Check for updates

Received: 20 February 2020 | Revised: 20 May 2020 | Accepted: 22 May 2020

DOI: 10.1002/cem.3269

SPECIAL ISSUE - RESEARCH ARTICLE

Journal of CHEMOMETRICS WILEY

Three-way clustering around latent variables approach with constraints on the configurations to facilitate interpretation

Véronique Cariou[1] | **Marie-Cécile Alexandre-Gouabau**[2] | **Tom F. Wilderjans**[3,4,5]

[1]StatSC, ONIRIS, INRAE, Nantes, 44322, France

[2]NUN, INRAE, CHU Nantes, UMR 1280, Physiopathologie des Adaptations Nutritionnelles (PhAN), Institut des maladies de l'appareil digestif (IMAD), Centre de Recherche en Nutrition Humaine Grand Ouest (CRNH GO), Nantes, 44000, France

[3]Methodology and Statistics Research Unit, Institute of Psychology, Faculty of Social and Behavioral Sciences, Leiden University, Pieter de la Court Building, Wassenaarseweg 52, Leiden, AK, 2333, The Netherlands

[4]Research Group of Quantitative Psychology and Individual Differences, Faculty of Psychology and Educational Sciences, KU Leuven, Tiensestraat 102, Leuven, 3000, Belgium

[5]Leids Universitair Medisch Centrum (LUMC), Leiden Institute for Brain and Cognition (LIBC), Leiden, RC, 2300, The Netherlands

Correspondence
Véronique Cariou, StatSC, ONIRIS, INRAE, Nantes 44322, France.
Email: veronique.cariou@oniris-nantes.fr

Correction added on 5 August 2020, after first online publication: Peer review history statement has been added.

Abstract

The set-up of comprehensive studies in life sciences involving a longitudinal dimension—as appears in time-scale metabolomics—calls for the use of dimension reduction techniques for three-way data structures (e.g., samples by variables by time points). For this purpose, a clustering around latent variables for three-way data approach, *CLV3W*, has been proposed. *CLV3W* aims at both partitioning the variables into nonoverlapping clusters and estimating within each cluster a rank-one Parafac model consisting of a latent component (resp. a weighting system) associated with the first mode (resp. third mode) and a vector of loadings reflecting the degree of closeness of each variable of the second mode to its cluster. In this paper, two constrained *CLV3W* models are discussed. First, a nonnegativity constraint is defined implying that clusters are composed of positively correlated variables. Second, it is proposed to constrain the weighting system to be the same for all clusters. These two constraints aim at providing more parsimonious models with configurations that are easier to interpret. The appropriateness of both constraints is evaluated in a simulation study and illustrated on two case studies pertaining to sensory evaluation and metabolomics data. Regarding the first case study, *CLV3W* yields the identification of two consumer segments together with one common emotional pleasantness dimension associated with coffee aromas. *CLV3W* analysis of human preterm breast milk metabolomics data provided three clusters of lipid species that are responsible for specific functions (i.e., milk fat globules membrane-constituents, fatty acid oxidation-products, lipid mediators as eicosanoids and endocannabinoids).

KEYWORDS

CLV3W, consumer segmentation, metabolomics, nonnegativity, Parafac

Correction added on 5 August 2020, after first online publication: Peer review history statement has been added.

 © 2020 John Wiley & Sons, Ltd.

1 | INTRODUCTION

From the last decade onwards, improvements in high-throughput omics technologies offered an amazing opportunity to capture the whole picture of global biological systems in a hypothesis-free and unbiased way.[1] Moreover, these holistic research methods can be applied at multiple levels of biological information for a comprehensive understanding of biological processes. In metabolomics, which is actually one of the most rapidly evolving "omics" tools measuring the small molecule composition of biofluids and tissues, representing these multiple levels of information leads to several data blocks. These latter ones, containing for the same set of samples information measured on different analytical platforms, are commonly analyzed by multiblock techniques.[2–6] Notwithstanding, a biological system can only be fully and dynamically understood when both considering its spatial and temporal dimension. Indeed, another key issue is to account for the longitudinal nature of the measurements often encountered in experimental designs.[1,7] When collecting data over time, it naturally leads to a three-way data structure with samples, metabolites, and time points as the three data modes. The handling of such data tables, with generally a few time points and a large set—easily more than 1000—of highly correlated metabolites, prevents from relying on multivariate time series analysis techniques. Therefore, to properly analyze such data, either dimension reduction or variables selection approaches are preferred with the ultimate goal of the identification of small groups of biomarkers.[7]

In parallel to ordination techniques, cluster analysis is often used to detect potential subsets of variables and the latent dimension underlying each (variable) cluster. In the context of two-way data, the clustering around latent variables (*CLV*) approach[8] has gained ground in sensometrics and chemometrics (for applications on spectral data, see Vigneau et al. and Cuny et al.[9,10]). For clustering variables within the scope of three-way data, Wilderjans and Cariou[11] proposed a clustering around latent variables approach (*CLV3W*) which extends *CLV* to a three-way data array, with the variables referring to the second mode. The *CLV3W* analysis consists of grouping the variables into Q nonoverlapping clusters and determining for each cluster a latent configuration corresponding to a rank-one Parafac model applied to the data of each cluster (i.e., the set of lateral data slices associated with the variables belonging to the cluster in question). More precisely, *CLV3W* simultaneously partitions the variables and determines for each cluster of variables: (a) a latent component associated with the first mode (e.g., samples, subjects), (b) a vector of loadings reflecting the degree of closeness of each variable to this latent component, and (c) a system of weights corresponding to the third mode. *CLV3W* has been introduced in sensometrics in order to either optimally weight underlying sensory dimensions in the context of conventional sensory profiling data[11] or to identify consumer segments on the basis of their multivariate ratings of a set of samples.[12] With regard to longitudinal metabolomics data, *CLV3W* fully takes the higher-order structure of the data into account as opposed to multiblock and multilevel approaches, which rely on unfolding the data leading to a substantial loss of information.

Nevertheless, in some situations, it does not make sense to group together variables that are negatively correlated. In a consumer evaluation context, where the aim is to provide a consumer segmentation given three-way data (e.g., samples by consumers by attributes), clusters need to group consumers that have similar product evaluation patterns, that is to say that are positively correlated together. To account for such user-knowledge, a nonnegativity constraint is imposed on the loadings, which implies that these loadings can only take positive values. As a consequence, it is ensured that the variables belonging to the same cluster are all positively correlated with their latent component. This is also relevant for clustering genes with the clusters grouping co-expressed genes (i.e., genes which are positively correlated).[13] Within the scope of longitudinal metabolomics studies, such a constraint aims at identifying subsets of metabolites that are positively correlated along the biological trajectory. In other situations, it can be expected that the weights do not differ across clusters. This can be incorporated in *CLV3W* by imposing a common weighting system across clusters, thus assuming that the clusters share the same weighting values for the third mode (i.e., occasions/time points) elements. As such, this constraint aims at simplifying the initial model to a more parsimonious one, which facilitate the interpretation of the obtained results. Moreover, adopting a more parsimonious model is also beneficial in terms of model parameter estimates. Indeed, using simpler models prevents overfitting the data, which can seriously deteriorate estimates of model parameters as noise is erroneously considered as signal.

The rest of the paper is organized as follows. In Section 2, we recall the original *CLV3W* model, criterion, and algorithm. Further in this section, we introduce the two different constraints proposed in this paper, with one constraint being imposed on the loadings and the other one on the weights. In Section 3, the initial *CLV3W* approach together

with the proposed constraints is evaluated in a simulation study. In Section 4, the constraints are illustrated on two case studies pertaining to sensometrics and chemometrics.

