLS, the predictor matrix can be mapped (by a kernel function) to a higher dimensional space in order to model potential non-linear relationships. The resulting kernel matrix K is of size $N \times N$. In AMOPLS, however, the kernel transformation is only used to base the OPLS decomposition on sample variance-covariance matrices, such as the ones used in AComDim.The multi-block characteristic of AMOPLS comes from its ability to analyse several matrices simultaneously [35].

It will be shown below that AMOPLS presents several similitudes with AComDim even if the general approach seems very different. Finally, it is important to note that the AMOPLS algorithm is applicable to balanced or unbalanced designs, but the structure of its response matrix implies that it is recommended to use it only for balanced or almost balanced designs to produce valuable results.

3.3.2. Model

Figure 6 presents the AMOPLS approach. Following the response matrix decomposition by ANOVA/GLM modelling, the effect matrices are used to compute a predictor matrix and a response matrix subsequently used in a KOPLS regression.

The detailed steps of the technique described in [35] and [38] are as follows:

1. Compute the matrix decomposition by ANOVA/GLM modelling: $Y = \hat{M}_0 + \sum_{f=1}^{F-1} \hat{M}_f + E$
2. Compute the $N \times (p-1)$ KOPLS response matrix Z as a concatenation of non-null score matrices extracted from each pure effect matrix.
   - Apply a SVD to each pure effect matrix $\hat{M}_f$ for f=1 to $F-1$: $\hat{M}_f = Q_f \Gamma_f V_f'$.
   - Compute the score matrix for each effect $Q_f \Gamma_f$ and gather the first non null columns in M_f^*. M_f^* is of size $N \times p_f$ where p_f is the number of model parameters linked to effect f in the GLM model.
   - Concatenate the resulting matrices to create the KOPLS response matrix Z:

     $$Z = \begin{pmatrix} M_1^* & M_2^* & \dots & M_{F-1}^* \end{pmatrix}$$

     The final number of variables of the response matrix Z is equal to $p-1$, the number of model parameters minus 1. Z is an orthogonal matrix when the design is balanced. When the design in unbalanced, some or all blocs M_i^* of Z are not orthogonal to each other anymore which may decrease the interpretability of the final results.
3. Compute the $N \times N$ KOPLS predictor matrix Ψ_f on the basis of the augmented effect matrices aggregated in a similar way to AComDim:
   - Compute the residual-augmented effect matrices $\tilde{M}_f = \hat{M}_f + E$ for f=1 to $F-1$.

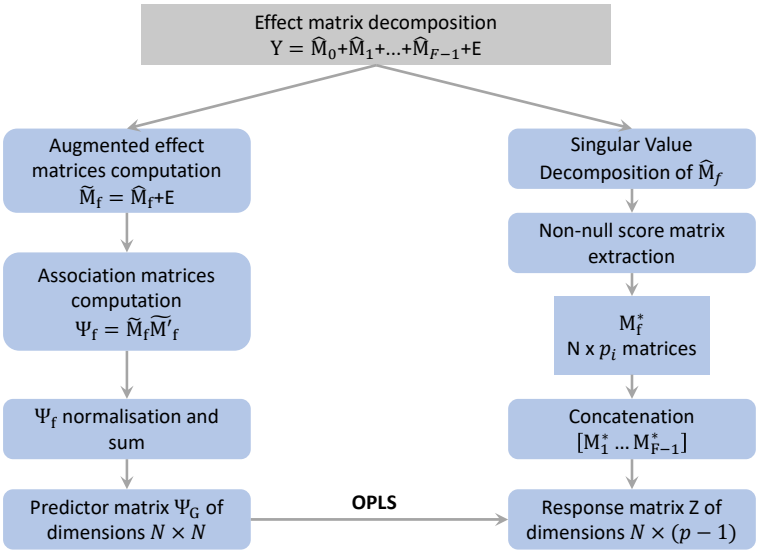

Figure 6: Overview of the AMOPLS approach

- Transform the F residual-augmented effect matrices in sample variance covariance matrices or "association" matrices $\Psi_f = \tilde{M}_f \tilde{M}'_f$ of size $N \times N$. This transformation is equivalent to a polynomial kernel transformation of order 1 in KOPLS.
- Divide each Ψ_f matrix by its Frobenius norm.
- Define the KOPLS predictor matrix Ψ_G as the sum $\Psi_G = \sum_{f=1}^{F} \Psi_f$ of these normalised matrices.

4. Fit the KOPLS model:
   - Identify the optimal number of orthogonal components for the OPLS of Z on Ψ_G in order to remove most of the non-significant information from the predictor matrix. The selection of the number of orthogonal components is discussed in Subsection 5.2.
   - Set the number of predictive components as the number of columns of the predictor matrix Ψ_G, i.e. $p-1$
   - Run a KOPLS regression explaining the response matrix Z by the predictor matrix Ψ_G. The KOPLS algorithm in [35] is simply an OPLS where the predictive matrix is the kernel transformation Ψ_f of the initial matrix.

3.3.3. Output

For a given number of components r, the percentage of explained variance can be computed in AMOPLS via the classic OPLS goodness-of-fit indicators $R2Y$ and $R2X$. $R2X$ gives a measure of the percentage of variance of Ψ_G explained by r OPLS components and $R2Y$ of the percentage of variance of Z explained by Ψ_G through these components.

A scree plot presenting the total percentage of variance explained for models with different numbers of

predictive components can be used to select the number of components of interest in the final model. This approach will be applied in Section 5.

The contribution of each model effect f to component r can be computed as $\delta_f^r = \frac{t'_r \Psi_f t_r}{\sum_{f=1}^{F} t'_r \Psi_f t_r}$ where t_r is the orthogonal or predictive score vector and $\sum_f \delta_f^r = 1$.

These contributions can also be used to build a measure of model term importance, the RSR (for *Residual Structure Ratio* [35]). For effect matrix f the RSR is defined as: $RSR_f = \delta_E^{o1}/\delta_f^{o1}$, the ratio between the contribution of the error to the first orthogonal component and the contribution of the effect f to this component. This ratio follows the same spirit as the F ratio defined in AComDim: since the first orthogonal component in AMOPLS represents the noise in the data which is orthogonal to the model effects, it is expected to be very similar to the first common component of AComDim. The contribution of an effect f to t_{o1}, δ_f^{o1} will then be close to δ_E^{o1} if f has no effect in the model and smaller than δ_E^{o1} when f has an effect. As in AComDim, a permutation test based on the RSR can be set up to assess the effect significances.

Scores and loadings can also be computed for each OPLS component and visualised and interpreted as for the other techniques. In [35], loadings are computed using the same approach as in the original AComDim article [26], but this paper uses the general form suggested in Section 3.2.3: $p_r = \sum_{f=1}^{F} \sqrt{\delta_f^r} \tilde{M}'_f t_r$.

4. Data

The data set used to illustrate the methods in the next section is a subset of the Urine-Citrate-Hippurate data set described in [39]. It has specifically been chosen in order to highlight and compare the properties of the different approaches.

The data contain metabolomic profiles obtained by ^{1}H-NMR spectroscopy of a pool of rat urine samples. These samples were prepared by spiking the pool with known concentrations of chemicals (Hippurate and Citrate) in order to obtain data based on an experimental design. Classic NMR data pretreatments steps were applied to the original FID's prior to statistical analysis (see details in [39]). The spectra have been reduced to 600 descriptors and normalised with a constant sum normalisation. This traditional approach in metabolomics is suitable here since an appropriate quantity of buffer has been added to the samples in order to avoid a decrease of urine signal in the presence of the spiked molecules. No variable scaling was applied as, in such NMR data, scaling can artificially inflate noisy spectral areas. Please note that scaling in the ASCA context must be applied with caution, taking into account the considerations raised in [40].

In the subset of this data set used here, and further referred to as the UCH data set, each of these factors has three levels: no added chemical (a concentration of 0), a medium concentration (2 mM for the Citrate factor (Qc/4) and 1 mM for the Hippurate factor (Qh/4)) and a high concentration (4 mM for the Citrate factor (Qc/2) and 2 mM for the Hippurate factor (Qh/2)). This leads to an experimental design with 9 possible combinations of concentrations illustrated in Figure 7(a). It is important to highlight than a linear increase in concentration usually leads to a linear increase in spectral intensity when using NMR. This means that the Hippurate and Citrate effects can be expected to be almost one-dimensional in the ANOVA model in spite of having three levels each.

Four samples of each of the nine combinations were prepared, then frozen and analysed (by sets of 18) over two days. Each day, the analysis of the two samples with the same Hippurate and Citrate proportions was done under two conditions: just after defreezing and one hour later in order to evaluate stability of samples over time. The UCH data selected for the paper originally contained 36 spectra: 9 combinations of concentrations $\times$ 2 time points $\times$ 2 replicates. However, two outliers were removed, resulting in 34 observations and a slightly unbalanced experimental design (see Figure 7(b)).

The resulting (34×600) spectral matrix is defined as the response matrix Y. In these spectra, 4 Hippurate peaks are located around 3.96, 7.54, 7.62 and 7.82 ppm and the aggregated Citrate peak is located around 2.6ppm [41] (see Figure 7(c)). Figure 7(d) presents the scores plot of the first two PCs for the response matrix. It shows that the information on the Citrate and Hippurate factor levels from the experimental design can be recovered, but only to some extent.