2 | *CLV3W* FOR THE CLUSTERING OF VARIABLES WITHIN THE SCOPE OF THREE-WAY DATA

2.1 | Structure of the data

The standardized notation introduced by Kiers[14] is adopted herein. Let us denote by $\mathbf{X}$ a $I \times J \times K$ data array where the jth variable ($j = 1, ..., J$) is associated with the lateral slice $\mathbf{X}_j$ (with size $I \times K$) of $\mathbf{X}$. The clustering of variables within the scope of a three-way array $\mathbf{X}$ consists of segmenting the set of lateral slices $\{\mathbf{X}_j | j = 1, ..., J\}$ into a small number of nonoverlapping clusters. Without loss of generality, we assume that all $\mathbf{X}_j$ (j = 1,...,J) are column-wise centered (i.e., across the elements of the first mode). Other pre-processing steps, such as standardization or block scaling, can be applied depending on the dataset under study.[11]

2.2 | *CLV3W* model and criterion

Given a $I \times J \times K$ three-way array $\mathbf{X}$, *CLV3W* aims at simultaneously clustering the J variables into Q nonoverlapping clusters G_q ($q = 1, ..., Q$) and determining Q cluster-specific latent variables $\mathbf{t}_1$, ..., $\mathbf{t}_q$, ..., $\mathbf{t}_Q$ along with cluster-specific weights $\mathbf{w}_q$ such that the following function is maximized:

$$g = \sum_{j=1}^{J} \sum_{q=1}^{Q} p_{jq} cov^2 \left(\mathbf{X}_j \mathbf{w}_q, \mathbf{t}_q \right), \tag{1}$$

with $\|\mathbf{w}_q\| = \|\mathbf{t}_q\| = 1$ for all q ($q = 1, ..., Q$), and p_{jq} denoting whether variable j is allocated ($p_{jq} = 1$) or not ($p_{jq} = 0$) to cluster G_q. In this initial model, $\mathbf{w}_q$ is the cluster-specific vector of weights that is constant for all variables belonging to G_q. Maximizing the *CLV3W* criterion is equivalent to minimizing the least squares loss function:

$$f = \sum_{j=1}^{J} \sum_{q=1}^{Q} p_{jq} \left\| \mathbf{X}_j - \alpha_{j(q)} \left(\mathbf{t}_q \mathbf{w}_q^\top \right) \right\|_F^2, \tag{2}$$

with all symbols as defined above, $\alpha_{j(q)}$ denoting the loading of variable j for cluster G_q (with $\alpha_{j(q)}$ being zero when variable j does not belong to G_q) and $\|\cdots\|_F^2$ denoting the squared Frobenius norm (i.e., sum of squared elements). It is worth noting that $\alpha_{j(q)}$ equals $cov(\mathbf{X}_j \mathbf{w}_q, \mathbf{t}_q)$ and thus reflects the degree of proximity between the latent variable $\mathbf{X}_j \mathbf{w}_q$ related to variable j and the latent component $\mathbf{t}_q$ of its associated cluster. Note that $\alpha_{j(q)}$ is undefined when variable j does not belong to cluster G_q; in that case, $\alpha_{j(q)}$ is taken equal to 0. The *CLV3W* criterion is equivalent to the *Clusterwise Parafac* criterion when clustering the elements of the second mode with Q clusters and one component in each cluster.[15] Moreover, this latter criterion clearly shows the equivalence of the *CLV3W* model with the *ParaFac with Optimally Clustered Variables* (*PFOCV*) model designed for clustering the elements of the first mode.[16]

To partition the variables, starting from an initial variable clustering (into Q clusters), *CLV3W* runs an alternating least squares (ALS) algorithm, which alternates between two updating steps as follows: (a) To update the cluster membership of a variable j, the criterion $f_{jq} = \left\| \mathbf{X}_j - \alpha_{j(q)} \left(\mathbf{t}_q \mathbf{w}_q^\top \right) \right\|_F^2$ is computed for each cluster G_q and variable j is assigned to the cluster G_q for which f_{jq} is minimal[11,15]; (b) After updating the cluster membership of all the variables, a Parafac model[17–19] with one component is carried out on each cluster G_q, that is to say on the three-way array $\underline{\mathbf{X}}^{(q)}$ associated with cluster G_q ($q = 1, ..., Q$). The three-way array $\underline{\mathbf{X}}^{(q)}$ is the array obtained by only taking the data slices $\mathbf{X}_j$ of $\underline{\mathbf{X}}$ associated to variables j belonging to G_q. Both steps together decrease—at least not increase—the loss function value and lead to updates for each cluster G_q and for all cluster-specific parameters $\mathbf{t}_q$, $\mathbf{w}_q$ and $\alpha_{j(q)}$. This procedure is repeated until convergence. For more detail, see Wilderjans and Cariou.[11]

To start the ALS algorithm, an initial variable partition into Q clusters should be specified. Such a partition can be either a user-defined partition reflecting some domain knowledge or a partition generated by the algorithm in a

random, rational or pseudo-rational way.[20] As the ALS algorithm is sensitive to the initial partition used and does not guarantee, even after convergence, to reach the global optimum solution, it is strongly recommended to implement a multi-start procedure (i.e., a similar advise is given for K-means, see[21,22]).

To determine the optimal number of clusters Q (i.e., model selection), it is advised to run the *CLV3W* algorithm with several increasing values of Q and selecting an optimal Q by means of some model selection procedure, like, for example, CHull.[23,24] In CHull, first, the models are determined that are lying at the boundary of the convex hull of a plot in which Q is plotted against a measure of model (mis)fit, which, for example, can be the sum of squared errors or the variance accounted for (VAF). The models not on the boundary are discarded as they have a worse balance between model (mis)fit and Q than the models at the boundary. In a second step, Q is identified by looking for the solution after which the boundary of the convex hull levels off. To this end, the *st*-scree-value is computed and the Q with the largest *st*-value is selected (for an illustration of CHull, see Section 4.2).

2.3 | Nonnegativity constraint on the cluster-specific loading vectors: *CLV3W-NN*

To ensure that variables belonging to the same cluster are homogeneous in the sense that these variables are as much related—in terms of a positive covariance—as possible with their cluster-specific latent variable, a nonnegativity constraint is imposed on the loadings $\alpha_{j(q)}$.[12] This model, denoted by *CLV3W-NN*, implies that for a variable j belonging to a particular cluster G_q, the weighted average over the third mode elements $\mathbf{X}_j\mathbf{w}_q$ is positively related to the latent variable $\mathbf{t}_q$, which is associated to the cluster in question. As it has been shown from the *CLV3W* optimization criterion that $\alpha_{j(q)}$ is equal to $cov(\mathbf{X}_j\mathbf{w}_q, \mathbf{t}_q)$, this constraint can be obtained by adding the following nonnegativity constraint to the *CLV3W* optimization criterion:

$$g_{NN} = \sum_{j=1}^{J}\sum_{q=1}^{Q} p_{jq} cov^2\left(\mathbf{X}_j\mathbf{w}_q, \mathbf{t}_q\right), \text{s.t.} cov\left(\mathbf{X}_j\mathbf{w}_q, \mathbf{t}_q\right) \geq 0. \tag{3}$$

To estimate the *CLV3W-NN* parameters, a similar ALS algorithm as for *CLV3W* can be used after implementing two changes: (a) in step 1 of the algorithm, the optimal nonnegative $\alpha_{j(q)}$ given $\mathbf{t}_q$ and $\mathbf{w}_q$ for each cluster G_q ($q = 1, ..., Q$) should be computed by means of a nonnegativity constrained linear regression[25,26]; (b) in step 2, the cluster-specific parameters should be updated by fitting a one-component *Parafac* model[17–19] with nonnegativity constraint on the loadings to each three-way array $\underline{\mathbf{X}}^{(q)}$ ($q = 1, ..., Q$).

2.4 | Equality constraint on the cluster-specific weighting system

In some situations it may occur or be expected that the distribution of the weights is basically the same across the different clusters. This similarity in weight distributions may indicate that the variables share the same overall behavior among the third mode units. In the metabolomics case, for example, this may imply that the metabolites (i.e., variables) have a very similar longitudinal trajectory irrespective of the cluster they belong to. As a consequence, imposing an equality (across clusters) constraint on the weights suggests that the clusters of variables differ only in the way the first mode units are scored by each variable (cluster) across all third mode units. Such a property leads to a simpler (more parsimonious) configuration (i.e., less parameters), which may be easier to interpret. This corresponds to the following constrained optimization criterion:

$$g_W = \sum_{j=1}^{J}\sum_{q=1}^{Q} p_{jq} cov^2\left(\mathbf{X}_j\mathbf{w}, \mathbf{t}_q\right), \tag{4}$$

in which $\mathbf{w}$ is kept constant across clusters, with the length of $\mathbf{w}$ set to one. Maximizing the previous criterion is equivalent to minimizing the least squares loss function:

$$f_W = \sum_{j=1}^{J}\sum_{q=1}^{Q} p_{jq} \left\| \mathbf{X}_j - \alpha_{j(q)}\left(\mathbf{t}_q\mathbf{w}^\top\right)\right\|_F^2. \tag{5}$$

To impose the equality constraint, a modification of step (2) of the original ALS algorithm for optimizing the *CLV3W* criterion is required. In particular, in step 2 of the ALS algorithm, after initializing $\mathbf{w}$ (and having an initial variable partition as obtained in step 1 of the ALS algorithm), the two following steps have to be alternated until convergence*:

1. Each $\mathbf{t}_q$ is updated as the left singular vector corresponding to the largest singular value of the $I \times \#G_q$ matrix (with $\#G_q$ denoting the number of elements in cluster G_q) that is obtained by horizontally concatenating the vectors $\mathbf{X}_j\mathbf{w}$ for all variables j that belong to cluster G_q; $\mathbf{X}_j\mathbf{w}$ is the weighted average of the third mode units associated with variable j.
2. The common weight $\mathbf{w}$ is updated as the left singular vector associated with the largest singular value from the $K \times J$ matrix that is obtained by horizontally concatenating the vectors $\mathbf{X}_j^\top \mathbf{t}_q$ for all j ($j = 1, ..., J$).

This parsimonious model involves determining a single vector of weights that is common to all clusters.

2.5 | Software

Functions to perform *CLV3W* and *CLV3W-NN* analyses have been implemented in Matlab (version 2014b) and in the R (version 3.4.4) package *ClustVarLV*.[27] The constraint on the common weighting system will soon be integrated in the *ClustVarLV* package.

3 | SIMULATION STUDY

3.1 | Problem

The goal of this section is to evaluate the performance of the unconstrained and constrained (i.e., nonnegativity and/or equality) *CLV3W* algorithms in a simulation study. In particular, the interest is in the recovery of the true clusters (i.e., true object partition), the true cluster-specific components (i.e., scores for the first mode), and associated true weighting system (i.e., scores for the third mode).

Moreover, we will investigate whether and how these recovery aspects depend on three data characteristics. The first data characteristic is the degree of congruence (i.e., similarity) between the cluster-specific weighting systems $\mathbf{w}_q^{true}$. This characteristic will be varied at three levels, with one level implying an equal weighting system across clusters. Regarding this characteristic, we expect that algorithmic performance will deteriorate when there is a large amount of congruence between the cluster-specific weighting systems as this implies that clusters are more similar (i.e., overlap) and thus are more difficult to disentangle from each other.[28,29] Moreover, enforcing the equality constraint when in reality the $\mathbf{w}_q^{true}$'s are not equal across clusters is also expected to decrease recovery performance. Further, as a second factor, it will be manipulated whether the variable loadings (i.e., scores for the second mode) are all nonnegative or can contain negative values. We expect that when the true loadings are nonnegative, all algorithms—also those not imposing nonnegativity—will perform more or less the same. When, however, the true loadings are not constrained to be nonnegative, we expect that algorithms enforcing nonnegativity will perform worse than algorithms without this constrained enforced. Finally, the third factor pertains to the amount of noise in the data, for which we hypothesize that algorithmic performance will deteriorate when the data become more noisy.[30–32]

3.2 | Design and procedure

As the goal was to keep the simulation study limited and focus on the effect of the nonnegativity and/or equality constraint, the size of the data set was kept constant by fixing the number of samples at 100 ($I = 100$), the number of variables at 250 ($J = 250$), and the number of occasions/time points at 7 ($K = 7$). Further, the number of variable clusters

*Note that when obtaining an initial $\mathbf{w}$, which can be generated in various ways (i.e., random, rational, pseudo-rational), it is important to respect the length one constraint: $\|\boldsymbol{w}\| = 1$. As the final estimate for $\mathbf{w}$ obtained at convergence may depend on the initial $\mathbf{w}$ used, it is recommended to employ a multi-start procedure in order to minimize the risk of retaining a local minimum.

was kept constant at 2 ($Q = 2$) with equally sized clusters. Next, the degree of congruence between the true cluster-specific components $\mathbf{t}_q^{true}$ ($q = 1,2$) was fixed at an intermediate level corresponding to a moderate amount of overlap between the components of .70 in terms of the Tucker congruence coefficient.[33] The chosen values for the fixed data aspects are realistic values for many empirical datasets.

The three factors that were introduced in Section 3.1 were systematically manipulated in a completely randomized three-factorial design:

1. Three degrees of congruence between the cluster-specific weighting systems $\mathbf{w}_q^{true}$ ($q = 1,2$): low congruence (.15 in terms of Tucker congruence), moderate congruence (.70) and perfect congruence (1). Note that perfect congruence implies an equality constraint on $\mathbf{w}_q^{true}$ (i.e., $\mathbf{w}_1^{true} = \mathbf{w}_2^{true} = \mathbf{w}^{true}$);
2. two levels for the nonnegativity of the loadings $\alpha_{j(q)}$: unconstrained versus constrained to be nonnegative (for both clusters);
3. two levels for the amount of noise in the data ε: .30 and .60. Data consisting of 30% of noise is realistic for empirical metabolomics data, whereas the 60% level was included to heavily challenge the method.

For a given combination of the levels of the three manipulated factors, a data set $\mathbf{X}$ was constructed as follows: First, a true partition matrix $\mathbf{P}^{true}$ ($J \times Q$) was constructed with $Q = 2$ equally sized clusters by assigning at random the correct number of variables to each cluster. Next, true loadings $\alpha_{j(q)}^{true}$ and true cluster-specific components $\mathbf{t}_q^{true}$ and weights $\mathbf{w}_q^{true}$ ($q = 1,2$) were generated. When nonnegativity was enforced, $\alpha_{j(q)}^{true}$, $\mathbf{t}_q^{true}$, and $\mathbf{w}_q^{true}$ were created by independently drawing entries from a uniform distribution from the interval [0 1]. When no nonnegativity constraint was imposed, entries were independently drawn from $U(-1,1)$ and thereafter, for each cluster separately, $\mathbf{t}_q^{true}$,$\mathbf{w}_q^{true}$ and the $\alpha_{j(q)}^{true}$'s were centered to a mean of zero. Note that nonnegative vectors were not centered as centering such a vector makes some of its values negative, which is unwanted here.

The amount of congruence between the different $\mathbf{t}_q^{true}$ and $\mathbf{w}_q^{true}$ ($q = 1,2$) was manipulated through the following two-step procedure. First, common base vectors $\mathbf{t}^{base}$ and $\mathbf{w}^{base}$ were generated in the same manner as explained above with an optional centering of the generated vector. Next, for each cluster separately, vectors $\mathbf{t}_q^{temp}$ and $\mathbf{w}_q^{temp}$ ($q = 1,2$) were simulated by drawing entries from $U(-2.5,2.5)$ or $U(-.60, .60)$ (and centering the vector) for a low (i.e., .15) and a moderate (.70) amount of overlap, respectively. Finally, $\mathbf{t}_q^{true}$ and $\mathbf{w}_q^{true}$ ($q = 1,2$) were obtained by adding $\mathbf{t}_q^{temp}$ and $\mathbf{w}_q^{temp}$ to $\mathbf{t}^{base}$ and $\mathbf{w}^{base}$, respectively, (e.g., $\mathbf{t}_q^{true} = \mathbf{t}^{base} + \mathbf{t}_q^{temp}$).

Next, $\mathbf{P}^{true}$, $\alpha_{j(q)}^{true}$, $\mathbf{t}_q^{true}$, and $\mathbf{w}_q^{true}$ ($q = 1,2$) were combined to determine the true array $\mathbf{T}$. Finally, a data array $\mathbf{X}$ was constructed by adding a noise array $\mathbf{E}$ (of size $100 \times 250 \times 7$) to $\mathbf{T}$. This noise array $\mathbf{E}$ was constructed in such a way that the data contained the requested amount of noise ε. To this end, the following three-step procedure was performed: (a) generating a noise matrix $\underline{\mathbf{E}}^{start}$ (of size $100 \times 250 \times 7$) by independently drawing numbers from $N(0,1)$ and centering $\underline{\mathbf{E}}^{start}$ along the first mode, (b) scaling $\underline{\mathbf{E}}^{start}$ such that $\|\underline{\mathbf{E}}^{start}\|_F^2 = \|\underline{\mathbf{T}}\|_F^2$ and (c) adding the scaled $\underline{\mathbf{E}}^{start}$ to $\mathbf{T}$ as $\underline{\mathbf{X}} = \underline{\mathbf{T}} + c\,\underline{\mathbf{E}}^{start}$. Regarding the last step, in order to ensure $\frac{\|\underline{\mathbf{E}}\|_F^2}{\|\underline{\mathbf{X}}\|_F^2}$ being equal to ε, c was set equal to $3/7$ and 1.5 for the 30% ($\varepsilon = .30$) and 60% ($\varepsilon = .60$) noise condition, respectively. It should be noted that by centering $\alpha_{j(q)}^{true}$, $\mathbf{t}_q^{true}$, and $\mathbf{w}_q^{true}$ and the error $\mathbf{E}$, each column for each slice of $\mathbf{X}$ has a mean of zero.