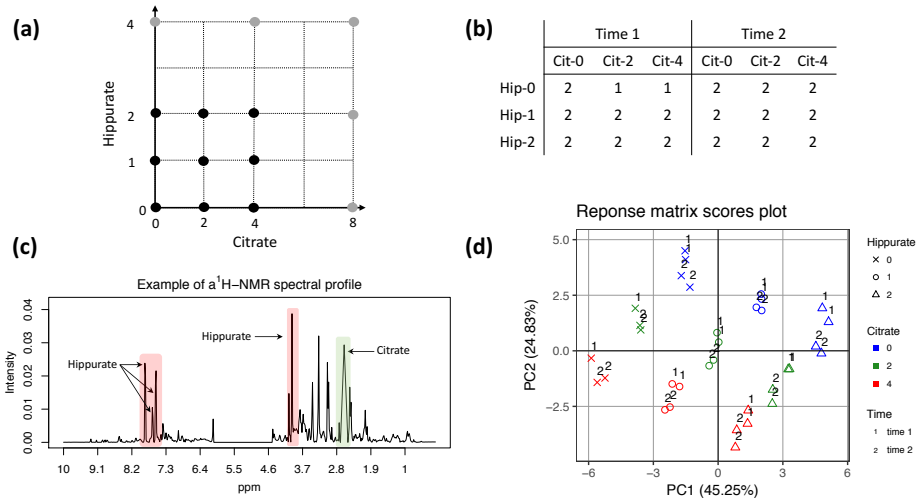

Figure 7: Presentation of the UCH data set (a) Experimental design (b) Sample sizes for each of the 9 combinations of Citrate, Hippurate and Time levels (c) Example of a ^{1}H-NMR spectral profile showing the Hippurate and Citrate peaks (d) Response matrix scores plots for the first two PCs of the UCH data set

5. Results

ASCA(-E), PARAFASCA, AComDim and AMOPLS were applied to the UCH data set. The results for each method are presented in parallel below with a particular effort to facilitate their comparison.

5.1. GLM decomposition

For all approaches, the following 3-way crossed ANOVA model was first fitted to the UCH data set. It includes three main factors, the concentrations of Hippurate (α) and Citrate (β) and the Time (γ), and four interactions, Hippurate-Citrate (H×C), Hippurate-Time (H×T), Citrate-Time (C×T), and Hippurate-Citrate-Time (H×C×T).

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} + \epsilon_{ijkl}$$

The model parameters and effect matrices were estimated by the GLM approach in order to take into account the unbalanced property of the design. For each response l ($l = 1, ...m$) this model contains 17 independent parameters (constant term excluded). The first useful result is displayed in Table 1: based on the type III Sum of Squares, it represents the percentage of variation (squared Frobenius norm) of the response matrix Y explained by each model effect in the GLM decomposition. It shows that the three main effects correspond to a meaningful percentage of variation in the data, the H×T interaction to a small but notable percentage and the three other interactions to a very small percentage of variation. Please note that, because the design is unbalanced, the effect matrices are not orthogonal to each other and the total variation of Y cannot be exactly decomposed into percentages of variation summing to 100% [5].

Table 1: Percentage of type III Sum of Squares for each effect matrix from the GLM decomposition

Hippurate	Citrate	Time	H:C	H:T	C:T	H:C:T	Error	**Total**
39.31	29.91	16.24	1.54	6.23	0.54	1.68	4.30	**99.74**

5.2. Components data decomposition and measure of component importance

All four approaches include factorial decompositions of combinations of the GLM effect matrices. This section presents, for each approach, the percentage of variation of the data explained by the calculated components and show that scree plots are a good complement to model term significance tests to decide how many components are informative in each model. It must be stressed that these percentages of variance cannot directly be compared across approaches because the initial decomposed matrices and decomposition algorithms are not identical.

For ASCA, a PCA decomposition was applied separately to each pure effect matrix $\hat{M}_f$ and to the residual matrix E. Table 2(a) shows that, for the three main effects, one component explains most or all of the variation. This was expected for the Time because it only has two levels (i.e. one degree of freedom). For the three-level Hippurate and Citrate factors, this is an indication that the main effect is probably linear and also has only one degree of freedom. The PCA decomposition of the H×T interaction is also almost one-dimensional and since Table 1 showed that other interaction terms are not very important in the model, the PCA decompositions of the corresponding interaction matrices are therefore not expected to contain substantial information. Keeping one component for each model effect can thus be considered enough in further steps. Finally, the residuals' decomposition shows that at least two PCs are necessary to study their behaviour. In conclusion, for ASCA, keeping 8 components, each directly linked to one model effect or to the two first residuals PCs, should allow understanding most of the information available in the data. Please note that these 8 components are not all orthogonal to each other because of the unbalanced structure of the design.

Table 2: Percentage of variance of the decomposed matrices explained by each dimension for ASCA, PARAFASCA, AComDim and AMOPLS. The total percentages of variance cannot directly be compared accross approaches because the initial decomposed matrices are not identical.

	Hippurate	Citrate	Time	H:C	H:T	C:T	H:C:T	Error
PC 1	97.71	98.22	100	44.01	93.92	90.76	47.23	48.54
PC 2	2.29	1.78	0	38.51	6.08	9.24	27.49	16.90
Total of $\hat{M}_f$	**100**	**100**	**100**	**82.52**	**100**	**100**	**74.72**	**65.44**

(a) ASCA-E effect matrix variance for PC 1 and PC 2.

comp. 1	comp. 2	comp. 3	comp. 4	**Total of Y_G**
38.03	28.27	16.64	11.63	**94.58**

(b) PARAFASCA variance decomposition for components 1 to 4.

CC 1	CC 2	CC 3	CC 4	CC 5	CC 6	**Total of Ψ_G**
20.44	21.54	20.39	17.63	2.435	8.543	**90.98**

(c) AComDim variance decomposition for common components 1 to 6.

R2X	R2X	R2X	R2X	R2X	**R2X**	**R2Y**	
OC 1	OC 2	PrC 1	PrC 2	PrC 3	PrC 4	**Total of Ψ_G**	**Total of Z**
26.84	12.31	8.13	7.79	7.85	6.62	**69.54**	**94.09**

(d) AMOPLS X and Y variance decomposition for orthogonal and predictive components.

PARAFASCA was applied in the spirit of the methodology described in Section 3.1. The centred model predictions for all factors levels $\hat{y}_{ijkl} - \hat{\mu}$ were gathered in a $(3 \times 3 \times 2 \times m)$ Y_G matrix and a 4-mode PARAFAC decomposition was applied to Y_G with different numbers of components R.

Figure 8(a) shows, on a scree plot, the percentage of variance of Y_G explained by models with 1 to 17 components (17 being the number of parameters in the model). It shows that a model with four components is best suited to these data. Table 2(b) shows the percentage of variation explained by each component. Please note that, as suggested in Section 3.1.3, an orthogonality constraint has been imposed on Mode 4 in the decomposition.

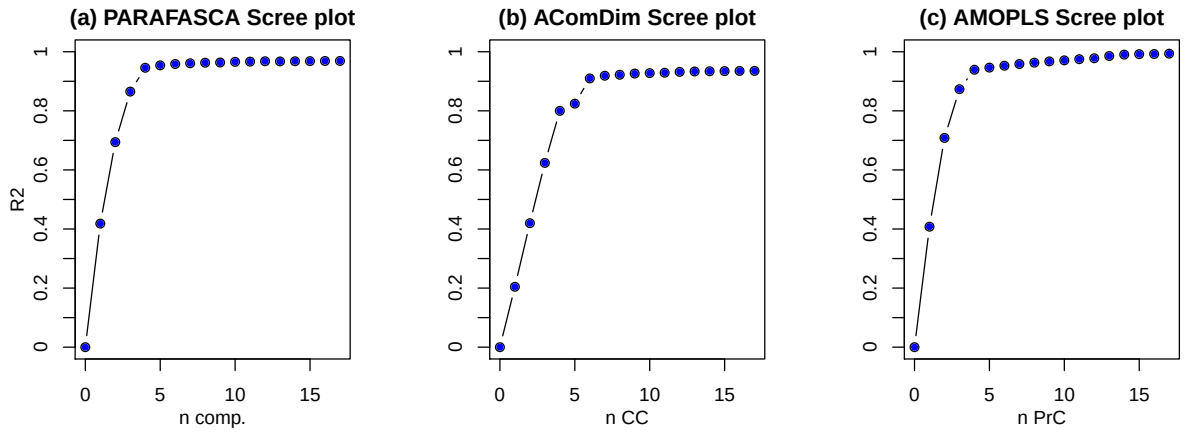

Figure 8: Scree plots based on R2 to select the optimal number of components for PARAFASCA (4 components), AComDim (6 Common Components) and AMOPLS (4 Predictive Components)

For AComDim, a screeplot was also used to visualise the information provided by each additional dimension. It shows that 6 components are very informative and explain together 91% of the information of the aggregated effect matrix Ψ_G.

The choice of the number of components in AMOPLS was done in two steps. First, a number of orthogonal components needed to be selected. [35] suggest applying a permutation test to decide how many orthogonal components to include but this is time-consuming for complex designs. It was therefore decided to reflect the dimensionality of the data while applying the principle of parsimony. The ASCA results show that the error is bi-dimensional and the authors of [35] advise selecting a small number of orthogonal components. A model with two orthogonal components was therefore chosen. Next, the maximal number of predictive components was set to 17, the number of parameters in the model, but their interpretation in the next sections will focus on the first four on the basis of the scree plot shown in Figure 8(c).

5.3. Components interpretation

In ASCA, the principal components calculated on each effect matrix are directly linked to the model effects and can then further be interpreted via their scores and loadings (see Sections 5.5 and 5.6). This section shows how this link can be visualised and quantified for the three approaches of interest.