The procedure described above was repeated 10 times for each combination of the levels of the three manipulated factors. As such, 10 (replication)× 3 (congruence) × 2 (nonnegativity) × 2 (amount of noise) = 120 different data arrays $\mathbf{X}$ were obtained. Subsequently, four *CLV3W* analyses with the true number of clusters $Q = 2$—not allowing clusters to be empty—were performed on each data array $\mathbf{X}$ with 50 starts. For a single data set, on average, the four *CLV3W* analyses (with 50 starts) took 25 minutes in total (20-30 minutes for almost all data sets), with no consistent differences in computation time between the four analyses. As no information was *a-priori* given with regard to the distribution of the two clusters, the starts were obtained by randomly assigning variables to one of the two clusters and by imposing only that none of the clusters was empty. The four *CLV3W* analyses, which are listed in the rows of Tables 1 and 2, result from the combination of (a) whether or not enforcing nonnegativity on the loadings and (b) whether or not restricting the weighting systems to be equal across clusters. For each analysis, the default centering step included in the *ClustVarLV* package was applied. This step, which is tightly associated with the maximization of the sum of the squared covariances in criterion (1), consists of setting that mean across the samples to zero for each combination of a variable and an occasion/time point. The simulation study was run in R (version 3.5.1).

TABLE 1 Mean adjusted rand index (ARI) for the obtained attribute partition by each CLV3W analysis (rows) for the four types of true data sets (columns)

	True data types			
	Common	**Common_NN**	**Specific**	**Specific_NN**
Equal_NoNN	.8964	.9021	.6731	.5861
Equal_NN	−.0029	.9081	−.0025	.5763
NoEqual_NoNN	.8897	.9022	.9095	.9221
NoEqual_NN	−.0029	.9073	−.0025	.8643

Note. The four true data types (columns) are (1) *Common* = only equality (but no nonnegativity) constraint, (2) *Common_NN* = equality and nonnegativity constraint, (3) *Specific* = no equality and no nonnegativity constraint, and (4) *Specific_NN* = only nonnegativity (but no equality) constraint. The four *CLV3W* analyses (rows) are (1) *Equal_NoNN* = with equality (but no nonnegativity) constraint, (2) *Equal_NNt* = with equality and nonnegativity constraint, (3) *NoEqual_NoNN* = no equality and no nonnegativity constraint, and (4) *NoEqual_NN* = with nonnegativity (but no equality) constraint.

TABLE 2 Mean Tucker congruence coefficient for the obtained cluster-specific components (first value) and weighting systems (second value) by each CLV3W analysis (rows) for the four types of true data sets (columns)

	True data types			
	Common	**Common_NN**	**Specific**	**Specific_NN**
Equal_NoNN	1/.9994	1/.9995	.8298/.9710	.8074/.9596
Equal_NN	1/.9293	1/.9995	.8199/.9021	.8078/.9600
NoEqual_NoNN	1/.9994	1/.9995	.9875/.9963	1/.9995
NoEqual_NN	1/.9293	1/.9995	.8154/.9006	.9757/.9899

Note. The four true data types (columns) are (1) *Common* = only equality (but no nonnegativity) constraint, (2) *Common_NN* = equality and nonnegativity constraint, (3) *Specific* = no equality and no nonnegativity constraint and (4) *Specific_NN* = only nonnegativity (but no equality) constraint. The four *CLV3W* analyses (rows) are (1) *Equal_NoNN* = with equality (but no nonnegativity) constraint, (2) *Equal_NN* = with equality and nonnegativity constraint, (3) *NoEqual_NoNN* = no equality and no nonnegativity constraint, and (4) *NoEqual_NN* = with nonnegativity (but no equality) constraint.

3.3 | Results

3.3.1 | Recovery of the true partition of the attributes

To determine the degree to which the underlying variable partition has been recovered by the *CLV3W* algorithm, we calculated the adjusted rand index (ARI[34]) between the true variable partition $\mathbf{P}^{true}$ and the variable partition retained by the algorithm $\mathbf{P}$. An ARI value of one is encountered when both partitions are identical, whereas ARI equals zero when recovery is at chance level.

In Table 1, the mean ARI across all 120 data sets is displayed for each of the four true data types (columns) when analyzed by each of the four *CLV3W* analyses (rows). From this table, it appears that when the true weighting systems are equal to each other and the true components can take positive and negative values (i.e., column "Common"), a *CLV3W* analysis with equality constraint (mean ARI of .8964) only slightly outperforms an analysis without the equality constraint (.8897), as long as no nonnegativity constraint is imposed. When, however, enforcing nonnegativity, the obtained partitions become random (i.e., at chance level) as the mean ARI's are close to zero.

For the second data type in which the true data both have equal weighting systems and nonnegative components (i.e., column "Common_NN"), it appears that all *CLV3W* analyses recover the true variable partition to a large extent (i.e., mean ARI around .90). Restricting the components to be nonnegative performs a little bit better than not enforcing this constraint, whereas imposing the equality constraint does not really seem to improve cluster recovery.

Regarding the third data type (i.e., column "Specific": true weighting systems varying across clusters and no nonnegativity of components), a good cluster recovery (i.e., mean ARI of .9095) is obtained when no constraint is imposed. Erroneously constraining weighting systems to be equal across clusters seriously deteriorates cluster recovery (.6731),

whereas incorrectly restricting components to be nonnegative has an enormous negative impact on cluster recovery (i.e., mean ARI around zero, which indicates chance level).

Finally, when data have specific weighting systems and nonnegative components (i.e., column "Specific_NN"), imposing the equality constraint seriously hampers cluster recovery (i.e., mean ARI around .58). When allowing for cluster-specific weighting systems, adding a nonnegativity constraint, however, does not seem to increase cluster recovery (i.e., mean ARI of .86 for with versus .92 for without nonnegativity constraint).

In general, *CLV3W* with and without constraints recovers the underlying variable partition to a very large extent (i.e., mean ARI around .90) as long as the imposed constraints are in correspondence with the structure underlying the data. Erroneously enforcing nonnegativity of components has a larger negative impact on cluster recovery than incorrectly assuming equality of weighting systems. Further, cluster recovery also deteriorates, as could be expected, when the data contain more noise and when clusters show more overlap (not shown).

3.3.2 | Recovery of the true cluster-specific components and weighting systems

The extent to which *CLV3W* with and without constraints recovers the true cluster-specific components $\mathbf{t}_q^{true}$ and weighting systems $\mathbf{w}_q^{true}$ was evaluated by computing the Tucker congruence coefficient[33] between the true parameters ($\mathbf{t}_q^{true}$ and $\mathbf{w}_q^{true}$) and the estimated parameters ($\mathbf{t}_q$ and $\mathbf{w}_q$) ($q = 1,2$). A Tucker congruence value of one indicates that the recovery of the true cluster-specific parameters is perfect, whereas a value of zero implies that there is no recovery at all.

From Table 2, it appears that, for the first data type (i.e., column "Common"), recovery is excellent (i.e., mean Tucker around .99) as long as no nonnegativity constraint is enforced (.96). As such, enforcing equality of weighting systems is not needed for *CLV3W* in this situation to yield good recovery performance. Incorrectly imposing the nonnegativity constraint has negative consequences in terms of recovery performance for the weighting systems (i.e., mean Tucker of .93) but not for the components (i.e., mean Tucker of 1).

Further, for the second data type (i.e., column "Common_NN"), all *CLV3W* analyses recover the cluster-specific parameters to the same very large extent (i.e., mean Tucker of .999).

Next, regarding the third data type (i.e., column "Specific"), recovery is excellent when analyzing the data with (the original) *CLV3W* without constraints (i.e., mean Tucker of .999). Adding the equality (.90) or the nonnegativity (.86) constraint seems to deteriorate recovery performance, with this effect being stronger for the components than for the weighting systems.

Regarding the fourth data type (i.e., column "Specific_NN"), recovery is excellent (i.e., mean Tucker around .985) as long as no equality constraint is imposed on the weighting systems (.88). Restricting components to be nonnegative does not seem to have an impact on recovery performance. Again, these (negative) effects are larger for the components than for the weighting systems.