Figure 9 presents the PARAFASCA (mean) scores for the three components and three modes A, B and C, corresponding respectively by construction to Hippurate, Citrate and Time factors. These results clearly show the link between scores for component 1 and the Hippurate effect, scores for component 2 and the Citrate effect and scores for component 3 and the Time effect. The PARAFAC decomposition recovers then perfectly the three main effects in the model and the response loadings for the first three components will therefore be easily interpretable. These scores also allow highlighting the linearity of the Hippurate and Citrate effects. The results for component 4 are less straightforward to interpret but point towards an H×T interaction effect. This is expected because this interaction is the 4th more important term in the GLM model (see Table 1) and because, in Figure 9, the Hippurate and Time effects are steep and corresponding scores change sign, while the Citrate effect is smaller and the corresponding scores remain positive. Its interpretation will be complemented by the scores and loadings plots presented in subsections 5.5 and 5.6.

Both AComDim and AMOPLS provide an index to quantify the contribution of the model effects to each dimension of the data space decomposition. These indices allow interpreting directly the loadings and scores in terms of their link with the model effects. They are represented in Figure 10 for the two methods. In AComDim, the contributions are based on the saliences and in AMOPLS the contributions are obtained with a formula similar to the AComDim saliences (see Section 3.3.3). For these two methods, a careful interpretation of the contributions in terms of proportions is possible: in AComDim, saliences sum to 100%

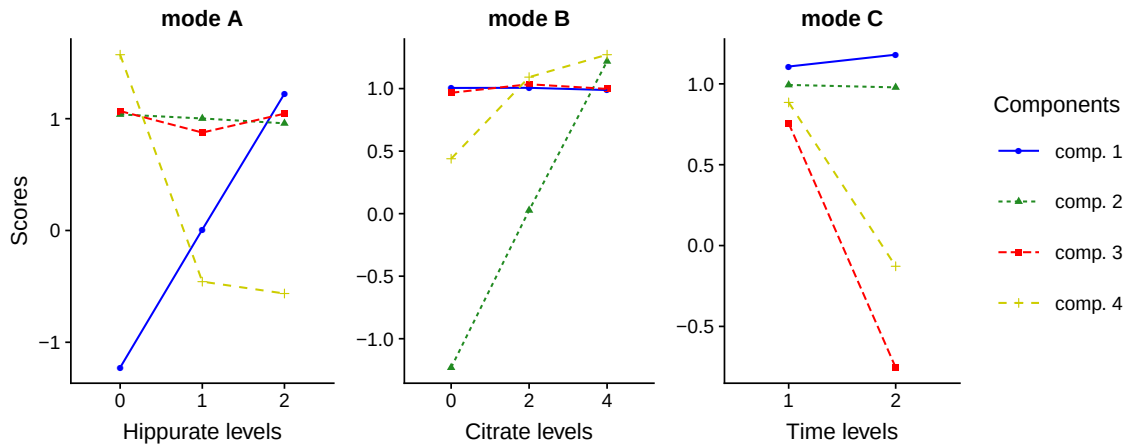

Figure 9: PARAFASCA scores for modes A, B and C.

for each model effect (over all dimensions) whereas the contributions are normed to one within a component in AMOPLS.

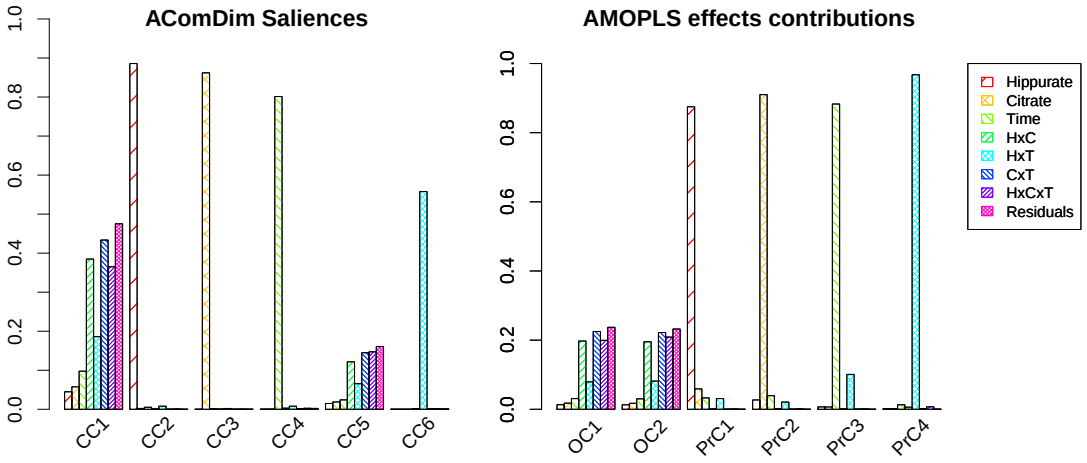

Figure 10: Model effects and residuals contributions for each AComDim and AMOPLS predictive components.

In AComDim, the first component is related to the residuals. This result could be expected because, in this method, the error matrix E appears eight times in the decomposed data matrix - one time in each residual-augmented effect matrix and one time alone - and therefore has a high weight. E thus represents "common information" in each data block used in the analysis.

The next three components clearly correspond to the Hippurate, Citrate and Time main effects. This ordering is consistent with the percentage of variance explained by these terms in Table 1. Component 5 seems to contain more information concerning model residuals, while component 6 clearly corresponds to the H×T interaction effect.

In AMOPLS, the two orthogonal components are linked to the residuals and their effect contribution profiles

are very similar to what is observed in AComDim. This is expected since the residual matrix E is orthogonal to the KOPLS response matrix based on pure effect matrices only. The effect contributions for the three predictive components and the H×T interaction are similar to the ones obtained in AComDim but, in AMOPLS, small contributions of other model effects appear and are linked to the fact that AMOPLS is very sensitive to unbalanced designs.
These contribution plots can finally be linked to Table 2(c) and (d) to show that variance explained by components linked to the main model effects are, in AComDim and AMOPLS, generally smaller than the real proportion of the variance explained by the main effects in the data (see Table 1). This is particularly true for AMOPLS. This decrease is clearly due to the overweighting of the error term in the decomposed data matrices and can be deduced mathematically.

5.4. Significance tests of model effects

Significance tests of the model effects (Table 3) were performed for ASCA, AComDim and AMOPLS. The test statistic is based on the type III Sum of Squares of the effects for ASCA, on a salience ratio for AComDim and on a contribution ratio (RSR) for AMOPLS (see Section 3). Permutations (n=1000) were implemented to assess the statistical significance of the test statistics according to the approach suggested in [9] : for main effects , responses were permuted between factor levels keeping other factor levels constant and, for interaction effects, Manly's approach was used. The p-values were estimated as the proportion of test statistics obtained from the random permutation response matrices which are larger than the test statistic obtained for the original response matrix. For AComDim, the p-values based on the $((N-1),(N-1))$ F-distribution recommended in [26] were also computed.

Table 3 shows that results from all approaches seem consistent: the Hippurate, Citrate, Time main effects and the H×T interaction are significant at the $\alpha = 0.05$ level. The other interactions are not significant across all methods. The interpretation of the significant effect will be done on the basis of scores and loadings in the next sections. Note also that for AComDim, the same test statistic is used for both the F and the permutation approaches but the p-values are calculated in different ways. In this example, the resulting p-values are consistent but one must note that, in spite of this, the densities of the test statistic under the null hypothesis in both cases (the $F((N-1),(N-1))$ density and the density obtained from permutations) are quite different. As an illustration, the density of the empirical distribution of the F-statistic under H_0 for the Hippurate effect is right-skewed, having minimum values of 1.01 and quantiles $q_{0.05} = 1.03$ and $q_{0.95} = 1.26$ whereas the F distribution with $(33, 33)$ df is more symmetric, centered around 1 with quantiles $q_{0.05} = 0.56$ and $q_{0.95} = 1.79$.
The results of these significance tests also confirm the adequacy of the number of components selected in Section 5.2 on the basis of the scree plots.

Table 3: Model effects significance tests for ASCA, AComDim & AMOPLS

	ASCA		AComDim			AMOPLS	
	Test statistic	p-val. permut.	Test statistic	p-val. F-test	p-val. permut.	Test statistic	p-val. permut.
Hippurate	9.15	< 0.001	10.59	< 0.001	< 0.001	17.97	< 0.001
Citrate	6.96	< 0.001	8.26	< 0.001	< 0.001	13.39	< 0.001
Time	3.78	< 0.001	4.87	< 0.001	< 0.001	7.68	< 0.001
H:C	0.36	0.146	1.24	0.274	0.366	1.20	0.499
H:T	1.45	<0.001	2.55	0.004	< 0.001	2.99	< 0.001
C:T	0.13	0.448	1.10	0.397	0.499	1.05	0.680
H:C:T	0.39	0.104	1.30	0.227	0.179	1.19	0.548

5.5. Scores

In the context of this paper, scores plots help visualise the amplitude of model effects directly on the data, identify outlying or specific observations, and detect potential patterns. The scores plots are presented as

scatter plot matrices for ASCA-E, AComDim and AMOPLS in Figures 11, 14 and 15. For each component, the model effect with the highest contribution is indicated and the observations are identified with respect to different factors in the upper and lower part of the scatter plot matrix. In the upper part, colours and markers differentiate Hippurate et Citrate levels and in the lower part, the observations are identified with respect to the Time. For PARAFASCA, score plot are different and their construction is explained below. For ASCA-E (Figure 11), 9 components have been kept to visualise the data and factor effects but the 3 components related to the non-significant interaction terms are less relevant. The PC1 scores for the three main effects Hippurate, Citrate and Time show a clear separation between the factors levels which is a confirmation of their significance and are consistent with the initial shape of the $3 \times 3 \times 2$ factorial design. On line 5 and column 1, the (H×T, Hippurate) scores plot shows a distinctive pattern which helps to interpret the interaction between Hippurate and Time. The observations for the two Time levels cross each other when Hippurate increases, which shows that the effect of Time is different for level 0 of Hippurate compared to levels 1 and 2. This will be much clear in the PARAFASCA score plot below. Finally, PC2 of residuals allows identifying two moderate outliers which will be confirmed by other methods' scores plots.

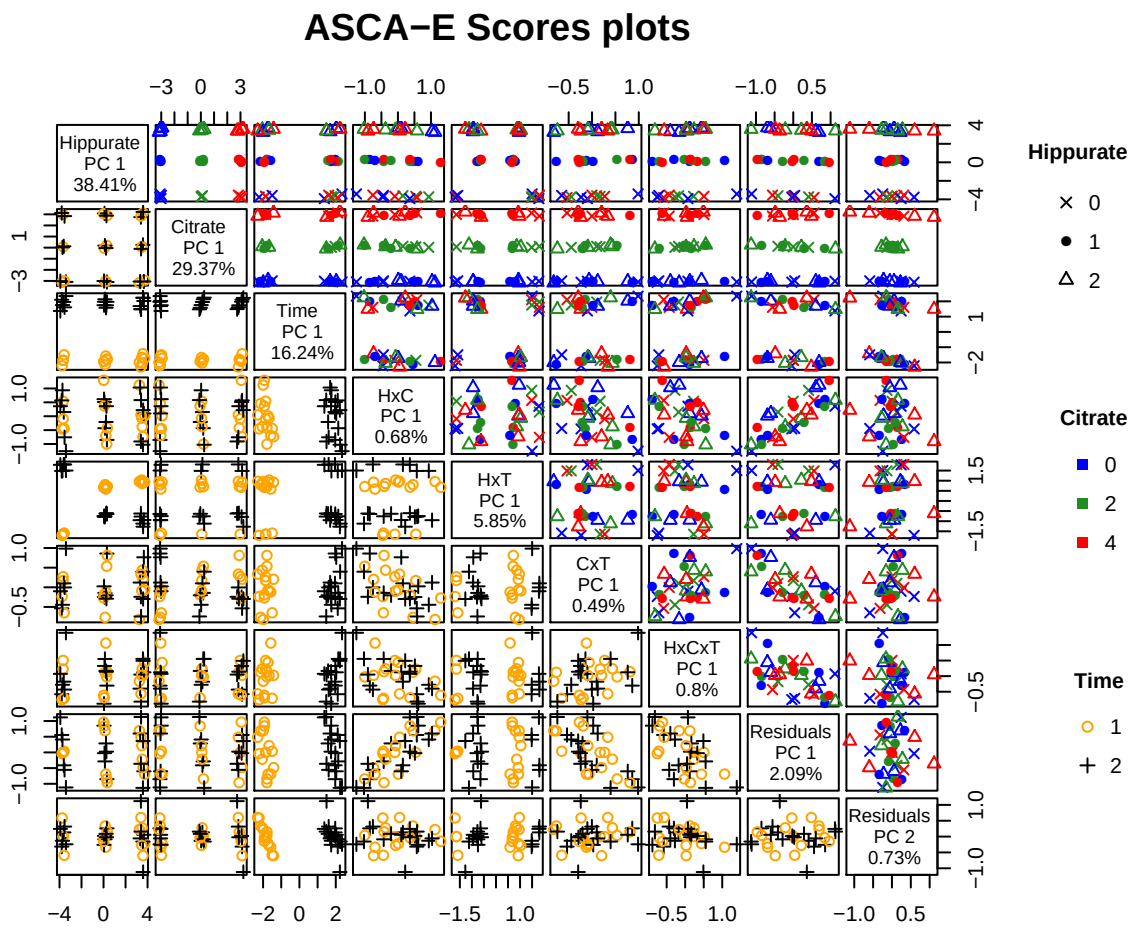

Figure 11: ASCA-E Scores plots.

Score plots for the full PARAFASCA are built in this section in a different way than in Figure 9. For each component r, the tensor product $u_r \Delta v_r \Delta w_r$ between the scores of the three first modes have been calculated and all numbers in the resulting three-way array represented on one graphic with respect to the

most important factor highlighted in Figure 9. These graphs (see Figure 12) show much more clearly that components 1, 2 and 3 are respectively linked to H, C and T alone and that component 4 is linked to the $H \times T$ interaction. Putting these graphics next to the loadings plots provided in the next section will allow understanding very clearly how the NMR spectra are affected by factor changes. To obtain a clearer view of the interaction effect, a three-way PARAFASCA was then applied to the $H \times T$ effect alone, which provides the interaction plot of Figure 13. This graph is very similar to the graph (5,1) obtained in the ASCA-E score plot matrix.

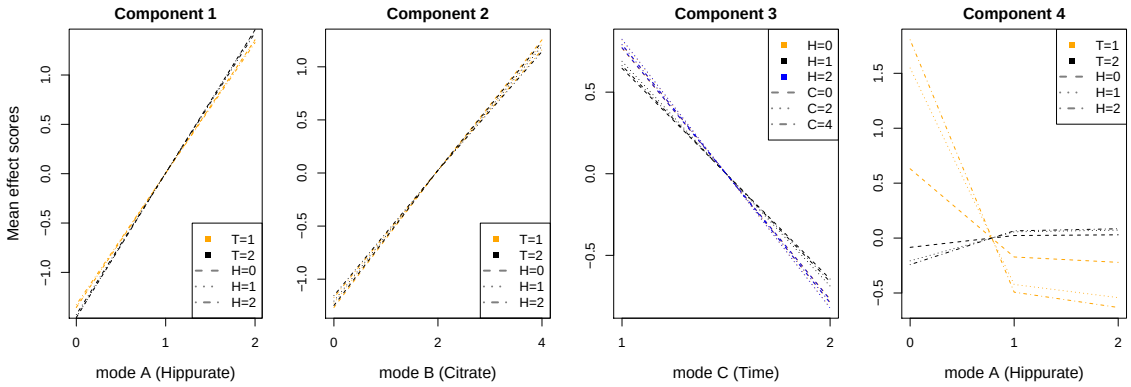

Figure 12: (Mean effect) score plots by component for the full PARAFASCA model

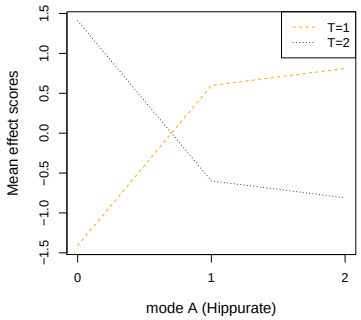

Figure 13: (Mean effect) score plots by component for the partial PARAFASCA model on H×T

AComDim scores plots are very similar to those of ASCA-E but present a more parsimonious image of this information. In Figure 14, the main effects of Hippurate, Citrate and Time are recovered very clearly by the common components 2, 3 and 4. The (H×T, Hippurate) scores plot shows the same pattern as in ASCA-E. Common Dimension 5, linked to residuals, allows visualising the same two outliers as in ASCA-E.

For AMOPLS, predictive Components 1 and 2 scores plots clearly correspond to the 9 combinations of concentrations available in the experimental design, but the results are slightly less clearly displayed than in AComDim and ASCA-E. This is due to the fact that AMOPLS has been designed for balanced design and is thus quite sensitive to unbalanced designs. Predictive components 3 and 4 respectively correspond to the Time effect and the H×T interaction. The same interaction pattern appears on the (H×T, Hippurate) scores plot. The two Orthogonal Components are mostly noise and OC2 lets appear the two outliers.

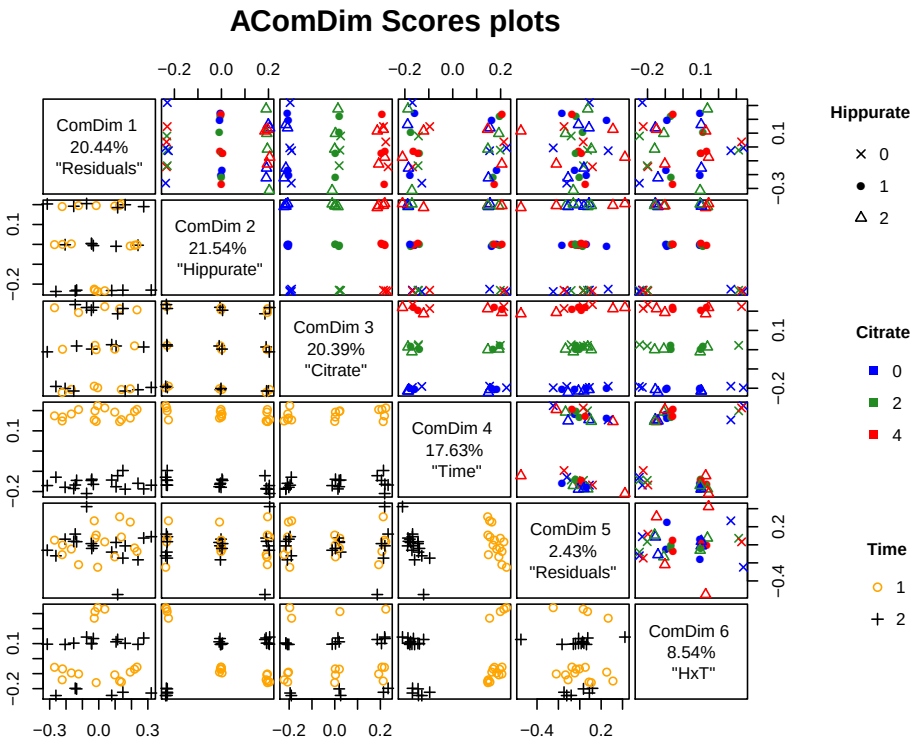

Figure 14: AComDim scores scatter plot matrix.