It can be concluded that, overall, constrained and unconstrained *CLV3W* recovers the cluster-specific components and weighting systems to a large extent when constraints are correctly enforced. Also, when cluster have more overlap and the data contain more noise, the deterioration in recovery is more pronounced (not shown). In sum, the decrease in cluster recovery is paralleled by a similar decrease in recovery of the components and weighting systems.

4 | CASE STUDIES

4.1 | Sensory case study: Emotions associated with coffee aromas

To illustrate the use of *CLV3W* with both constraints, we consider a case study pertaining to consumer emotions in which subjects rated a variety of 12 coffee aromas on 15 affective terms. Aromas were presented in pillboxes labeled with a random three-digit code. Eighty-four participants were asked to complete each rating on a 5-point rating scale. More details regarding the dataset with both the lists of aromas and affective terms can be found in Cariou and Wilderjans.[12]

The three-way Aromas × Consumers × Emotions data set was analyzed to identify potential segments of consumers, herewith varying the number of clusters Q from 1 to 9. Before analyzing, each products by emotions matrix $\mathbf{X}_j$ was column-wise centered and scaled to the same inertia, resulting in each data block having the same amount of influence

on the analysis. As it is important to lump together consumers who rate similarly the different products along the set of emotion terms, a nonnegativity constraint was imposed on the consumer scores. The evolution of the loss criterion against the number of clusters (see left-hand panel of Figure 1) led to retain a partition with two clusters as an elbow is observed in the plot at $Q = 2$.

The weighting system for the emotion terms for each cluster is presented in Figure 2 in ascending order. Remarkably, the distribution of the weights is basically the same for both clusters. This suggests to use the additional constraint of a common weighting scheme across the clusters (as proposed in Section 2.4).

Therefore, the *CLV3W-NN* model with the common weights constraint was fitted to the coffee aromas dataset. Again, the loss criterion against the number of clusters (see right-hand panel of Figure 1) led to retain a partition with two clusters (i.e., elbow at $Q = 2$). The first cluster was composed of 42 consumers, whereas the second one grouped 38 consumers. Two subjects were set aside as they had loading values equal to zero for both clusters. It is worth noting that this latter aspect is not due to the common weighting system constraint but rather to the nonnegativity constraint. The VAF obtained with this constrained model is equal to 22.5%, which is only a little smaller than the value of 22.7% for the unconstrained model. As such, the more parsimonious constrained model is preferred above the unconstrained one as it leads to a reduction in the number of parameters at the price of only a very small reduction in VAF. Finally, this negligible difference in VAF's demonstrates the relevance of the additional common weights constraint and

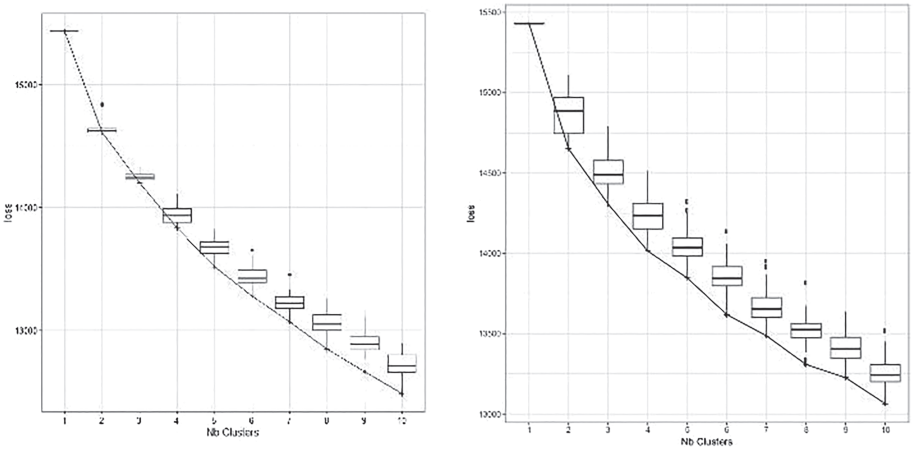


FIGURE 1 Evolution of the clustering around latent variables for three-way data (CLV3W)-NN loss value (A) versus the CLV3W-NN loss value with a common weights constraint (B) across increasing numbers of clusters varying from 1 up to 10 for the consumer data; boxplots indicate the variability in loss functions values encountered across 50 random starts

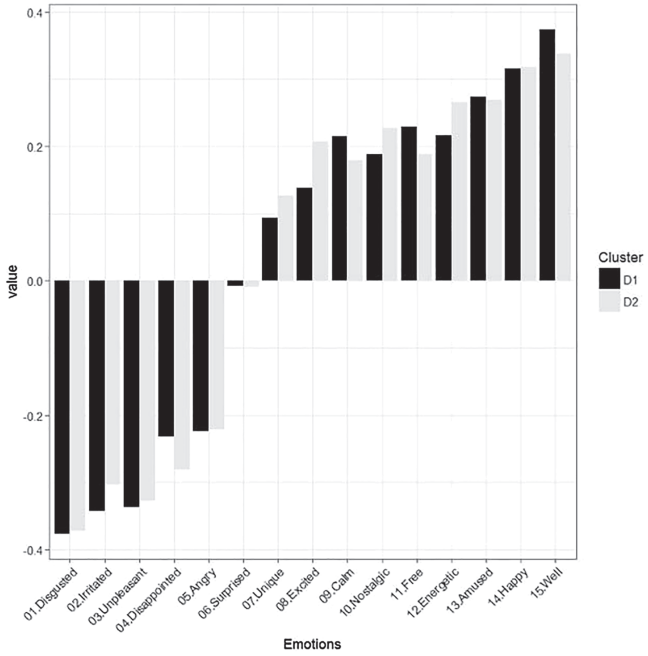


FIGURE 2 Attribute weights for the two-cluster clustering around latent variables for three-way data (CLV3W)-NN solution for the consumer data. D1 (resp. D2) corresponds to the weighting system of the first (resp. second) cluster

indicates that the two clusters of consumers mainly differ in the way the consumers rated the set of products. On the whole, all consumers weighted the different emotion terms similarly when rating the products.

When inspecting the common weights system presented in Figure 3, it appears that the distribution clearly reflects a bipolar pleasantness/unpleasantness dimension. As such, a single emotion dimension going from pleasant to unpleasant is underlying the ratings of all subjects irrespective of the cluster to which they belong. With regard to such a distribution, one can highlight three groups of emotional terms: (a) unpleasantness with the largest negative values (e.g., angry, unpleasant, disappointed, irritated, disgusted), (b) neutral emotion terms with a weight close to zero (e.g., surprised), and (c) those emotion terms corresponding to pleasantness emotions (i.e., nostalgic, calm, excited, unique, energetic, free, well, and happy).

Compared with the initial *CLV3W* algorithm, it should be stressed that using a single vector of weights common to all clusters yields a configuration that is easier to interpret. Indeed, the vertical slices of the three-way array can be thereafter aggregated on the basis of this overall weighting scheme and concatenated into a two-way matrix. When performing principal component analysis (PCA) on this matrix, the variable clusters can be easily represented in scores and loadings plots as shown in Figure 4.

Figure 4 depicts the first factorial plane for both clusters from a PCA on the matrix resulting from the aggregation of the data array along the third mode, according to the common weighting system. In Figure 4, both the products and the subjects—represented as arrows—are displayed. When inspecting the product scores, three sets of coffee aroma products can be identified. A first set of products, consisting of Basmati rice, cedar, earth, and medicinal, has a negative score for both clusters. These products are associated with negative emotions such as disgusted and unpleasant. Secondly, apricot, flower, coffee, and lemon aromas are encountered with positive scores for both clusters, and these aromas are related to pleasantness emotions. It appears that consumers (clusters) more or less agree on the rating of both sets of products. The remaining products demonstrate the main differences between the two consumer clusters in the evaluation of the aromas. Although hay, hazelnut, honey, and vanilla yield positive emotions with regard to the first consumer cluster, they yield negative emotions for the second consumer cluster. On the contrary, coriander seed yields positive emotions with regard to the second consumer cluster, whereas it yields negative emotions for the first cluster.

4.2 | Metabolomics case study: Mother milk lipidome dataset

This case study is part from a larger study whose aim was to fully characterize human breast milk composition in the context of prematurity and to investigate its beneficial effect on preterm infant postnatal growth.[35] As energy intake has been identified as a main independent factor associated with infant growth until the age of 2 years,[36] with lipids being the main contributor of energy requirements of healthy infants, the present preterm breast milk phenotyping was focused on lipidomics data. In this context, the application of *CLV3W* aimed at (a) identifying subsets of metabolites (all lipid species) that are highly positively correlated and, simultaneously, (b) exhibiting the longitudinal trajectory associated with each group of metabolites.