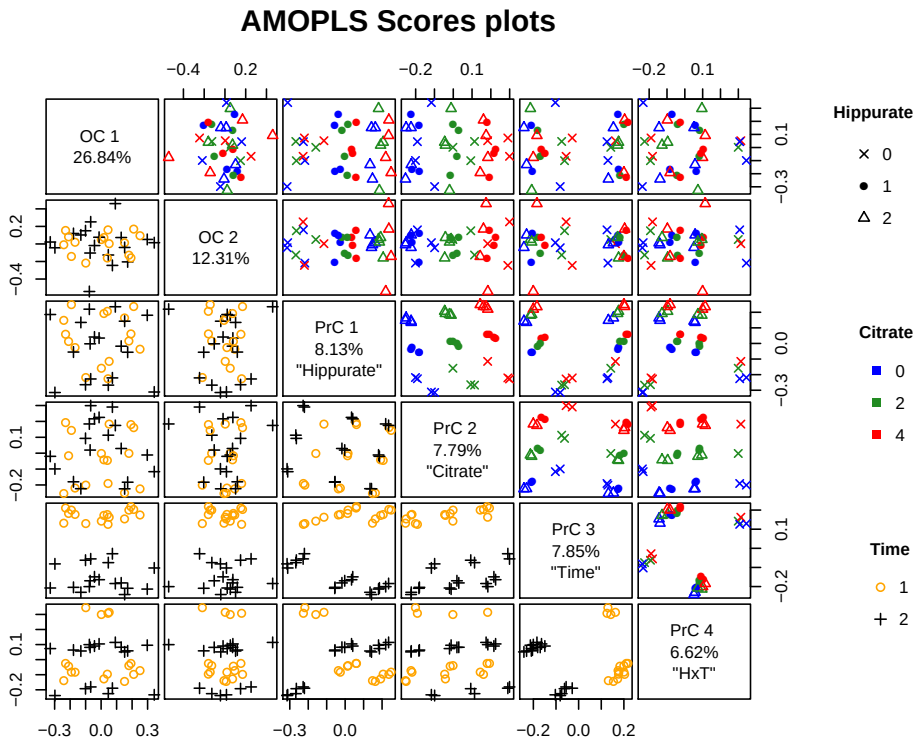

Figure 15: AMOPLS scores scatter plot matrix.

5.6. Loadings

Loadings plots provide a way to visualise the link between the model effects and the m responses (the 600 descriptors from the spectral matrix). Figure 16 presents, for each method of interest, the loadings for the components corresponding to the model main effects Hippurate, Citrate, Time and H×T interaction and are very similar accros methods. The loadings for the Hippurate and Citrate main effects clearly show the expected molecule peaks: 4 peaks for Hippurate and one for Citrate (this peak has been aggregated in the preprocessing step). The Time effect is located at the border of the water zone and the interaction effect shows alternating peaks around 3 ppm typically due to peak shifting across spectra.

The Figure 12 score plot combined with these loading plots provide a very clear way to understand how factor changes affect the spectra in the three-factor design.

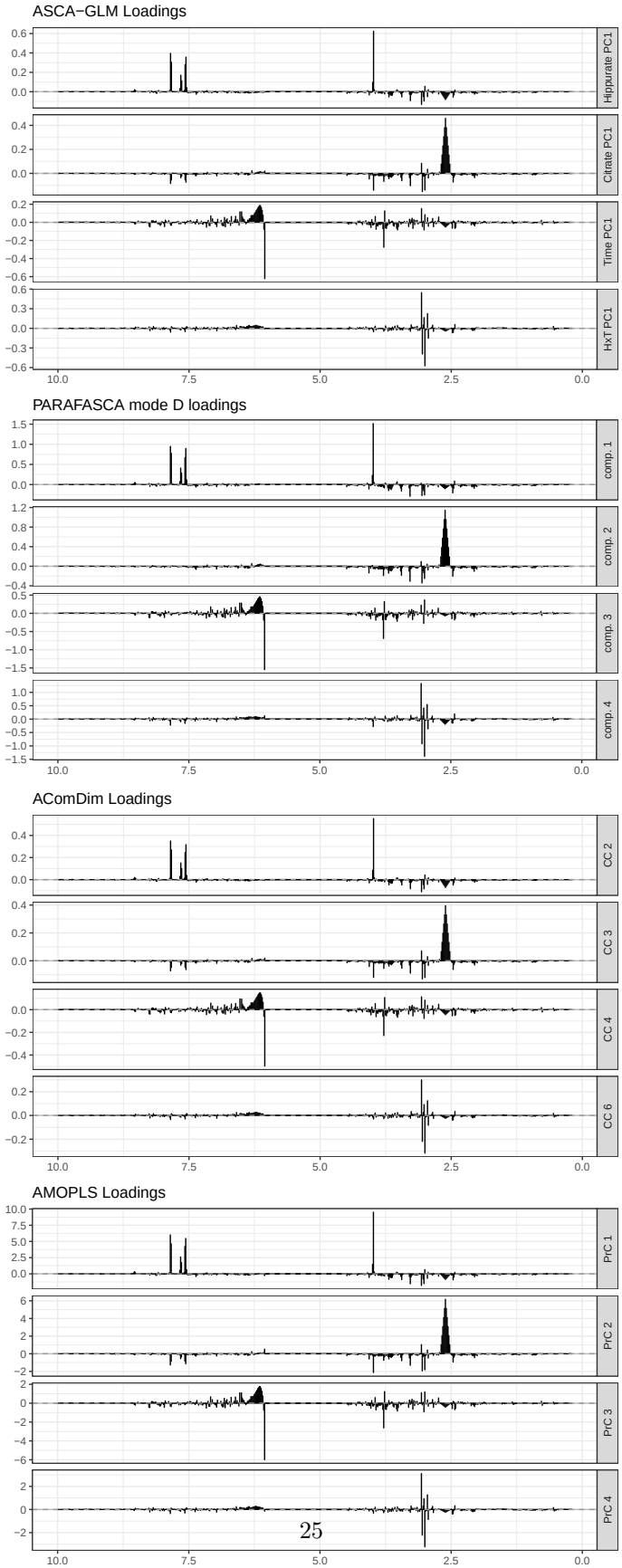

Figure 16: Loadings plots for all methods calculated for the components corresponding to the three main effects Hippurate, Dilution and Citrate.

6. Discussion and conclusions

This paper aimed at presenting together three "global" methods suitable for the analysis of Wide Multi-Response Experimental Design (WMRED) data, compare them to the traditional ASCA/APCA approaches and extend them to unbalanced designs thanks to an effect matrices estimation by GLM instead of ANOVA modeling. These techniques first decompose the response matrix with respect to the ANOVA/GLM model before analysing the resulting effect matrices with one multivariate technique to interpret the model results. It has been shown that the main advantages of these methods are their ability to evaluate the effect of all model terms on the responses under a common framework with an automatic ranking of their importance, and that they provide several interesting tools to ease the interpretation of the components generated by the multivariate step.

The case study has illustrated that these methods provide interesting tools to complement ASCA and APCA, especially when the multifactorial model has a large number of terms. On the other hand, these global approaches can induce a remixing of the model effects because they are recombined in a common object, although the experience of the authors of this paper and the related literature show that they are quite robust to this problem especially in the balanced case. This will be discussed further below.

The objective of this paper was not to select one method over the others but to present under a common framework, analyse objectively, and illustrate on a case study similar approaches in parallel. All techniques lead to useful and interpretable results and provide a decomposition of the data into components linked to specific model effects, but they differ in subtle ways. Table 4 presents some of the main differences between techniques.

Table 4: Comparison of the three techniques

Technique	**PARAFASCA**	**AComDim**	**AMOPLS**
Data type	D-dimensional array	Multi-block	Multi-block
Matrices used	Effects estimation related to the question of interest	Residual-augmented effect matrices	Pure effect matrices (response) and residual-augmented effect matrices (predictors)
Residual matrix	Not included	Included	Included for the predictor matrix, excluded for the reponse matrix
Underlying methodology	Parallel Factor Analysis	Common Dimension analysis	Kernel Orthogonal Partial Least Squares
Unbalanced Design	Applicable	Applicable	Applicable for no or minimally unbalanced data

PARAFASCA can be seen as an elegant and parsimonious way to complement the results provided by ASCA to visualise all or some factor effects, especially in the presence of interactions. PARAFASCA, being a multi-way method, provides results that are closer to the initial structure of the interaction terms in the model and works directly on the effect matrices estimated by the ANOVA/GLM model without including model residuals in the analysis. Thanks to these two properties, it can provide a cleaner and more readable view of factor effects than other methods.

For complex factor behaviors or if too many model terms are included in the PARAFASCA array, the method could induce effect remixing, but the case study of this paper illustrate at least that, even on a full three-factor cross design, PARAFASCA may provide very elegant results. Note also that the ALS algorithm can be unstable when the chosen number of components is to high.

In AComDim, the interpretation is a strength: it is facilitated by the use of saliences, which allow easily identifying the links between the model effects and common components. AComDim could in theory be affected by effect remixing but the case study of this paper, other experiences by its authors and the results

on balanced data found in [26] always show very pure profiles for common components not interpreted as noise.

Some additional results provided in Appendix 7.3 confirm this and also show that, even under high unbalance, the common components remain mainly related to only one model term. The use of residual-augmented effect matrices in AComDim also deserves some comments: residuals are added to each effect matrix and therefore have a big weight in the ComDim decomposition. The percentage of variance explained by the residuals is thus overestimated and the cleanliness of the scores and loadings may be slightly impacted but, on the other hand, this extra noise has the important advantage to hide non significant model terms into the noise and automatically extract the important ones in the analysis. Lastly, in AComDim, the F-test suggested in the original article is not suitable in this context, although the alternative permutation test suggested in this paper can alleviate this problem. AMOPLS is the only supervised method, using the pure effect matrices to guide the components construction. In this respect, it is close to AoV-PLS and ANOVA-TP. This technique can therefore potentially lead to better components and avoid effect mixing in the balanced case due to the orthogonal structure of the O-PLS response matrix. On this last aspect, Appendix 7.3 show that, unfortunately, in the highly unbalanced case, the AMOPLS components may be very affected, unlike what was observed in AComDim. Data imputation could then be considered if one want to apply AMOPLS on unbalanced designs. Note also that the AMOPLS algorithm may seem fairly complex for users who are not familiar with PLS/OPLS, but this paper shows that some steps of the procedure can be simplified. On the other aspects, AMOPLS tools, results and properties are very close to those provided by AComDim.

The choice of the number of components is also an important matter to consider in AComDim, PARAFASCA and AMOPLS. In AComDim, things are close to the spirit of PCA : they are extracted one by one and one can choose a sufficiently large number of components to be sure to extract all the information. Complementary tools like a scree plot may then be used to decide how many are of interest. In PARAFASCA the components profiles change for different number of components extracted by the ALS algorithm. It is then recommended to compare the results obtained with several model dimensions and chose an adequate number on the basis of a scree plot, the profils of the obtained score and loading plots and by remaining parsimonious in order to avoid the instability of the results. In AMOPLS, the choice of the number of orthogonal components may be simply based on the dimensionality of the residuals identified via ASCA and empirical results also show that the choice of one or two orthogonal components in AMOPLS often does not lead to major changes in the results. Next, the choice of the number of (useful) predictive components can be guided, as in AComDim, by the use of a scree plot.

The three tested methods rely on the choice of joint components for all the model effect matrices, which is not the case in ASCA when separate PCA are applied to each effect matrix instead of an integrated SCA analysis. This last method was not the aim of this paper and would probably not have led to different results in the case study, but its robustness to data unbalance and effect mixing would gain to also be tested in further works.

In conclusion, the three methods discussed have shown their usefulness to analyse Wide Multi-Response Experimental Design (WMRED) data, while each has its own advantages and disadvantages. They allow for a fast and parsimonious analysis when the number of model terms is large, while remaining within the ASCA/APCA framework in terms of available information. Besides, the GLM approach helps analyse the data on the same basis as the ANOVA approach, but may be extended to unbalanced designs without biasing the estimation. Depending on its data and context, researchers may choose to apply an ASCA/APCA approach or one of the techniques described here in order to reach their objectives.

7. Appendices

7.1. PARAFASCA algorithm

One of the most common algorithms to compute a PARAFAC decomposition is the alternating least squares algorithm. Its objective is to derive iteratively each loading/score matrix by finding a least square solution to equations relating the matrix of interest and the unfolded array built from the loading/score

matrices related to the other models. An important PARAFAC step is the unfolding of the array Y_G to obtain two-dimensional matrices. Figure 17 shows the possible unfolded arrays obtained from Y_G and used in the computation of the PARAFAC components when the number of modes is 3.

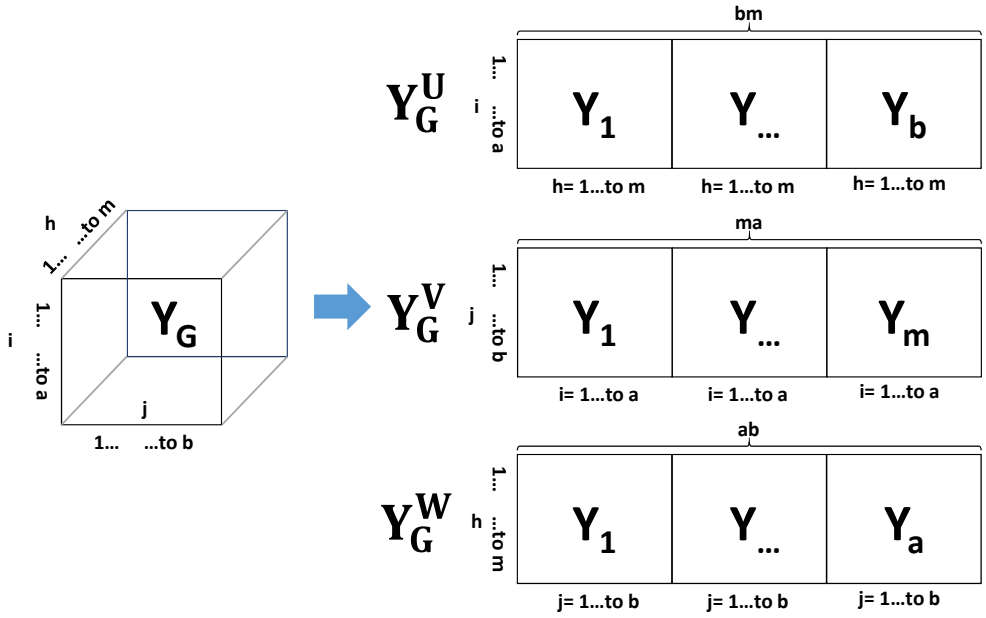

Figure 17: Visualisation of the unfolded PARAFASCA arrays adapted from [17]

The full algorithm includes the following steps [17, 19]:

■ Initialisations

- Compute Y_G, the three-dimensional array resulting from concatenating the residual-augmented effect matrices (Figure 4).
- Decide on a number of components R to compute.
- Initialise loading/scores matrices V_0 and W_0 of size $b \times R$ and $m \times R$ respectively. Various options can be used, among which random initialisation or an SVD-based initialisation.

■ START OF THE LOOP with k, the number of iterations until convergence and r, the component, with r=1 to R

- Unfold Y_G to an $a \times (m.b)$ matrix Y_G^U and define Z_{Uk} as the $R \times (m.b)$ matrix with row vectors $v'_r \otimes w'_r$ where v_r and w_r are respectively the r^{th} columns of V_{k-1} and W_{k-1}.
- Estimate matrix U_k (from V_{k-1} and W_{k-1}) by least square regression: $U_k = Y_G^U Z'_{Uk}(Z_{Uk}Z'_{Uk})^{-1}$.
- Unfold Y_G to an $b \times (m.a)$ matrix Y_G^V and define Z_{Vk} as the $R \times (m.a)$ matrix with row vectors $u'_r \otimes w'_r$ where u_r and w_r are respectively the r^{th} columns of U_k and W_{k-1}.
- Estimate matrix V_k (from U_k and W_{k-1}) by least square regression: $V_k = Y_G^V Z'_{Vk}(Z_{Vk}Z'_{Vk})^{-1}$.

- Unfold Y_G to an $N \times (b.a)$ matrix Y_G^W and define Z_{Wk} as the $R \times (b.a)$ matrix with row vectors $u'_r \otimes v'_r$ where u_r and v_r are respectively the r^{th} columns of U_k and V_k.
- Estimate matrix W_k (from U_k and V_k) by least square regression based on $W_k = Y_G^W Z'_{Wk}(Z_{Wk}Z'_{Wk})^{-1}$
- Compute the loss function $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{h=1}^{m}(y_{ijh} - \sum_{r=1}^{R} u_{ir}v_{jr}w_{hr})^2$.
- If the loss function does not decrease by more than a specified threshold value, the algorithm is considered as having converged and the loop ends. Otherwise, the algorithm goes back to the start of the loop $(k+1)$ using the new matrices U_k, V_k and W_k.

■ END OF THE LOOP when convergence has taken place

■ The loadings/scores matrices are the matrices U, V and W computed in the last iteration of the loop.

Note that this algorithm can be slow to converge but some enhancements are proposed in the related literature (see [17]).

7.2. AComDim algorithm

The iterative algorithm used in ComDim can be detailed as follows [28, 26]:

■ Initialisations

- Calculate the residual-augmented effect matrices $\tilde{M}_f = \hat{M}_f + E$ with $f = 1$ to $F-1$ and define $\tilde{M}_F = E$
- Center each matrix by column and normalise it by its Frobenius norm.
- Compute initial sample variance-covariance matrices $\Psi_f^1 = \tilde{M}_f\tilde{M}'_f$ for f=1 to F.

■ START OF LOOP 1 : repeat for each common dimension r to be extracted with r=1 to R

- Initialise the saliences λ_f^r to 1 for f=1 to F.
- START OF LOOP 2: repeat until convergence
  - Compute the weighted sum of the Ψ_f^r matrices $\Psi_G^r = \sum_{f=1}^{F}\lambda_f^r\Psi_f^r$.
  - Decompose Ψ_G^r by singular value decomposition into $Q_r\Gamma_r Q'_r$.
  - Assign the value of the first eigenvector of Ψ_G^r (the first column vector of Q_r) to the vector q_r.
  - Update the saliences $\lambda_f^r = q'_r\Psi_f^r q_r$ with f=1 to F.
  - Compute the loss function $\sum_{f=1}^{F}||\Psi_f^r - \lambda_f^r q_r q'_r||^2$.
  - Evaluate the convergence of the loss function. If the loss function does not decrease by more than a specified threshold value (for instance 10^{-20}), the algorithm is considered as having converged.
  - If the algorithm has not converged, go back to the beginning of LOOP 2 using the new values of the saliences λ_f^r.
- END OF LOOP 2: when convergence has taken place
- The final values of q_r and λ_f^r are respectively defined as the r-th common dimension and the r-th salience of matrix f with f=1 to F.
- Compute the deflated matrices: $\Psi_f^{r+1} = (I_n - q_r q'_r)\Psi_f^r(I_n - q_r q'_r)'$ for f=1 to F and go back to the beginning of LOOP 1 to compute the next common component on the deflated matrices Ψ_f^{r+1}, $f = 1...F$.