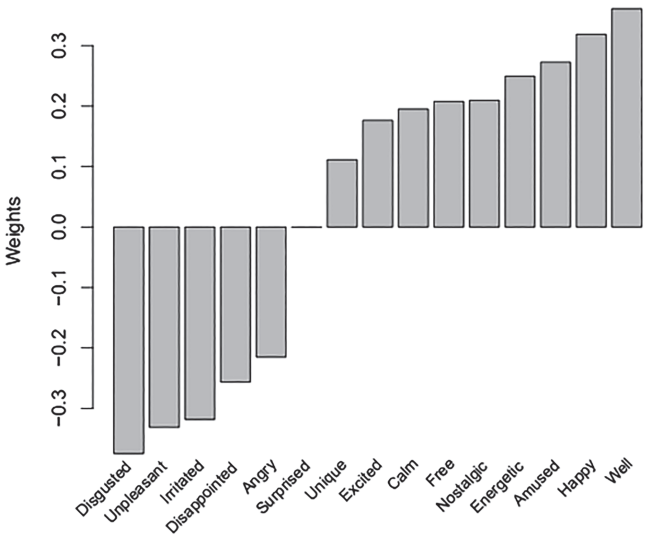


FIGURE 3 Attribute weights for the two-cluster solution with equality and nonnegativity constraints, for the consumer data

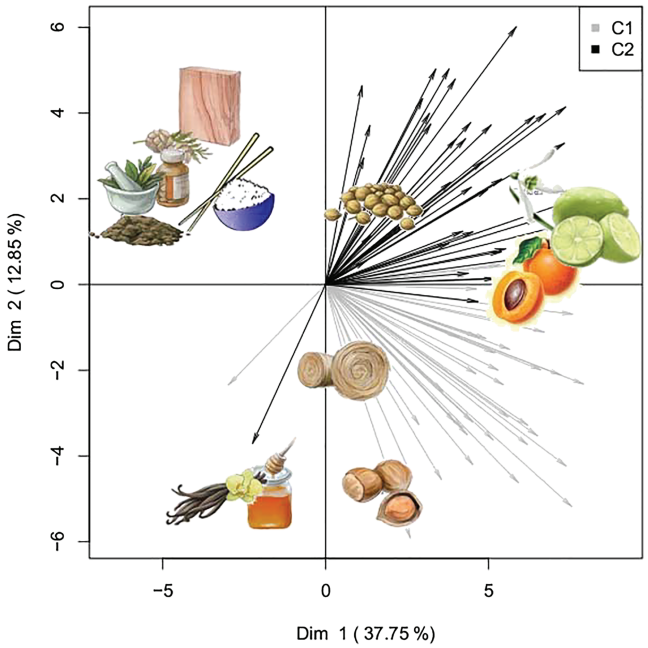


FIGURE 4 Configuration of the products from the principal component analysis (PCA) of the aggregated array, for the two-cluster solution with equality and nonnegativity constraints, for the consumer data

The dataset corresponded to eight preterm infants, selected from the 118 infants enrolled in the mono-centric prospective LACTACOL cohort (registered at www.clinicaltrails.gov as #NCT01493063) in the University Hospital of Nantes, France. During their hospital stay, these preterm infants received, for a minimum of 28 days, their own mother's breast milk only. Weekly breast milk samples were collected for each mother during the infant hospitalization. To characterize milk composition, their comprehensive lipidomic signatures were obtained using liquid chromatography-electrospray ionization (positive/negative)-high-resolution-mass spectrometry (LC-ESI HR-MS)-based cross-platform phenotyping. Before performing any multivariate analyses, these raw data were preprocessed with the open-source XCMS[37] software combined with CAMERA[38] and finally a filtering step made it possible to retain 3451 and 903 (*m/z*; RT) features (metabolites) in positive and negative mode, respectively. In the present pilot study, we focused on the 903 features acquired in negative mode. For further details on this study, see Alexandre-Gouabau et al.[35] As hospitalization stays varied among infants, five time points were kept from week 2 to week 6. Eight infants were considered, resulting in the analysis of a three-way array of eight infants × 903 metabolites × five time points.

Lipidomic LC-ESI-HRMS variables were analyzed with *CLV3W*, considering also the two constraints. In Figure 5, for *CLV3W* and each combination of the two additional constraints, the evolution of the loss function is depicted against the number of clusters varying from one to 15. Obviously, when only selecting a single cluster, the loss values are equal for both the solution with and without common weights constraint. More interestingly, the difference in loss values induced by the nonnegativity constraint is quite negligible (especially when comparing the models without common weight constraint) whatever the number of clusters considered. This clearly indicates that metabolites within each cluster are mainly positively correlated to their latent component—and to each other—across the five time points. Comparing the values of the loss function between the unconstrained *CLV3W* solution and the common weights *CLV3W* solution, it appears that imposing such a constraint implies a substantial decrease in the quality of the model irrespective of the number of clusters (for $Q > 1$) inspected. This demonstrates undeniably that the set of metabolites under study does not have the same kinetics of lactation over time, which is in agreement with the large breast milk variability, and in particular its lipid component, reported during the first 3 months of lactation.[39,40] In the remainder, the *CLV3W-NN* solution without equal weights constraint is therefore retained and discussed.

Applying CHull, as can be seen in Figure 6, a solution with three clusters is retained both when taking the *CLV3W-NN* loss function as model misfit measure (see also the line with red triangles in Figure 5) as when taking the VAF as model fit measure. The *st*-scree values for the *CLV3W-NN* loss function (i.e., *st*-values of 0, 1.97, 2.73, 1.29, 1.68, 1.41, 1.03, 1.24, 1.49, and 0 for $Q = 1, 2, 3, 4, 6, 8, 9, 12, 13$, and 15, respectively) and the VAF (i.e., *st*-values of 0, 1.67, 4, 1.5, 1, 2, 1, 1, and 0 for $Q = 1, 2, 3, 5, 6, 9, 11, 13$ and 15, respectively) indicate that the optimal number of clusters Q equals three. As an indication, with 903 metabolites to partition into three clusters and given a multi-start procedure with 50 initializations, the computational cost of the *CLV3W* algorithm varied from 2.4 minutes for the model with

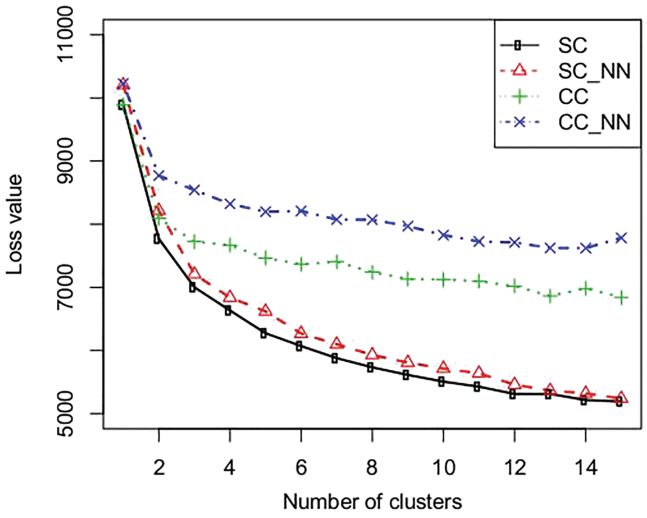


FIGURE 5 Evolution of the clustering around latent variables for three-way data (CLV3W) (SC) loss value together with those obtained with CLV3W-NN (SC_NN) versus CLV3W with the common weights constraint (CC, CC_NN), for increasing numbers of clusters varying from 1 up to 15, for the metabolomics data

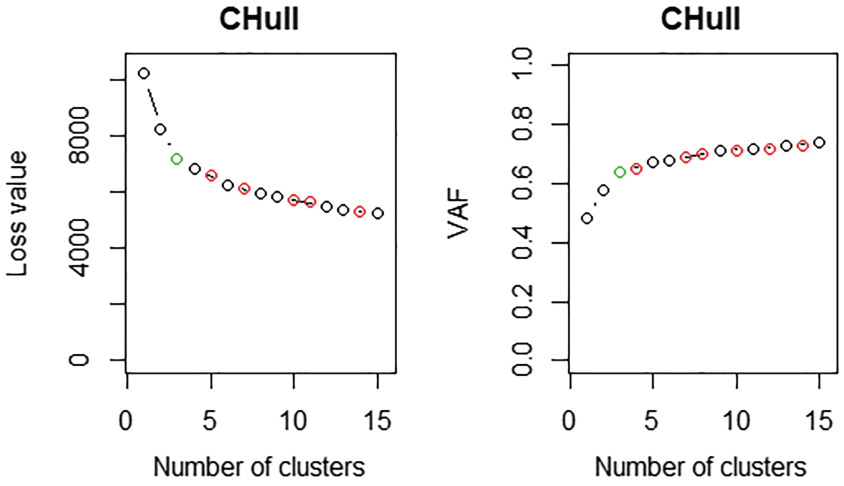


FIGURE 6 CHull analysis for the metabolomics data set based on plotting the number of clusters Q against the clustering around latent variables for three-way data (CLV3W)-NN loss value (left panel) and the variance accounted for (VAF) (right panel). Each model with a different Q is presented by a black (when at the boundary of the convex hull, marked by a black line) or red (not at the boundary) circle. The optimal model is indicated by a green circle

nonnegativity and common weights constraints to 13.6 minutes for the model with no constraints on a standard computer (16 Go RAM).

The partition of the metabolites led to three clusters composed of 327, 412, and 160 metabolites, respectively. It is worth noting that four metabolites revealed loading values equal to zero whatever the cluster considered. The VAF corresponding to the three cluster solution is equal to 64%. Figure 7 depicts the first factorial plane from a greedy-Parafac procedure[41] (i.e., sequential rank-one decomposition with orthogonality constraint on the first mode) carried out on the three-way array. One can clearly see that the first cluster is oriented along the first bisector, the second one is mainly oriented along the first Parafac dimension, whereas the third one is mainly oriented along the second dimension. The VAF for the first two dimensions equals 68%.

An inspection of the distributions of time point weights across the three clusters, which are displayed in Figure 8, shows positive weights for all clusters whatever the lactation time point. The first cluster shows a down-weighting of the second lactation point. With regard to the second cluster, one can see an increasing trajectory from the first point of lactation to the fourth one and a relative down-weighting of the fifth time point. On the contrary, the third cluster shows a decrease in the weighting of the successive lactation points from the first to the fifth one and finally an up-weighting of the fifth lactation time point.

On the whole, the fact that two opposite lactaction time trajectories are exhibited by the *CLV3W* with nonnegativity constraint procedure explains to a large extent the poor performance of *CLV3W* with a common weights constraint. To conclude, such an analysis makes it possible to identify subgroups of metabolites that have specific lactation time trajectories and that are highly positively correlated within each subgroup. Interestingly, on the basis of the few annotated lipid species,[35] the first cluster was mainly composed of milk fat globules membrane-structural constituents, whereas the second cluster included many mediators or signaling lipids (such as eicosanoids and endocannabinoids) involved in

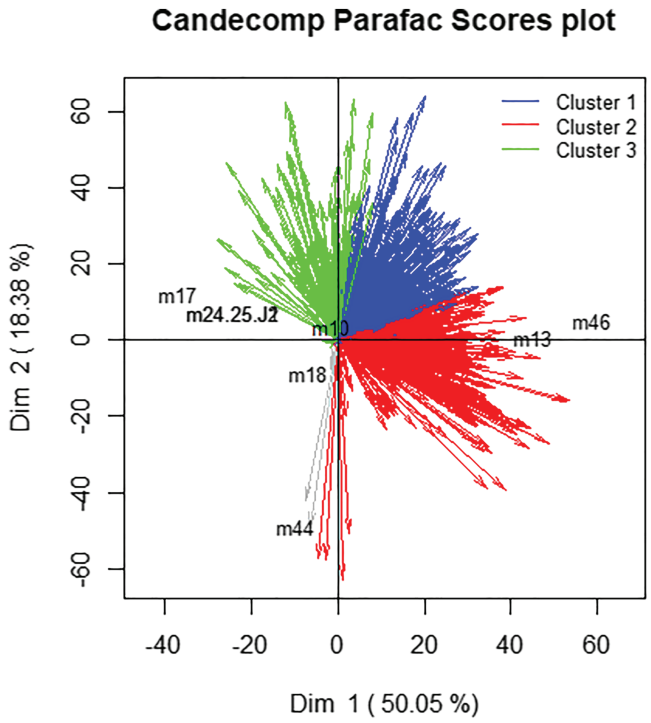


FIGURE 7 Configuration of the observations, with the three groups of metabolites (represented as arrows) regressed onto the greedy-Parafac first two components associated with the first mode. Gray arrows represent metabolites set aside from the three-cluster solution for the metabolomics data

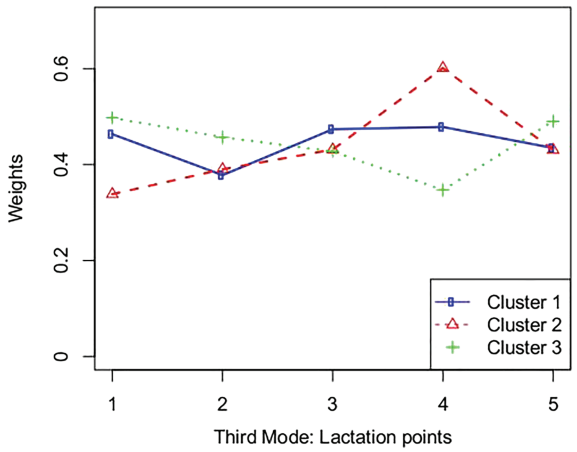


FIGURE 8 Weighting systems associated with the third mode of the three-cluster clustering around latent variables for three-way data (CLV3W)-NN solution, for the metabolomics data

energy homeostasis, insulin sensitivity, defense against oxidative stress or gastrointestinal tract infection, or inflammation. Their up-weighting during the first month of lactation is in agreement with their putative beneficial effects in health outcomes for preterm infant. Finally, the third cluster brought together several fatty acid oxidation-products, whose down-weighting may illustrate decreased mammary gland energy requirements for the establishment of lactation and production of mature milk. Ongoing work should focus on the further characterization and validation of these subgroups.

5 | CONCLUSION

In the clustering of variables framework, given a three-way data structure, Wilderjans and Cariou[11] proposed a clustering strategy, *CLV3W*, that aims at identifying clusters of variables and simultaneously a latent component and a weighting system associated to each cluster. Based on reasons of interpretation and substantive grounds, two constraints have been proposed and integrated into *CLV3W* resulting in (a) a model which make it possible to take some user-knowledge regarding the correlation between variables within a cluster (i.e., nonnegativity constraint) into account and (b) a more parsimonious model (i.e., common weights system). The usefulness of the constraints has been

illustrated both in simulated data and in two case studies pertaining to sensory analysis and metabolomics. In the latter case study, the application of constrained *CLV3W* facilitated the discovery of metabolite groups that show differential time trajectories in lactation profile.

Ongoing research concerns the determination of the number of clusters. As in the classical case, this can be done with a generalized version of the scree test of Cattell,[42] in which, for the solutions under consideration, the loss function value—which functions as a (mis)fit measure—is plotted against the number of clusters (i.e., model complexity). Other strategies that generalize cluster validation criteria, like the Silhouette[43] and Calinski-Harabasz index,[44] to the case of three-way structures would help the user in determining the optimal number of clusters for a data set at hand. In the same vein, further research is needed regarding criteria to help the user to decide whether or not imposing the common weights system and/or the nonnegativity constraint is appropriate for the data at hand.

Finally, further comparisons are needed with other clustering strategies for three-way data, such as simultaneous decomposition and clustering models[45,46] and model-based clustering.[47,48]

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/CEM.3269.

ORCID

Véronique Cariou https://orcid.org/0000-0003-4091-5910
Marie-Cécile Alexandre-Gouabau https://orcid.org/0000-0002-9966-9593
Tom F. Wilderjans https://orcid.org/0000-0002-1677-4938

REFERENCES

1. Boccard J, Rudaz S. Harnessing the complexity of metabolomic data with chemometrics. *J Chemometr*. 2014;28(1):1-9.
2. Wold S, Kettaneh N, Tjessem K. Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *J Chemometr*. 1996;10(5-6):463-482.
3. Smilde AK, Westerhuis JA, de Jong S. A framework for sequential multiblock component methods. *J Chemometr*. 2003;17(6):323-337.
4. Acar E, Bro R, Smilde AK. Data fusion in *metabolomics* using coupled matrix and tensor factorizations. *Proc IEEE*. 2015;103(9): 1602-1620.
5. Biancolillo A, Næs T. The sequential and orthogonalized PLS regression for multiblock regression: theory, examples, and extensions. In: Cocchi M, ed. *Data Handling in Science and Technology*. Vol. 31. Oxford, UK: Elsevier; 2019:157-177.
6. Cariou V, Bouveresse DJR, Qannari EM, Rutledge DN. ComDim methods for the analysis of multiblock data in a data fusion perspective. In: Cocchi M, ed. *Data Handling in Science and Technology*. Vol.31. Oxford, UK: Elsevier; 2019:179-204.
7. Smilde AK, Westerhuis JA, Hoefsloot HCJ, et al. Dynamic metabolomic data analysis: a tutorial review. *Metabolomics*. 2010;6(1):3-17.
8. Vigneau E, Qannari EM. Clustering of variables around latent components. *Communications in Statistics—Simulation and Computation*. 2003;32(4):1131-1150.
9. Vigneau E, Sahmer K, Qannari EM, Bertrand D. Clustering of variables to analyze spectral data. *J Chemometr*. 2005;19(3):122-128.
10. Cuny M, Vigneau E, Le Gall G, Colquhoun I, Lees M, Rutledge DN. Fruit juice authentication by 1 H NMR spectroscopy in combination with different chemometrics tools. *Anal Bioanal Chem*. 2008;390(1):419-427.
11. Wilderjans TF, Cariou V. CLV3W: a clustering around latent variables approach to detect panel disagreement in three-way conventional sensory profiling data. *Food Quality and Preference*. 2016;47:45-53. https://doi.org/10.1016/j.foodqual.2015.03.013
12. Cariou V, Wilderjans TF. Consumer segmentation in multi-attribute product evaluation by means of non-negatively constrained CLV3W. *Food Quality and Preference*. 2018;67:18-26. https://doi.org/10.1016/j.foodqual.2017.01.006
13. HastieT, Tibshirani R, Eisen MB, et al. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol*. 2000;1(2):research0003.1. https://doi.org/10.1186/gb-2000-1-2-research0003
14. Kiers HAL. Towards a standardized notation and terminology in multiway analysis. *J Chemometr*. 2000;14(3):105-122.
15. Wilderjans TF, Ceulemans E. Clusterwise Parafac to identify heterogeneity in three-way data. *Chemom Intel Lab Syst*. 2013;129:87-97.
16. Krijnen WP. *The analysis of three-way arrays by constrained PARAFAC methods*. Leiden, The Nederlands: DSWO Press; 1993.
17. Hitchcock FL. The expression of a tensor or a polyadic as a sum of products. *J Math Phys*. 1927;6(1):164-189.
18. Carroll JD, Chang JJ. Analysis of individual differences in multidimensional scaling via an N-way generalization of “Eckart-Young” decomposition. *Psychometrika*. 1970;35(3):283-319.
19. Harshman RA. Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics*. 1970;16:1-84.
20. Ceulemans E, Van Mechelen I, Leenen I. The local minima problem in hierarchical classes analysis: an evaluation of a simulated annealing algorithm and various multistart procedures. *Psychometrika*. 2007;72(3):77-391. https://doi.org/10.1007/s11336-007-9000-9
21. Steinley D. Local optima in K-means clustering: what you don't know may hurt you. *Psychol Methods*. 2003;8(3):294-304.
22. Steinley D. Profiling local optima in K-means clustering: developing a diagnostic technique. *Psychol Methods*. 2006;11(2):178-192.

23. Ceulemans E, Kiers HAL. Selecting among three-mode principal component models of different types and complexities: a numerical convex hull based method. *British Journal of Mathematical & Statistical Psychology*. 2006;59(1):133-150. https://doi.org/10.1348/000711005X64817
24. Wilderjans TF, Ceulemans E, Meers K. CHull: a generic convex hull based model selection method. *Behav Res Methods*. 2013;45(1):1-15. https://doi.org/10.3758/s13428-012-0238-5
25. Bro R, de Jong S. A fast non-negativity-constrained least squares algorithm. *J Chemometr*. 1997;11(5):393-401.
26. Smilde AK, Bro R, Geladi P. Visualization. In: *Multi-Way Analysis with Applications in the Chemical Sciences*. Chichester, England: John Wiley & Sons; 2004:175-220.
27. Vigneau E, Chen M, Qannari EM. ClustVarLV: an R package for the clustering of variables around latent variables. *The R Journal*. 2015;7(2):134-148.
28. Wilderjans TF, Ceulemans E, Kuppens P. Clusterwise HICLAS: a generic modeling strategy to trace similarities and differences in multi-block binary data. *Behav Res Methods*. 2012;44(2):532-545. https://doi.org/10.3758/s13428-011-0166-9
29. De Roover K, Ceulemans E, Timmerman ME, Vansteelandt K, Stouten J, Onghena P. Clusterwise simultaneous component analysis for analyzing structural differences in multivariate multiblock data. *Psychol Methods*. 2012;17(1):100-119. https://doi.org/10.1037/a0025385
30. Wilderjans TF, Ceulemans E, Van Mechelen I. The SIMCLAS model: simultaneous analysis of coupled binary data matrices with noise heterogeneity between and within data blocks. *Psychometrika*. 2012;77(4):724-740. https://doi.org/10.1007/S11336-012-9275-3
31. Wilderjans TF, Ceulemans E, Van Mechelen I, van den Berg RA. Simultaneous analysis of coupled data matrices subject to different amounts of noise. *British Journal of Mathematical and Statistical Psychology*. 2011;64(2):277-290. https://doi.org/10.1348/000711010X513263
32. Brusco MJ, Cradit JD. Conpar: a method for identifying groups of concordant subject proximity matrices for subsequent multidimensional scaling analyses. *Journal of Mathematical Psychology*. 2005;49(2):142-154. https://doi.org/10.1016/j.jmp.2004.11.004
33. Tucker LR. *A Method for Synthesis of Factor Analysis Studies. Personnel Research Section Rapport # 984*. Washington, DC, USA: Department of the Army; 1951.
34. Hubert L, Arabie P. Comparing partitions. *Journal of Classification*. 1985;2(1):193-218.
35. Alexandre-GouabauMC, Moyon T, Cariou V, et al. Breast milk lipidome is associated with early growth trajectory in preterm infants. *Nutrients*. 2018;10(2):E164. https://doi.org/10.3390/nu10020164
36. Hiltunen H, Loyttyniemi E, Isolauri E, Rautava S. Early nutrition and growth until the corrected age of 2 years in extremely preterm infants. *Neonatology*. 2018;113(2):100-107.
37. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem*. 2006;78(3):779-787.
38. Kuhl C, Tautenhahn R, Bottcher C, Larson TR, Neumann S. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal Chem*. 2011;84(1):283-289.
39. Bardanzellu, F, Peila, C, Fanos, V, Coscia, A. Clinical insights gained through metabolomic analysis of human breast milk. *Expert Review of Proteomics* 2019: accepted.
40. Neville MC, Picciano MF. Regulation of milk lipid secretion and composition. *Annu Rev Nutr*. 1997;17(1):159-184.
41. Kolda TG, Bader BW, Kenny JP. Higher-order web link analysis using multilinear algebra. Fifth IEEE International Conference on Data Mining (ICDM'05) (pp. 8), Houston, TX, USA. 2005. https://doi.org/10.1109/ICDM.2005.77
42. Cattell RB. The meaning and strategic use of factor analysis. In: Cattell RB, ed. *Handbook of multivariate experimental psychology*. Chicago, IL, USA: Rand McNally; 1966:174-243.
43. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*. 1987;20:53-65. https://doi.org/10.1016/0377-0427(87)90125-7
44. Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*. 1985;50(2):159-179.
45. Vichi M, Rocci R, Kiers HAL. Simultaneous component and clustering models for three-way data: within and between approaches. *Journal of Classification*. 2007;24(1):71-98.
46. LlobellF, Cariou V, Vigneau E, Labenne A, Qannari EM. Analysis and clustering of multiblock datasets by means of the STATIS and CLUSTATIS methods. *Application to Sensometrics Food Quality and Preference*. 2020;79:103520.
47. Basford KE, McLachlan GJ. The mixture method of clustering applied to three-way data. *Journal of Classification*. 1985;2(1):109-125.
48. Viroli C. Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing*. 2011;21(4):511-522.

How to cite this article: Cariou V, Alexandre-Gouabau M-C, Wilderjans TF. Three-way clustering around latent variables approach with constraints on the configurations to facilitate interpretation. *Journal of Chemometrics*. 2021;35:e3269. https://doi.org/10.1002/cem.3269