■ END OF LOOP 1 when $r = R$.

The process is iterated until a pre-selected number of common dimensions has been computed. The percentage of variance explained by each common component can guide the choice of the number of dimensions to analyse.

7.3. Factor mixing in ACOMDIM and AMOPLS for different degrees of data balanceness

Table 5: Observation frequencies in the balanced and highly unbalanced datasets.

		Balanced design			Highly unbalanced design			
		Citrate				Citrate		
	Hippurate	0	2	4	Hippurate	0	2	4
Time = 1	0	2	2	2	0	1	1	1
	1	2	2	2	1	2	2	2
	2	2	2	2	2	1	1	2
Time = 2	0	2	2	2	0	2	2	1
	1	2	2	2	1	1	2	2
	2	2	2	2	2	2	2	2

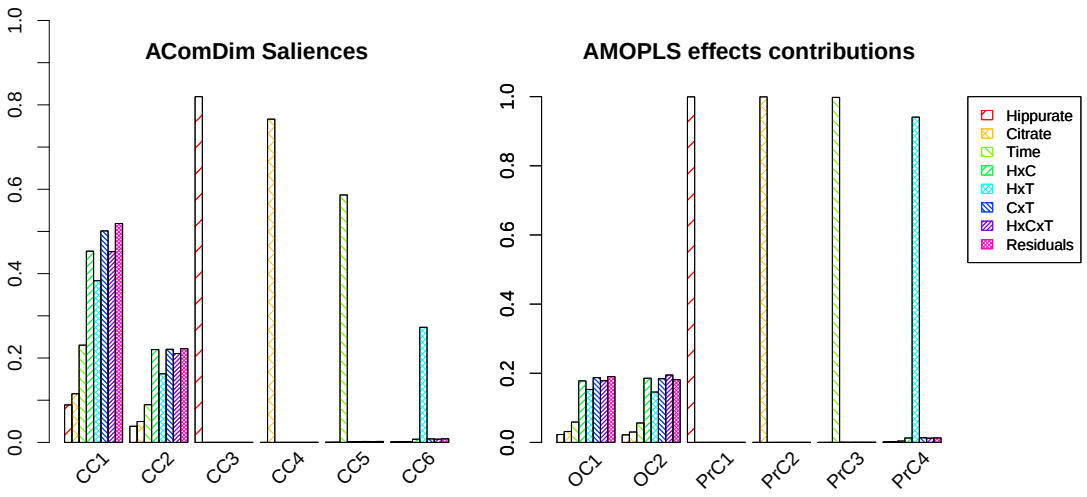

Figure 18: Model effects and residuals contributions for each AComDim and AMOPLS predictive components for a balanced dataset.

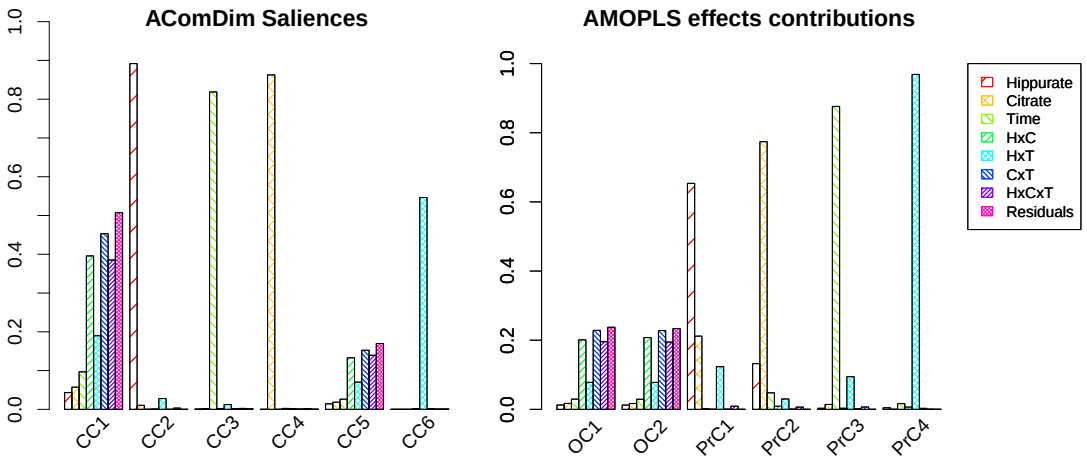

Figure 19: Model effects and residuals contributions for each AComDim and AMOPLS predictive components for a highly unbalanced dataset.

Acknowledgements

The authors thank the Université catholique de Louvain and in particular the Statistical Methodology and Computing Support (SMCS) for their support, Julien Boccard and Michel Thiel for fruitful discussions and sharing codes, and Eli Lilly company for providing the data used in this paper. The second author gratefully acknowledges funding from the Belgian Fund for Scientific Research (F.R.S.-FNRS) with a FRIA grant. Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is also gratefully acknowledged. Reviewers of this article are finally thanked for all their very relevant and stimulating comments.

Softwares

The R software environment was exclusively used (http://www.R-project.org). Data and R code are available on request.

Conflict of Interest

Authors declare that they have no conflict of interest.

Compliance with ethical requirements

This study implies the use of a pool of urine of healthy rats collected in agreement with standard ethical rules.

References

[1] A. K. Smilde, J. J. Jansen, H. C. J. Hoefsloot, R.-J. A. N. Lamers, J. van der Greef, M. E. Timmerman, ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data, Bioinformatics 21 (13) (2005) 3043–3048. doi:10.1093/bioinformatics/bti476.
URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bti476

[2] P. d. B. Harrington, N. E. Vieira, J. Espinoza, J. K. Nien, R. Romero, A. L. Yergey, Analysis of variance-principal component analysis: A soft tool for proteomic discovery, Analytica Chimica Acta 544 (1-2) (2005) 118–127. doi:10.1016/j.aca.2005.02.042.
URL http://linkinghub.elsevier.com/retrieve/pii/S0003267005002692

[3] A. El Ghaziri, E. M. Qannari, T. Moyon, M.-C. Alexandre-Gouabau, AoV-PLS: a new method for the analysis of multivariate data depending on several factors, Electronic Journal of Applied Statistical Analysis 8 (2) (2015) 214–235.
URL http://siba-ese.unile.it/index.php/ejasa/article/view/14988

[4] F. Marini, D. de Beer, E. Joubert, B. Walczak, Analysis of variance of designed chromatographic data sets: The analysis of variance-target projection approach, Journal of Chromatography A 1405 (2015) 94–102. doi:10.1016/j.chroma.2015.05.060.
URL http://linkinghub.elsevier.com/retrieve/pii/S0021967315007839

[5] M. Thiel, B. Fraud, B. Govaerts, ASCA+ and APCA+: Extensions of ASCA and APCA in the analysis of unbalanced multifactorial designs, Journal of Chemometrics 31 (6). doi:10.1002/cem.2895.
URL http://onlinelibrary.wiley.com.proxy.bib.ucl.ac.be:8888/doi/10.1002/cem.2895/abstract

[6] J. Neter, M. H. Kutner, C. J. Nachtsheim, Applied linear statistical models, 4th Edition, The Irwin series in production operations management, Irwin, Chicago, 1996.

[7] J. ten Berge, H. Kiers, V. van der Stel, Simultaneous components analysis, Statistica Applicata 4 (4) (1992) 377–392.

[8] J. J. Jansen, H. C. J. Hoefsloot, J. van der Greef, M. E. Timmerman, J. A. Westerhuis, A. K. Smilde, ASCA: analysis of multivariate data obtained from an experimental design, Journal of Chemometrics 19 (9) (2005) 469–481. doi:10.1002/cem.952.
URL http://dx.doi.org/10.1002/cem.952

[9] M. Anderson, C. T. Braak, Permutation tests for multi-factorial analysis of variance, Journal of statistical computation and simulation 73 (2) (2003) 85–113.
URL http://www.tandfonline.com/doi/abs/10.1080/00949650215733

[10] D. J. Vis, J. A. Westerhuis, A. K. Smilde, J. van der Greef, Statistical validation of megavariate effects in asca, BMC Bioinformatics 8 (1) (2007) 322. doi:10.1186/1471-2105-8-322.
URL https://doi.org/10.1186/1471-2105-8-322

[11] G. Zwanenburg, H. C. Hoefsloot, J. A. Westerhuis, J. J. Jansen, A. K. Smilde, ANOVA-principal component analysis and ANOVA-simultaneous component analysis: a comparison: ANOVA-PCA and ASCA: a comparison, Journal of Chemometrics 25 (10) (2011) 561–567. doi:10.1002/cem.1400.
URL http://doi.wiley.com/10.1002/cem.1400

[12] J. J. Jansen, R. Bro, H. C. J. Hoefsloot, F. W. J. van den Berg, J. A. Westerhuis, A. K. Smilde, PARAFASCA: ASCA combined with PARAFAC for the analysis of metabolic fingerprinting data, Journal of Chemometrics 22 (2) (2008) 114–121. doi:10.1002/cem.1105.
URL http://dx.doi.org/10.1002/cem.1105

[13] F. L. Hitchcock, The Expression of a Tensor or a Polyadic as a Sum of Products, Journal of Mathematics and Physics 6 (1-4) (1927) 164–189. doi:10.1002/sapm192761164.
URL http://onlinelibrary.wiley.com/doi/10.1002/sapm192761164/abstract

[14] 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. doi:10.1007/BF02310791.
URL https://link.springer.com/article/10.1007/BF02310791

[15] R. Harshman, Foundations of the PARAFAC procedure: Models and conditions for an" explanatory" multi-modal factor analysis, UCLA Working Papers in Phonetics 16 (1) (1970) 84.

[16] G. Favier, A. L. d. Almeida, Overview of constrained PARAFAC models, EURASIP Journal on Advances in Signal Processing 2014 (1) (2014) 142. doi:10.1186/1687-6180-2014-142.
URL https://link.springer.com/article/10.1186/1687-6180-2014-142

[17] R. Bro, PARAFAC. Tutorial and applications, Chemometrics and Intelligent Laboratory Systems 38 (2) (1997) 149–171. doi:10.1016/S0169-7439(97)00032-4.
URL http://www.sciencedirect.com/science/article/pii/S0169743997000324

[18] N. K. M. Faber, R. Bro, P. K. Hopke, Recent developments in CANDECOMP/PARAFAC algorithms: a critical review, Chemometrics and Intelligent Laboratory Systems 65 (1) (2003) 119–137. doi:10.1016/S0169-7439(02)00089-8.
URL http://www.sciencedirect.com/science/article/pii/S0169743902000898

[19] G. Tomasi, R. Bro, A comparison of algorithms for fitting the PARAFAC model, Computational Statistics & Data Analysis 50 (7) (2006) 1700–1734. doi:10.1016/j.csda.2004.11.013.
URL http://www.sciencedirect.com/science/article/pii/S0167947304003895

[20] L. Lenhardt, M. D. Dramianin, PARAFAC: A tool for the analysis of phosphor mixture luminescence, Journal of Luminescence 170 (2016) 136–140. doi:10.1016/j.jlumin.2015.10.030.
URL http://linkinghub.elsevier.com/retrieve/pii/S0022231315303434

[21] M. Horochowska, I. Stanimirova, B. Czarnik-Matusewicz, Studying the influence of enflurane, isoflurane, and sevoflurane on the DPPC lipid bilayer using the analysis of variance and parallel factor analysis, Chemometrics and Intelligent Laboratory Systems 153 (2016) 146–152. doi:10.1016/j.chemolab.2016.03.003.
URL http://www.sciencedirect.com/science/article/pii/S0169743916300399

[22] R. Romano, T. Naes, P. B. Brockhoff, Combining analysis of variance and three-way factor analysis methods for studying additive and multiplicative effects in sensory panel data, Journal of Chemometrics 29 (1) (2015) 29–37, cEM-14-0017.R1. doi:10.1002/cem.2659.

URL http://dx.doi.org/10.1002/cem.2659

[23] A. K. Smilde, R. Bro, P. Geladi, Multi-way analysis with applications in the chemical sciences, J. Wiley, Chichester, West Sussex, England ; Hoboken, NJ, 2004.

[24] R. A. Harshman, M. E. Lundy, PARAFAC: Parallel factor analysis, Computational Statistics & Data Analysis 18 (1) (1994) 39–72. doi:10.1016/0167-9473(94)90132-5.
URL http://linkinghub.elsevier.com/retrieve/pii/0167947394901325

[25] R. Bro, H. A. L. Kiers, A new efficient method for determining the number of components in PARAFAC models, Journal of Chemometrics 17 (5) (2003) 274–286. doi:10.1002/cem.801.
URL http://doi.wiley.com/10.1002/cem.801

[26] D. Jouan-Rimbaud Bouveresse, R. C. Pinto, L. Schmidtke, N. Locquet, D. Rutledge, Identification of significant factors by an extension of ANOVA-PCA based on multi-block analysis, Chemometrics and Intelligent Laboratory Systems 106 (2) (2011) 173–182. doi:10.1016/j.chemolab.2010.05.005.
URL http://linkinghub.elsevier.com/retrieve/pii/S016974391000081X

[27] E. M. Qannari, I. Wakeling, H. J. H. MacFie, A hierarchy of models for analysing sensory data, Food Quality and Preference 6 (4) (1995) 309–314. doi:10.1016/0950-3293(95)00033-X.
URL http://www.sciencedirect.com/science/article/pii/095032939500033X

[28] E. M. Qannari, I. Wakeling, P. Courcoux, H. J. H. MacFie, Defining the underlying sensory dimensions, Food Quality and Preference 11 (1-2) (2000) 151–154. doi:10.1016/S0950-3293(99)00069-5.
URL http://www.sciencedirect.com/science/article/pii/S0950329399000695

[29] G. Mazerolles, M. F. Devaux, E. Dufour, E. M. Qannari, P. Courcoux, Chemometric methods for the coupling of spectroscopic techniques and for the extraction of the relevant information contained in the spectral data tables, Chemometrics and Intelligent Laboratory Systems 63 (1) (2002) 57–68.
URL http://www.sciencedirect.com/science/article/pii/S0169743902000369

[30] A. El Ghaziri, V. Cariou, D. N. Rutledge, E. M. Qannari, Extension of ComDim for the analysis of (K+ 1) datasets; Application in Sensometrics (2016).
URL http://agrostat2016.sfds.asso.fr/wp-content/uploads/2016/03/3.04_E.Quannari.pdf

[31] S. Amat, N. Dupuy, J. Kister, D. Rutledge, Development of near infrared sensors: Detection of influential factors by the AComDim method, Analytica Chimica Acta 675 (1) (2010) 16–23. doi:10.1016/j.aca.2010.06.037.
URL http://linkinghub.elsevier.com/retrieve/pii/S0003267010008305

[32] R. Korifi, S. Amat, C. Rebufa, V. Labed, D. Rutledge, N. Dupuy, AComDim as a multivariate tool to analyse experimental design application to gamma-irradiated and leached ion exchange resins, Chemometrics and Intelligent Laboratory Systems 141 (2015) 12–23. doi:10.1016/j.chemolab.2014.12.003.
URL http://linkinghub.elsevier.com/retrieve/pii/S0169743914002561

[33] R. Korifi, J. Plard, Y. Le Dreau, C. Rebufa, D. Rutledge, N. Dupuy, Highlighting metabolic indicators of olive oil during storage by the AComDim method, Food Chemistry 203 (2016) 104–116. doi:10.1016/j.foodchem.2016.01.137.
URL http://linkinghub.elsevier.com/retrieve/pii/S0308814616301364

[34] G. Mazerolles, M. Hanafi, E. Dufour, D. Bertrand, E. Qannari, Common components and specific weights analysis: A chemometric method for dealing with complexity of food products, Chemometrics and Intelligent Laboratory Systems 81 (1) (2006) 41–49. doi:10.1016/j.chemolab.2005.09.004.
URL http://linkinghub.elsevier.com/retrieve/pii/S0169743905001504

[35] J. Boccard, S. Rudaz, Exploring Omics data from designed experiments using analysis of variance multiblock Orthogonal Partial Least Squares, Analytica Chimica Acta 920 (2016) 18–28. doi:10.1016/j.aca.2016.03.042.
URL http://linkinghub.elsevier.com/retrieve/pii/S0003267016303920

[36] M. Rantalainen, M. Bylesjo, O. Cloarec, J. K. Nicholson, E. Holmes, J. Trygg, Kernel-based orthogonal projections to latent structures (K-OPLS), Journal of Chemometrics 21 (7-9) (2007) 376–385. doi:10.1002/cem.1071.
URL http://doi.wiley.com/10.1002/cem.1071

[37] J. Trygg, S. Wold, Orthogonal projections to latent structures (O-PLS), Journal of Chemometrics 16 (3) (2002) 119–128. doi:10.1002/cem.695.
URL http://doi.wiley.com/10.1002/cem.695

[38] J. Boccard, D. N. Rutledge, A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock Omics data fusion, Analytica Chimica Acta 769 (2013) 30–39. doi:10.1016/j.aca.2013.01.022.
URL http://linkinghub.elsevier.com/retrieve/pii/S0003267013001700

[39] R. Rousseau, Statistical contributrion to the analysis of metabonomic data in 1h-NMR spectroscopy, Ph.D. thesis, Universit Catholique de Louvain, Louvain-la-Neuve, Belgium (2011).

[40] M. E. Timmerman, H. C. J. Hoefsloot, A. K. Smilde, E. Ceulemans, Scaling in ANOVA-simultaneous component analysis, Metabolomics 11 (5) (2015) 1265–1276. doi:10.1007/s11306-015-0785-8.
URL http://link.springer.com/10.1007/s11306-015-0785-8

[41] D. S. Wishart, C. Knox, A. C. Guo, R. Eisner, N. Young, B. Gautam, D. D. Hau, N. Psychogios, E. Dong, S. Bouatra, R. Mandal, I. Sinelnikov, J. Xia, L. Jia, J. A. Cruz, E. Lim, C. A. Sobsey, S. Shrivastava, P. Huang, P. Liu, L. Fang, J. Peng, R. Fradette, D. Cheng, D. Tzur, M. Clements, A. Lewis, A. De Souza, A. Zuniga, M. Dawe, Y. Xiong, D. Clive, R. Greiner, A. Nazyrova, R. Shaykhutdinov, L. Li, H. J. Vogel, I. Forsythe, HMDB: a knowledgebase for the human metabolome, Nucleic Acids Research 37 (Database) (2009) D603–D610. doi:10.1093/nar/gkn810.
URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkn810