



HAL
open science

Benchmarking multiblock methods with canonical factorization

Stéphanie Bougeard, Caroline Peltier, Benoît Jaillais, Jean-Claude Boulet,
Mohamed Hanafi

► **To cite this version:**

Stéphanie Bougeard, Caroline Peltier, Benoît Jaillais, Jean-Claude Boulet, Mohamed Hanafi. Benchmarking multiblock methods with canonical factorization. *Chemometrics and Intelligent Laboratory Systems*, 2024, 254, pp.105240. 10.1016/j.chemolab.2024.105240 . hal-04945817

HAL Id: hal-04945817

<https://hal.science/hal-04945817v1>

Submitted on 18 Feb 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Benchmarking multiblock methods with canonical factorization

Stéphanie Bougeard ^{a*}, Caroline Peltier ^b, Benoit Jaillais ^c, Jean-Claude Boulet ^d, Mohamed Hanafi ^c

^a *Epidemiology and Welfare, ANSES, Ploufragan, France*

^b *Centre des Sciences du Goût et de l'Alimentation, CNRS, INRAE - PROBE research infrastructure, ChemoSens facility, Institut Agro, University of Bourgogne, Dijon, France*

^c *Oniris, INRAE, StatSC, Nantes, France*

^d *SPO, INRAE - PROBE research infrastructure, PFP polyphenols analytical facility, Institut Agro Montpellier, Univ. Montpellier, France*

Abstract. Data measured on the same observations and organized in blocks of variables — from different measurement sources or deduced from topics specified by the user — are common in practice. Multiblock exploratory methods are useful tools to extract information from data in a reduced and interpretable common space. However, many methods have been proposed independently and the users are often lost in selecting the appropriate one, especially as they do not always lead to the same results or because outputs do not have the same form. For this purpose, the data decomposition by canonical factorization was introduced thus applied to some widely-used methods, CPCA, MCOA, MFA, STATIS and CCSWA. The methods were compared on simulated (resp. real) data whose structure is controlled (resp. known). Theoretical and practical results pinpoint that the block-structure must be carefully explored beforehand. The number of block-variables and the block-variance distribution along dimensions impacts the choice of the block-scaling. The observation-structure within and between blocks impacts the choice of the method. CPCA or MCOA mix common and specific information, STATIS highlights common structure only whereas CCSWA focuses on specific information. To enable these diagnoses, methods and proposed comparison tools are available on **R**, **Matlab** or **Galaxy**.

Keywords. data integration, factorization, multiblock data decomposition, exploratory multiblock analysis, principal component analysis

Highlights.

- Canonical factorization gives a unified framework to compare multiblock methods.
- Multiblock methods retrieve the block-structure with specificities in terms of block-scaling and importance given to common and specific information.
- Exploring data before choosing the multiblock method is highly recommended.
- Methods and comparison tools are available in **R**, **Matlab** and **Galaxy**.

1 Introduction

Data sets organized in blocks of variables measured on the same observations are common in practice. Blocks may correspond to (quantitative) variables from different measurement sources or to topics specified by the user. For example in omics data, blocks consist of variables obtained from different techniques (e.g., proteomics, metabolomics). In food science, relationships between physico-chemical measurements, microbiological characterization and sensory attributes can be explored. The aims of exploratory analysis of multiblock data are manifold: (i) jointly reduce the dimensions of multiple blocks, (ii) investigate relationships between blocks, (iii) and between variables within and between blocks, and (iv) recover within-block variation to highlight common and specific block-structure in a common space. For this purpose, multiblock exploratory methods are appropriate tools. Many methods have been proposed independently in the literature within different application frameworks: Generalized Canonical Correlation Analysis (GCCA) [5] and its popular case SUMCOR [17], Generalized Procrustes Analysis (GPA) [12] MAXBET [37] Consensus Principal Component Analysis (CPCA) [42], Multiblock Principal Component Analysis (MBPCA) [42], Multiple Factorial Analysis (MFA) [9], Structuration de Tableaux A Trois Indices de la Statistique (STATIS) [20], Multiple CO-inertia Analysis (MCOA) [7], Hierarchical Principal Component Analysis (HPCA) [41, 44], COMDIM also known as Common Components and Specific Weights Analysis (CCSWA) [27], or SUM-PCA [32]. Users are often lost in choosing the appropriate method because outputs do not have the same form or do not lead to the same results.

Compared to the number of methods, their interconnection has been little investigated. Some of them (e.g., MBPCA, HPCA) were not described with a criterion but with an iterative algorithm, while some others (e.g., MCOA, MFA) derived from eigendecomposition. An integrative analysis of some methods (SUM-PCA, PCovR, MFA, STATIS, SCA-P) based on Simultaneous Component Analysis (SCA) has been proposed [38]. A monotonicity property of HPCA was revealed and an optimization criterion was presented to show equivalence between HPCA and CCSWA [15]. A new formulation of CCSWA was introduced with a criterion similar to that of MCOA or CPCA [14]. New properties of CPCA revealed its connection with MCOA and PCA [16]. Some methods (e.g., GCCA, SUMCOR) can be considered as special cases of Regularized Generalized Canonical Correlation Analysis (RGCCA) when the concatenated block is considered as the dependent one [35]. GCCA, CCSWA and HPCA were considered in a unified framework [33]. Despite these clarifications, users still need to compare methods theoretically and, most importantly, get a practical guide to choose the appropriate method.

Our first aim is to reformulate the outputs of widely-used methods, CPCA, MCOA, MFA, STATIS and CCSWA (equivalent to COMDIM, ACCPS, HPCA). For this purpose, we proposed to introduce data decomposition by canonical factorization to each method. Benefits are twofold: (i) standardization of method outputs — comparable to those of standard PCA — and (ii), for a given method, relation between overall- and block outputs. Our second aim is to simulate different data and compare multiblock methods on these outputs. Without any relevant model, it is not possible to simulate multiblock data properly. Thanks to canonical factorization, it is. Our final aim is to apply multiblock methods to real data in order to move towards a clear user guideline. The rest of this article is organized as follows. In Section 2, the notion of canonical factorization is given (Section 2.2) and then applied to multiblock methods (Section 2.3). The

56 way to simulate and compare methods is given (Section 2.4). In Section 3, multiblock methods
57 were compared on a simulation study (Section 3.1) and on data pertaining to multiple data
58 integration for food (Section 3.2). In Section 4, results are summarized and perspectives for
59 future work are drawn.

60 2 Method

61 2.1 Notations

62 Matrices are denoted by bold upper-case letters (\mathbf{X}) and vectors by bold lower-case letters (\mathbf{x}).
63 \mathbf{X}^T denotes the transpose operation of a matrix \mathbf{X} . For a square matrix \mathbf{X} , $\text{trace}(\mathbf{X})$ is the sum of
64 diagonal elements of \mathbf{X} and $\text{diag}(\mathbf{x})$ is a diagonal matrix whose diagonal elements are elements
65 of \mathbf{x} . $\|\mathbf{X}\|$ is the Frobenius norm of \mathbf{X} . Multiblock data are described with N observations and
66 K blocks ($\mathbf{X}_1, \dots, \mathbf{X}_K$) of (J_1, \dots, J_K) variables, with $J = \sum_{k=1}^K J_k$. The concatenated data is $\mathbf{X} =$
67 $[\mathbf{X}_1 | \dots | \mathbf{X}_K]$, and matrix rank is $\text{rank}(\mathbf{X}) = H$. Without loss of generality, variables are assumed to
68 be column-centered. The concatenated data can be decomposed with components and loading
69 by $\mathbf{X} = \tilde{\mathbf{T}}\tilde{\mathbf{P}}^T$ (PCA) or with standardized ones by $\mathbf{X} = \mathbf{T}\mathbf{D}\mathbf{P}^T$ (SVD). Let $\mathbf{T} = [\mathbf{t}^{(1)}, \dots, \mathbf{t}^{(H)}]$ be
70 standardized common components of size ($N \times H$) ($\|\mathbf{t}^{(h)}\| = 1$ for $h = 1, \dots, H$ with $\mathbf{t}^{(h)}\mathbf{t}^{(h')T} = 0$
71 for $h = 1, \dots, H, h' = 1, \dots, H, h \neq h'$), \mathbf{D} the diagonal matrix of scaling of size ($H \times H$) and
72 $\mathbf{P} = [\mathbf{p}^{(1)}, \dots, \mathbf{p}^{(H)}]$ standardized common loadings of size ($J \times H$) ($\|\mathbf{p}^{(h)}\| = 1$ for $h = 1, \dots, H$).

73 2.2 Canonical factorization of multiblock data

74 **Rationale** The canonical factorization of data by a multiblock method consists of assigning a
75 standardized decomposition to each block, concatenated data included. Canonical factoriza-
76 tion is the decomposition of ($K + 1$) blocks by a product of matrices with various constraints,
77 depending on multiblock methods. This seek a common mathematical concept, of which each
78 multiblock method — associated with a criterion to be optimized — is a particular implemen-
79 tation associated with unique data decompositions. The originality of canonical factorization
80 is manifold: (i) it decomposes multiblock data into common and thus comparable parameters,
81 (ii) it highlights relationships between overall analysis (decomposition of \mathbf{X}) and block-analyses
82 (decompositions of $\mathbf{X}_1, \dots, \mathbf{X}_K$), (iii) common and block-parameters have a statistical and geo-
83 metrical interpretation which clarifies the strategy adopted by methods for analysing multiblock
84 data.

85 **Proposal** The canonical factorization of multiblock data is the decomposition of the ($K + 1$)
86 blocks — ($\mathbf{X}_1, \dots, \mathbf{X}_K$) and \mathbf{X} — following:

$$\left\{ \begin{array}{l} \mathbf{X} = \mathbf{T}\mathbf{D}\mathbf{P}^T \\ \mathbf{X}_k = \mathbf{T}\mathbf{D}_k\mathbf{P}_k^T \quad \text{for } k = (1, \dots, K) \\ \text{with the constraints } \mathbf{T}^T\mathbf{T} = \mathbf{I} \quad \text{and} \quad \|\mathbf{p}^{(h)}\| = \|\mathbf{p}_k^{(h)}\| = 1 \quad \text{for } h = (1, \dots, H), \end{array} \right. \quad (1)$$

87 with \mathbf{T} the standardized and orthogonal common components, \mathbf{P} the standardized (not neces-
 88 sarily orthogonal) common loadings, \mathbf{P}_k the standardized block-loadings, \mathbf{D} and \mathbf{D}_k the diagonal
 89 scaling matrices with property $\mathbf{D}^2 = \sum_{k=1}^K \mathbf{D}_k^2$ that relate common and block-analyses.

90 **Proof** The overall data are decomposed into $\mathbf{X} = \tilde{\mathbf{T}}\tilde{\mathbf{P}}^T$, e.g., with a PCA. The two matrices are
 91 standardized following $\tilde{\mathbf{T}} = \mathbf{T}\mathbf{D}_T$ and $\tilde{\mathbf{P}}^T = \mathbf{D}_P\mathbf{P}^T$, the scaling matrices \mathbf{D}_T and \mathbf{D}_P being both
 92 diagonal of size $(H \times H)$. The canonical factorization of \mathbf{X} is given by:

$$\mathbf{X} = \mathbf{T}(\mathbf{D}_T\mathbf{D}_P)\mathbf{P}^T = \mathbf{T}\mathbf{D}\mathbf{P}^T \quad (2)$$

with the constraints $\mathbf{T}^T\mathbf{T} = \mathbf{I}$ and $\|\mathbf{p}^{(h)}\| = 1$ for $h = (1, \dots, H)$,

93 with $\mathbf{D} = \mathbf{D}_T\mathbf{D}_P$. If $\mathbf{X} = [\mathbf{X}_1 | \dots | \mathbf{X}_K]$, it follows $\mathbf{X} = \tilde{\mathbf{T}}[\tilde{\mathbf{P}}_1^T | \dots | \tilde{\mathbf{P}}_K^T]$ with $\tilde{\mathbf{P}}^T = [\tilde{\mathbf{P}}_1^T | \dots | \tilde{\mathbf{P}}_K^T]$, the
 94 non-standardized loading matrix inheriting data structure of \mathbf{X} . It derives $\mathbf{X}_k = \tilde{\mathbf{T}}\tilde{\mathbf{P}}_k^T$, blocks
 95 being decomposed into the same orthogonal basis $\tilde{\mathbf{T}}$. Standardizations are also applied to
 96 $\mathbf{X}_k = \mathbf{T}(\mathbf{D}_T\mathbf{D}_{P_k})\mathbf{P}_k^T$, the $(K + 1)$ scaling matrices \mathbf{D}_T and $(\mathbf{D}_{P_1}, \dots, \mathbf{D}_{P_K})$ being all diagonal of size
 97 $(H \times H)$. The canonical factorization of each block \mathbf{X}_k is given by:

$$\mathbf{X}_k = \mathbf{T}(\mathbf{D}_T\mathbf{D}_{P_k})\mathbf{P}_k^T = \mathbf{T}\mathbf{D}_k\mathbf{P}_k^T \quad (3)$$

with the constraints $\mathbf{T}^T\mathbf{T} = \mathbf{I}$ and $\|\mathbf{p}_k^{(h)}\| = 1$ for $h = (1, \dots, H)$, $k = (1, \dots, K)$,

98 with $\mathbf{D}_k = \mathbf{D}_T\mathbf{D}_{P_k}$. To demonstrate relationships between \mathbf{D} and \mathbf{D}_k :

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{D}_T[\tilde{\mathbf{P}}_1^T | \dots | \tilde{\mathbf{P}}_K^T] \\ \mathbf{X} &= \mathbf{T}\mathbf{D}[(\mathbf{D}_T\mathbf{D}^{-1}\mathbf{D}_1)\mathbf{P}_1^T | \dots | (\mathbf{D}_T\mathbf{D}^{-1}\mathbf{D}_K)\mathbf{P}_K^T] \end{aligned}$$

99 Because $\|\mathbf{p}^{(h)}\| = 1$, it follows for a given dimension h , $\|\mathbf{p}\|^2 = \mathbf{p}^T\mathbf{p} = \sum_{k=1}^K \left(\frac{\mathbf{d}_T\mathbf{d}_k}{\mathbf{d}}\right)^2 \mathbf{p}_k^T\mathbf{p}_k$. Because
 100 $\|\mathbf{p}_k^{(h)}\| = 1$, it follows $\sum_{k=1}^K (\mathbf{D}_T\mathbf{D}^{-1}\mathbf{D}_k)^2 = \mathbf{I}$, then $\mathbf{D}^2 = \sum_{k=1}^K (\mathbf{D}_T\mathbf{D}_k)^2$. For normalized common
 101 components ($\|\mathbf{t}^{(h)}\| = 1$ for $h = 1, \dots, H$), $\mathbf{D}^2 = \sum_{k=1}^K \mathbf{D}_k^2$.

102 **Interpretation** For a given method, canonical factorization allows us to project the \mathbf{X}_k columns
 103 into a common space spanned by the (orthogonal and normalized) common components \mathbf{T} ,
 104 considered as a common model for blocks. Common and block parameters are related with
 105 each others. Because canonical factorization of data by multiblock methods have the same
 106 normalized format (Eq. 1), methods can be compared on the same parameters (i.e., \mathbf{T} , \mathbf{D}_k). The
 107 property $\mathbf{D}^2 = \sum_k \mathbf{D}_k^2$ means that the variance of concatenated data is the sum of block-variances.

108 2.3 Canonical factorization of data by multiblock methods

109 The canonical factorization of concatenated multiblock data by PCA is introduced as a reference.
 110 Then, multiblock methods are presented by: (i) their original algorithm or criterion, (ii) a
 111 criterion related to PCA of concatenated data, and (iii) their canonical factorization.

112 2.3.1 PCA of concatenated data

113 The PCA of $\mathbf{X} = [\mathbf{X}_1 | \dots | \mathbf{X}_K]$ — with standardized components — is based, for the first dimen-
 114 sion, on criterion [19]:

$$\arg \max_{\mathbf{t}, \mathbf{w}_1, \dots, \mathbf{w}_K} \sum_{k=1}^K \text{cov}^2(\mathbf{t}, \mathbf{X}_k \mathbf{w}_k) \quad \text{with the constraints} \quad \|\mathbf{t}\| = \|\mathbf{w}_k\| = 1. \quad (4)$$

115 The solution is given by $\mathbf{w}_k = \frac{\mathbf{X}_k^T \mathbf{t}}{\|\mathbf{X}_k^T \mathbf{t}\|}$. While replacing \mathbf{w}_k in Eq. (4), \mathbf{T} are either (normalized)
 116 eigenvectors of $(\mathbf{X}\mathbf{X}^T)$ or can be sought by the deflation of \mathbf{X}_k into successive components \mathbf{t} .
 117 Equivalently, \mathbf{t} is the first left singular vectors of \mathbf{X} and $\mathbf{w} = [\mathbf{w}_1 | \dots | \mathbf{w}_K]$ — with $\mathbf{t} = \mathbf{X}\mathbf{w}$ — is
 118 deduced from the right singular vectors of \mathbf{X} up to a block-scaling. Therefore, the decomposition
 119 of \mathbf{X}_k onto the orthogonal basis is given by $\mathbf{X}_k = \mathbf{X}_k^{(1)}$ and $\mathbf{X}_k^{(h+1)} = [\mathbf{I} - \mathbf{t}^{(h)} \mathbf{t}^{(h)T}] \mathbf{X}_k^{(h)}$ for each
 120 dimension $h = (1, \dots, H)$. It derives:

$$\begin{aligned} \mathbf{X}_k &= \sum_{h=1}^H \mathbf{t}^{(h)} \mathbf{t}^{(h)T} \mathbf{X}_k \quad \text{for } k = (1, \dots, K) \\ \Leftrightarrow \mathbf{X}_k &= \mathbf{T} \mathbf{T}^T \mathbf{X}_k \\ \Leftrightarrow \mathbf{X}_k &= \mathbf{T} \mathbf{D}_k \mathbf{W}_k^T \end{aligned} \quad (5)$$

121 with \mathbf{D}_k the scaling matrix of \mathbf{W}_k . From Eqs. (2) and (3), it follows:

$$\begin{cases} \mathbf{T} &= [\mathbf{t}^{(1)}, \dots, \mathbf{t}^{(H)}] \\ \mathbf{D}_k &= \text{diag}(\|\mathbf{X}_k^T \mathbf{t}^{(1)}\|, \dots, \|\mathbf{X}_k^T \mathbf{t}^{(H)}\|) & \text{for } k = (1, \dots, K) \\ \mathbf{P}_k^T &= [\mathbf{w}_k^{(1)T}, \dots, \mathbf{w}_k^{(H)T}] & \text{for } k = (1, \dots, K) \\ \mathbf{D} &= \text{diag}\left(\sqrt{\sum_k \|\mathbf{X}_k^T \mathbf{t}^{(1)}\|^2}, \dots, \sqrt{\sum_k \|\mathbf{X}_k^T \mathbf{t}^{(H)}\|^2}\right) \\ \mathbf{P}^T &= [\mathbf{D}_1 \mathbf{D}^{-1} \mathbf{P}_1^T | \dots | \mathbf{D}_K \mathbf{D}^{-1} \mathbf{P}_K^T]. \end{cases} \quad (6)$$

122 All multiblock methods look for common components \mathbf{T} , orthogonal to each other, on which
 123 blocks $(\mathbf{X}_1, \dots, \mathbf{X}_K)$ are decomposed. Diagonal elements of \mathbf{D}_k , $\|\mathbf{X}_k^T \mathbf{t}\|$ — scaling parameters of
 124 block-loadings \mathbf{W}_k — are interpreted as block-variance explained by common components. It
 125 helps us to understand how multiblock methods partition block-variability along dimensions
 126 of common space. The diagonal elements of \mathbf{D} are interpreted as the total variance explained
 127 by common component.

128 2.3.2 Consensus Principal Component Analysis (CPCA)

129 CPCA has been proposed by [42] in chemometrics by extending NIPALS algorithm to several
 130 blocks. The algorithm proceeds in two steps. First, the parameters — common components,
 131 common loadings, block-components and block-loadings — are computed according to a pro-
 132 cedure repeated until convergence. A standardized (random) vector \mathbf{t} is considered as a starting
 133 point. The columns in \mathbf{X}_k are regressed on \mathbf{t} , leading to block-loadings $\mathbf{w}_k^T = (\mathbf{t}^T \mathbf{t})^{-1} \mathbf{t}^T \mathbf{X}_k$ in turn

134 standardized. The block-components are computed by $\mathbf{t}_k = \mathbf{X}_k \mathbf{w}_k$. The common axis is com-
 135 puted by $\boldsymbol{\omega}^T = (\mathbf{t}^T \mathbf{t})^{-1} \mathbf{t}^T [\mathbf{t}_1 | \dots | \mathbf{t}_K]$, thus composed of K vectors $\boldsymbol{\omega}_k^T = (\mathbf{t}^T \mathbf{t})^{-1} \mathbf{t}^T \mathbf{t}_k$. The common
 136 component is updated by setting $\mathbf{t} = [\mathbf{t}_1 | \dots | \mathbf{t}_K] \boldsymbol{\omega}^T$. After convergence and in a second step,
 137 higher-order parameters are computed by deflating \mathbf{X}_k blocks with respect to common com-
 138 ponents. The procedure has monotonicity properties [16] and optimizes the PCA problem (4).
 139 This leads to canonical factorization of \mathbf{X} and $(\mathbf{X}_1, \dots, \mathbf{X}_K)$ by CPCA according to Eq. (6).

140 2.3.3 Multiple CO-inertia Analysis (MCOA)

In ecological field, MCOA has been proposed [7, 16] as an alternative to MAXVAR [5]. It is
 based on the same problem as CPCA (Eq. 4). For the first-order solution, the parameters \mathbf{w}_k and
 \mathbf{t} are the same as the ones of CPCA. The difference between MCOA and CPCA is the deflation
 step, with respect to block-loadings \mathbf{w}_k (MCOA) and to common component \mathbf{t} (CPCA). The
 common components are orthogonal to each other (MCOA and CPCA). Due to block-deflation,
 the maximum MCOA dimension is $H' = \min [\text{rank}(\mathbf{X}_1, \dots, \mathbf{X}_K)]$. Without any loss of generality
 but keeping in mind that \mathbf{X} and \mathbf{X}_k reconstructions are not complete and contain residual terms,
 canonical factorization of $(K + 1)$ data sets is performed on a H' -dimensional space:

$$\begin{aligned} \mathbf{X}_k &= \sum_{h=1}^{H'} \mathbf{t}^{(h)} \mathbf{t}^{(h)T} \mathbf{X}_k + \mathbf{R}_k \quad \text{for } k = (1, \dots, K) \\ \mathbf{X} &= \sum_{h=1}^{H'} \mathbf{t}^{(h)} \mathbf{t}^{(h)T} \mathbf{X} + \mathbf{R} \end{aligned}$$

141 Therefore, canonical factorization of \mathbf{X} and $(\mathbf{X}_1, \dots, \mathbf{X}_K)$ by MCOA is given by the same param-
 142 eters as CPCA (Eq. 6) but in a H' -dimension space.

143 2.3.4 Multiple Factorial Analysis (MFA)

144 MFA has been proposed by [9] and is wildly used in sensometrics [25]. It is known as a PCA of
 145 concatenated blocks, each of which being scaled with the inverse of $\sqrt{\lambda_k^{(1)}}$, the square root of the
 146 first eigenvalue of block-PCA, namely $\mathbf{X}_\lambda = [(1/\sqrt{\lambda_1^{(1)}})\mathbf{X}_1 | \dots | (1/\sqrt{\lambda_K^{(1)}})\mathbf{X}_K]$. This block-scaling
 147 gives more weight to blocks with the lowest within-block correlation. MFA (with standardized
 148 components) can be presented with following criterion:

$$\arg \max_{\mathbf{t}, \mathbf{w}_1, \dots, \mathbf{w}_K} \sum_{k=1}^K \text{cov}^2 \left(\mathbf{t}, \frac{1}{\sqrt{\lambda_k^{(1)}}} \mathbf{X}_k \mathbf{w}_k \right) \quad \text{with the constraints } \|\mathbf{t}\| = \|\mathbf{w}_k\| = 1. \quad (7)$$

149 The solution is given by $\mathbf{w}_k = \frac{1/\sqrt{\lambda_k^{(1)}} \mathbf{X}_k^T \mathbf{t}}{\|1/\sqrt{\lambda_k^{(1)}} \mathbf{X}_k^T \mathbf{t}\|} = \frac{\mathbf{X}_k^T \mathbf{t}}{\|\mathbf{X}_k^T \mathbf{t}\|}$. Because MFA is a PCA of weighted concate-
 150 nated matrix \mathbf{X}_λ , \mathbf{T} are eigenvectors of $\left(\sum_{k=1}^K \frac{1}{\lambda_k^{(1)}} \mathbf{X}_k \mathbf{X}_k^T \right)$. From Eq. (7) — H being the rank of
 151 $\left(\sum_{k=1}^K \frac{1}{\lambda_k^{(1)}} \mathbf{X}_k \mathbf{X}_k^T \right)$ — canonical factorization of \mathbf{X} and $(\mathbf{X}_1, \dots, \mathbf{X}_K)$ by MFA is given by Eq. (6).

152 2.3.5 Structuration de Tableaux A Trois Indices de la Statistique (STATIS)

153 STATIS has been proposed by [20] and is also widely used in sensometrics where a compromise
 154 (between judges) is mainly sought. The method assumes that information behind K blocks is
 155 contained in K matrices of scalar products $(\mathbf{X}_1\mathbf{X}_1^T, \dots, \mathbf{X}_K\mathbf{X}_K^T)$. The method is based on a two-step
 156 procedure. The first (inter-structure) one searches for a compromise matrix denoted \mathbf{S} , solution
 157 of:

$$\arg \min_{\mathbf{S}, \alpha_1, \dots, \alpha_K} \sum_{k=1}^K \|\mathbf{X}_k\mathbf{X}_k^T - \alpha_k\mathbf{S}\|^2 \quad \text{with the constraint} \quad \sum_{k=1}^K \alpha_k^2 = 1. \quad (8)$$

158 The solution is given by $\mathbf{S} = \sum_{k=1}^K \alpha_k \mathbf{X}_k \mathbf{X}_k^T$ with $\alpha = (\alpha_1, \dots, \alpha_K)$ the first eigenvector of
 159 $\mathbf{C} = [\text{trace}(\mathbf{X}_k \mathbf{X}_k^T \mathbf{X}_{k'} \mathbf{X}_{k'}^T) / \sqrt{\text{trace}(\mathbf{X}_k \mathbf{X}_k^T) \text{trace}(\mathbf{X}_{k'} \mathbf{X}_{k'}^T)}]$ of dimension $(K \times K)$ [2, 3, 30]. The second
 160 (compromise) step consists in looking for common components \mathbf{T} from the eigen decomposition
 161 of \mathbf{S} . Because $\mathbf{S} = \sum_{k=1}^K \alpha_k \mathbf{X}_k \mathbf{X}_k^T$ — with $\alpha_k \geq 0$ [26] — it follows $\mathbf{S} = \mathbf{X}_\alpha \mathbf{X}_\alpha^T$ with concatenated
 162 block-weighted matrix $\mathbf{X}_\alpha = [\sqrt{\alpha_1} \mathbf{X}_1 | \dots | \sqrt{\alpha_K} \mathbf{X}_K]$. Then, common (normalized) components \mathbf{T}
 163 can be found by PCA of \mathbf{X}_α :

$$\arg \max_{\mathbf{t}, \mathbf{w}_1, \dots, \mathbf{w}_K} \sum_{k=1}^K \text{cov}^2(\mathbf{t}, \sqrt{\alpha_k} \mathbf{X}_k \mathbf{w}_k) \quad \text{with the constraints} \quad \|\mathbf{t}\| = \|\mathbf{w}_k\| = 1. \quad (9)$$

164 The solution is given by $\mathbf{w}_k = \frac{\sqrt{\alpha_k} \mathbf{X}_k^T \mathbf{t}}{\|\sqrt{\alpha_k} \mathbf{X}_k^T \mathbf{t}\|} = \frac{\mathbf{X}_k^T \mathbf{t}}{\|\mathbf{X}_k^T \mathbf{t}\|}$ and \mathbf{T} the eigenvectors of $(\sum_{k=1}^K \alpha_k \mathbf{X}_k \mathbf{X}_k^T)$. From Eq.
 165 (9) — H being the rank of $(\sum_{k=1}^K \alpha_k \mathbf{X}_k \mathbf{X}_k^T)$ assuming non-trivial cases where $\alpha_k \neq 0$ — canonical
 166 factorization of \mathbf{X} and $(\mathbf{X}_1, \dots, \mathbf{X}_K)$ by STATIS is given by Eq. (6).

167 **Block-weight proportionality property** From Eq. (8) and eigen decomposition of \mathbf{S} , following
 168 equality holds:

$$(H-1) \sum_{k=1}^K \|\mathbf{X}_k \mathbf{X}_k^T\|^2 + \sum_{k=1}^K \|\mathbf{X}_k \mathbf{X}_k^T - \alpha_k \mathbf{S}\|^2 = \sum_{k=1}^K \sum_{h=1}^H \|\mathbf{X}_k \mathbf{X}_k^T - \alpha_k \lambda^{(h)} \mathbf{t}^{(h)} \mathbf{t}^{(h)T}\|^2.$$

169 For a given dimension h and a given block \mathbf{X}_k , its block-weight is equal to $\alpha_k \lambda^{(h)}$, with $\lambda^{(h)}$ the
 170 h th eigenvalue associated with eigen decomposition of $(\sum_{k=1}^K \alpha_k \mathbf{X}_k \mathbf{X}_k^T)$. This means that for a
 171 given block \mathbf{X}_k , its H block-weights $(\alpha_k \lambda^{(1)}, \dots, \alpha_k \lambda^{(H)})$ are proportional along dimensions. This
 172 property is important to be known by the user because it has strong practical consequences
 173 further illustrated in the simulation study.

174 2.3.6 Common Components and Specific Weights Analysis (CCSWA)

175 CCSWA has been proposed as an alternative to STATIS while relaxing block-weight propor-
 176 tionality property along dimensions, and used in sensometrics for analysis of standard or free

177 profile data [4, 13, 23, 27]. For a given dimension, CCSWA determines a common component \mathbf{t}
 178 and K block-salience α_k , solutions of problem:

$$\arg \min_{\mathbf{t}, \alpha_1, \dots, \alpha_K} \sum_{k=1}^K \|\mathbf{X}_k \mathbf{X}_k^T - \alpha_k \mathbf{t} \mathbf{t}^T\|^2 \quad \text{with the constraint} \quad \|\mathbf{t}\| = 1. \quad (10)$$

179 The solution is solved by an Alternating Least Square algorithm reiterated until convergence
 180 of criterion (10). In the first stage, a standardized (random) salience vector $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$
 181 is considered. The common component \mathbf{t} is computed as the first normalized eigenvector
 182 of $\sum_{k=1}^K \alpha_k \mathbf{X}_k \mathbf{X}_k^T$. The block-salience are updated by $\alpha_k = \|\mathbf{X}_k^T \mathbf{t}\|^2$ for $k = (1, \dots, K)$. After
 183 convergence and in a second stage, higher-order parameters are computed by deflating \mathbf{X}_k
 184 blocks with respect to common components.

185 **Property** CCSWA's salience are different from the weights α_k computed in STATIS — and
 186 not necessarily proportional — from one dimension to another because CCSWA is based, for a
 187 given dimension h , on problem:

$$\arg \min_{\mathbf{t}, \alpha_1, \dots, \alpha_K} \sum_{k=1}^K \|\mathbf{X}_k^{(h)} \mathbf{X}_k^{(h)T} - \alpha_k^{(h)} \mathbf{t}^{(h)} \mathbf{t}^{(h)T}\|^2 \quad \text{with the constraint} \quad \|\mathbf{t}^{(h)}\| = 1, \quad (11)$$

188 the salience being derived from common components by $\alpha_k^{(h)} = \|\mathbf{X}_k^{(h)T} \mathbf{t}^{(h)}\|^2$.

189 Hanafi and Qannari [14] demonstrated that CCSWA optimizes problem:

$$\arg \max_{\mathbf{t}, \mathbf{w}_1, \dots, \mathbf{w}_K} \sum_{k=1}^K \text{cov}^4(\mathbf{t}, \mathbf{X}_k \mathbf{w}_k) \quad \text{with the constraints} \quad \|\mathbf{t}\| = \|\mathbf{w}_k\| = 1. \quad (12)$$

190 The solution is given by $\mathbf{w}_k = \frac{\mathbf{X}_k^T \mathbf{t}}{\|\mathbf{X}_k^T \mathbf{t}\|}$ and \mathbf{T} the eigenvectors of $\sum_{k=1}^K \alpha_k \mathbf{X}_k \mathbf{X}_k^T$. The salience are *a*
 191 *posteriori* derived from $\alpha_k = \|\mathbf{X}_k^T \mathbf{t}\|^2$. From Eq. (12), canonical factorization of \mathbf{X} and $(\mathbf{X}_1, \dots, \mathbf{X}_K)$
 192 by CCSWA is presented by Eq. (6). Details and proofs are given in [14], especially Appendix 4.

193 2.3.7 Summary

194 All of the methods examined can be decomposed in the same way using Eq. (1). However, they
 195 summarize information provided by blocks in different ways. Their parameters \mathbf{T} , \mathbf{P}_k , \mathbf{D}_k , \mathbf{P} and
 196 \mathbf{D} , differ from one method to another, due to space deformation and specific rotations of \mathbf{T} in
 197 \mathbb{R}^N space. From Eqs. (4), (7), (9), (12), these differences result from: (i) specific block-weightings
 198 (for MFA and STATIS), (ii) covariance power in criterion (power 4 for CCSWA and 2 otherwise)
 199 and (iii) deflation procedure (on \mathbf{w}_k for MCOA and on \mathbf{t} otherwise). All these exploratory
 200 multiblock methods can be viewed as special cases of rGCCA with specific options and while
 201 considering concatenated dataset \mathbf{X} to be explained (i.e., 'super-block') [11, 35]. Some details are
 202 given in Appendix A. The theoretical specificities have practical consequences that lead users to
 203 clarify two issues before choosing multiblock method. These issues derive from block structure
 204 which must be explored beforehand, e.g., with K block-PCA. (i) The first issue is related to

205 variable-structure and concerns block-variance distribution along dimensions. If block(s) with
 206 a scattered variance along dimensions are more important than others, then MFA — or another
 207 method with a MFA-like block-scaling — should be chosen. Otherwise, a standard block-scaling
 208 (e.g., with inertia) should be applied. (ii) The second issue is related to observation-structure.
 209 If the aim is to emphasize only common structure, STATIS — and its block-scaling identical
 210 along dimensions (Eq. 9) — should be chosen. If the aim is to consider what is specific to each
 211 block along dimensions, CCSWA — and its specific block-scaling (Eq. 11) — should be chosen.
 212 Otherwise, if the aim is to mix common and specific information along dimensions, CPCA
 213 or MCOA (Eq. 4) should be chosen, the latter method giving a more condensed information
 214 (section 2.3.3).

215 2.4 Comparison of multiblock exploratory methods

216 **How to compare multiblock methods?** In case of multiblock data, the common structure
 217 to all blocks is observation-structure, variables being different from one block to another. The
 218 components \mathbf{T} represent this common structure. In a linear approach, there is a duality between
 219 observation- and variable-structure. This can be shown with the following equivalent canonical
 220 factorizations of \mathbf{X}_k , while taking into account that components and loadings are normalized
 221 and respectively orthogonal to each other ($\mathbf{T}^T \mathbf{T} = \mathbf{P}_k^T \mathbf{P}_k = \mathbf{I}$):

$$\begin{aligned} \mathbf{X}_k &= \mathbf{T} (\mathbf{P}_k^T \mathbf{P}_k) \mathbf{D}_k \mathbf{P}_k^T = \mathbf{T} \mathbf{P}_k^T \mathbf{R}_k^2 & \text{for } k = (1, \dots, K) & \quad (13) \\ \mathbf{X}_k &= \mathbf{T} \mathbf{D}_k (\mathbf{T}^T \mathbf{T}) \mathbf{P}_k^T = \mathbf{S}_k \mathbf{T} \mathbf{P}_k^T & \text{for } k = (1, \dots, K), & \end{aligned}$$

222 with $\mathbf{R}_k^2 = \mathbf{P}_k \mathbf{D}_k \mathbf{P}_k^T$ the square correlation matrix between (standardized) block-variables and
 223 $\mathbf{S}_k = \mathbf{T} \mathbf{D}_k \mathbf{T}^T$ the block scalar product matrix of observations into common space. The correla-
 224 tions between block-variables (\mathbf{R}_k) — mainly interpreted by users — are carried by observa-
 225 tion-structure (\mathbf{S}_k).

226 **Simulation model** Few studies have been devoted to multiblock data simulation and canoni-
 227 cal factorization highlights a model — common to several widely-used methods — that makes
 228 it possible. A simulation procedure of controlled observation-structure — associated with given
 229 components sought by canonical factorization — is proposed to compare multiblock methods.
 230 Each block is simulated separately, with a given observation-structure chosen identical or dif-
 231 ferent from other blocks (e.g., separation into more or less separated clusters, noise). For each
 232 block, components \mathbf{T}_k are simulated with a given observation-structure and a given eigenvalues
 233 \mathbf{D}_k . From \mathbf{D}_k , a correlation-matrix \mathbf{R}_k is derived [39]. From Eq. (13), \mathbf{P}_k loadings are derived
 234 from eigenvectors of $\mathbf{R}_k^2 = \mathbf{P}_k \mathbf{D}_k \mathbf{P}_k^T$. Finally, data are computed with $\mathbf{X}_k = \mathbf{T}_k \mathbf{D}_k \mathbf{P}_k^T$. The K blocks
 235 are simulated in the same way — each with a given controlled structure — then $\mathbf{X} = [\mathbf{X}_1 | \dots | \mathbf{X}_K]$
 236 is obtained.

237 **Comparison criteria** Multiblock methods were compared with three criteria. (i) Since com-
 238 mon components \mathbf{T} were comparable but might differ from one method (M_1) to another

239 (M_2), they were compared pairwise by the absolute value of the correlation coefficient $\rho^{(h)} =$
 240 $|\text{cor}(\mathbf{t}_{M_1}^{(h)}, \mathbf{t}_{M_2}^{(h)})|$ for a given dimension h . The closer the $\rho^{(h)}$ value was to 1 (or 0), the more similar
 241 (or different) the components were. (ii) Although common components may differ across
 242 methods, block-variance explained by them was an appropriate criterion to measure how components
 243 partition variance of blocks along dimensions. The second criterion came from diagonal
 244 elements of $\mathbf{D}_k^{(h)2} = \|\mathbf{X}_k^T \mathbf{t}^{(h)}\|^2$. This highlights the direction of block observation-structure on
 245 common structure \mathbf{T} . (iii) The third criterion measures whether simulated observation-structure
 246 was recalled by components \mathbf{T} . For this purpose, K adjusted-Rand index (Rand 1971) was com-
 247 puted, for each block k and dimension h ($\text{ARI}_k^{(h)}$), between expected (=simulated) and observed
 248 observation-structures. The observed structure comes from a K -means clustering applied to
 249 each component $\mathbf{t}^{(h)}$. Among the K computed ($\text{ARI}_1^{(h)}, \dots, \text{ARI}_K^{(h)}$), the maximum value was
 250 kept ($\text{ARI}^{(h)}$). The closer the ARI value was to 1 (or 0), the more the method recovered (or did
 251 not recover) expected observation-structure.

252 3 Application

253 The analyses were performed using R [28]. CPCA, MFA and STATIS come from the SVD of
 254 appropriately scaled (eventually block-scaled) matrix. MCOA came from the `ade4` package with
 255 `'mcoa'` function [8] and CCSWA from the `RGCCA` package with `rgcca` function and `'hpca'` options
 256 [11]. Simulated data were obtained from `monte` function (`fungible` package; [40]). The ARI
 257 criterion was computed with `adjustedRandIndex` function of `mclust` package [31] associated
 258 with `kmeans` function (`stats` package). The choice of optimal block-clustering was obtained
 259 from the most frequent results among 30 indices that evaluate clustering performance between
 260 2 and 10 clusters with `NbClust` package [6]. The PCA plots were obtained with the `FactoMineR`
 261 [21] and `factoextra` packages [18].

262 3.1 Simulation study

263 The goal was to explain similarities and differences between multiblock methods according to
 264 three key questions for users. How do methods behave in case of noise-blocks (scenario **S1**)? In
 265 case of blocks with common and specific observation-structures (scenario **S2**)? In case of blocks
 266 with each specific observation-structures (scenario **S3**)? CPCA, MCOA and CCSWA were used
 267 without any block-scaling, whereas MFA and STATIS were specifically scaled. To measure
 268 ability of multiblock methods to explore data structure, components were simulated from
 269 random data and labelled 'RANDOM'. The simulated data consisted of $N = 90$ observations
 270 measured on $J = 35$ (standardized) variables organized in $K = 4$ blocks with $J_k = (10, 10, 10, 5)$.
 271 The **S1** data were composed of a block with a specific observation-structure (\mathbf{X}_1 with three
 272 well-separated clusters of 30 observations each, 80% of this structure being spanned on the first
 273 dimension and 20% on the second one) and three blocks of noise. The **S2** data consisted of two
 274 blocks with the same observation-structure (\mathbf{X}_1 and \mathbf{X}_2) and two blocks with each a different
 275 structure. The **S3** data consisted of four blocks each with a different observation-structure.
 276 For the latter two scenarios, the four structures consisted of three well-separated clusters each
 277 of 30 observations, with 80% (or 75%, 70%, 65%) of this structure being spanned on the first

278 dimension and 20% (or 25%, 30%, 35%) on the second one. These values were chosen to fix
279 the order of blocks on components. The result of an additional scenario (how do methods
280 behave when all blocks share the same observation-structure) is shown in Sup. Mat. B. For
281 each scenario, 50 data were simulated and mean values of criteria were given.

282 3.1.1 S1: How do multiblock methods behave in case of noise-blocks?

283 The block-PCAs in Fig. 1(A) illustrate the structure of \mathbf{X}_1 in three clusters and noise structures of
284 other blocks. Fig. 1(B) shows that multiblock methods, especially CPCA, MCOA, STATIS and
285 CCSWA, perform better than the RANDOM method. The difference between CPCA and MCOA
286 — due to deflation procedure — is shown in Fig. 1(C) (mean $\rho=1$ for the first dim., 0.24 for the
287 fourth dim.). The \mathbf{X}_1 variance is recovered on the first dimension, the other block-variances
288 are shifted to higher-order dimensions (Fig. 1(B)). CPCA, MCOA, STATIS and CCSWA find
289 the same first component (mean $\rho=0.99-1$; Fig. 1(C)) and recover informative structure of \mathbf{X}_1 on
290 this dimension (meanARI=0.83-0.88; Fig. 1(D)). MFA does not recover the \mathbf{X}_1 structure on this
291 dimension (meanARI=0.12). This method is disadvantaged for this noise-scenario by its block-
292 scaling which gives less importance to block(s) containing information on the first dimension
293 (here, \mathbf{X}_1 with $\lambda_1^{(1)} = 79.4\%$; Fig. 1(A)).

294 3.1.2 S2: How do multiblock methods behave in case of blocks with common and specific 295 observation-structures?

296 The block-PCAs in Fig. 2(A) illustrate the common structure of \mathbf{X}_1 and \mathbf{X}_2 in three clusters
297 ($ARI(\mathbf{X}_1, \mathbf{X}_2) = 1$) and the specific structures of the two other blocks (e.g., $ARI(\mathbf{X}_1, \mathbf{X}_3) \approx 0.01$).
298 Fig. 2(B) shows that all methods recover common structure of \mathbf{X}_1 and \mathbf{X}_2 (MeanARI=0.93-
299 0.95; Fig. 2(D)) with the same first component (Mean $\rho=0.99-1$; Fig. 2(C)). However, there are
300 differences for higher-order dimensions, especially for STATIS and CCSWA. According to its
301 block-scaling stable along dimensions, STATIS gives importance to \mathbf{X}_1 and \mathbf{X}_2 on all dimensions
302 (Fig. 2(B)), the method being only interested in what is common to blocks. Conversely and
303 according to its block-scaling which can vary across dimensions, CCSWA highlights what is
304 common (\mathbf{X}_1 and \mathbf{X}_2 on the first dimension, then focuses on other specific block-structures \mathbf{X}_3
305 then \mathbf{X}_4) on higher-order dimensions. Consequently, CCSWA recovers informative structure
306 of blocks along dimensions (meanARI=0.98-0.85-0.71 for the first three dimensions; Fig. 2(D)).
307 The same conclusions were obtained when clusters are slightly separated or when block-cluster
308 sizes differ.

309 3.1.3 S3: How do multiblock methods behave in case of blocks with each specific observation- 310 structures?

311 The block-PCAs in Fig. 3(A) illustrate the specific structures of all blocks. The multiblock
312 methods follow different strategies here. Due to its specific block-scaling, MFA includes the \mathbf{X}_4
313 block from the first dimension. Because of its rank specificity, MCOA concentrates information
314 on the first dimension. The other methods partition the four block-variance along all dimensions
315 with different components, hence different strategies.

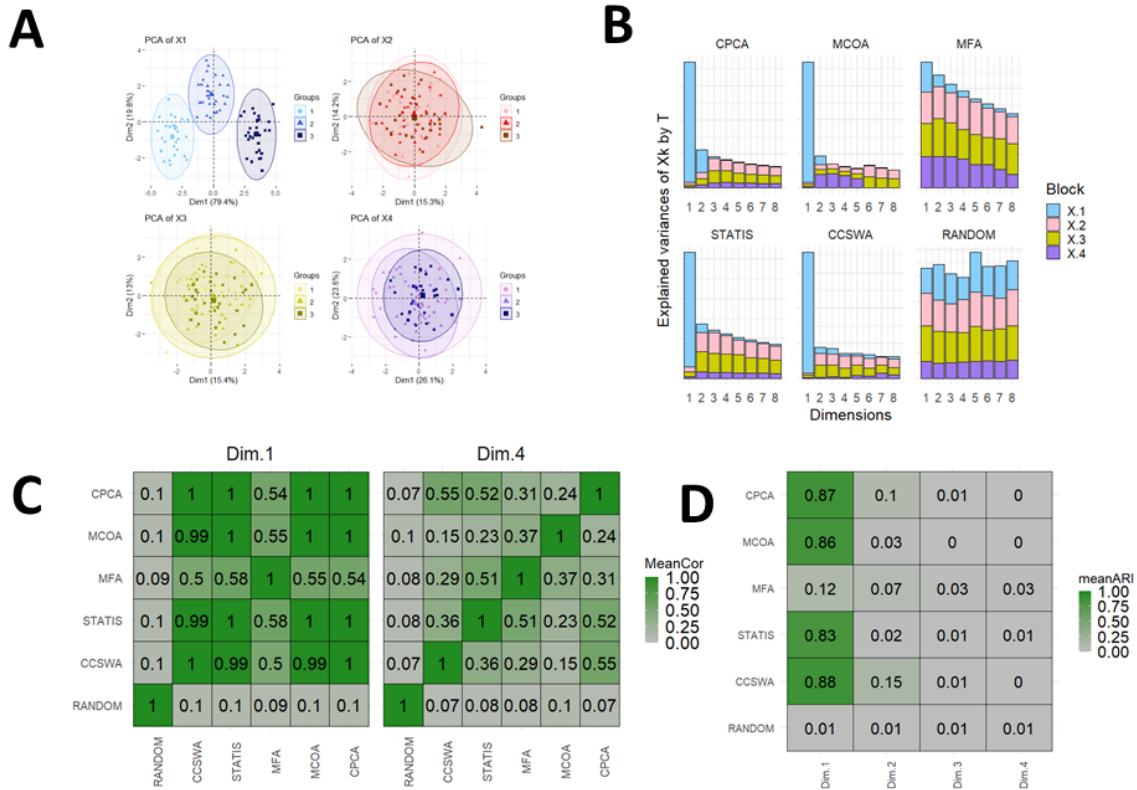


Figure 1: How do multiblock methods behave in case of noise-blocks (S1)? **A**: Simulated blocks (block-PCA), **B**: Block-variances provided by components (D_k^2), **C**: Common-components' similarity ($\rho^{(h)}$), **D**: Ability to recover the structure (ARI).

316 3.2 Exploratory data integration for food data

317 3.2.1 Data, pre-processing and aim

318 The data came from the 'AlimaSSenS' project [1], which aimed to provide a range of foods
 319 adapted to chewing difficulties of elderly, combining pleasure and comfort of eating with nutri-
 320 tional efficiency. Data were collected on 73 subjects, aged between 67 and 87 years old, with poor
 321 or good dental health. The subjects evaluated three meat products (minced chicken, shredded
 322 beef and shredded chicken) twice. During these evaluations, 46 variables were collected and
 323 organized in three blocks. The X_1 block concerned mouth comfort when eating products (28
 324 variables). The X_2 block came from *in vivo* aroma release and perception of foods from nasal
 325 space with PTR-ToF-MS (= Proton Transfer Reaction with Mass Spectrometry; 5 variables). The
 326 X_3 block concerned oral food processing which includes mastication, salivation, bowl forma-
 327 tion, enzyme digestion and swallowing of food bowls (13 variables). The concatenated data set

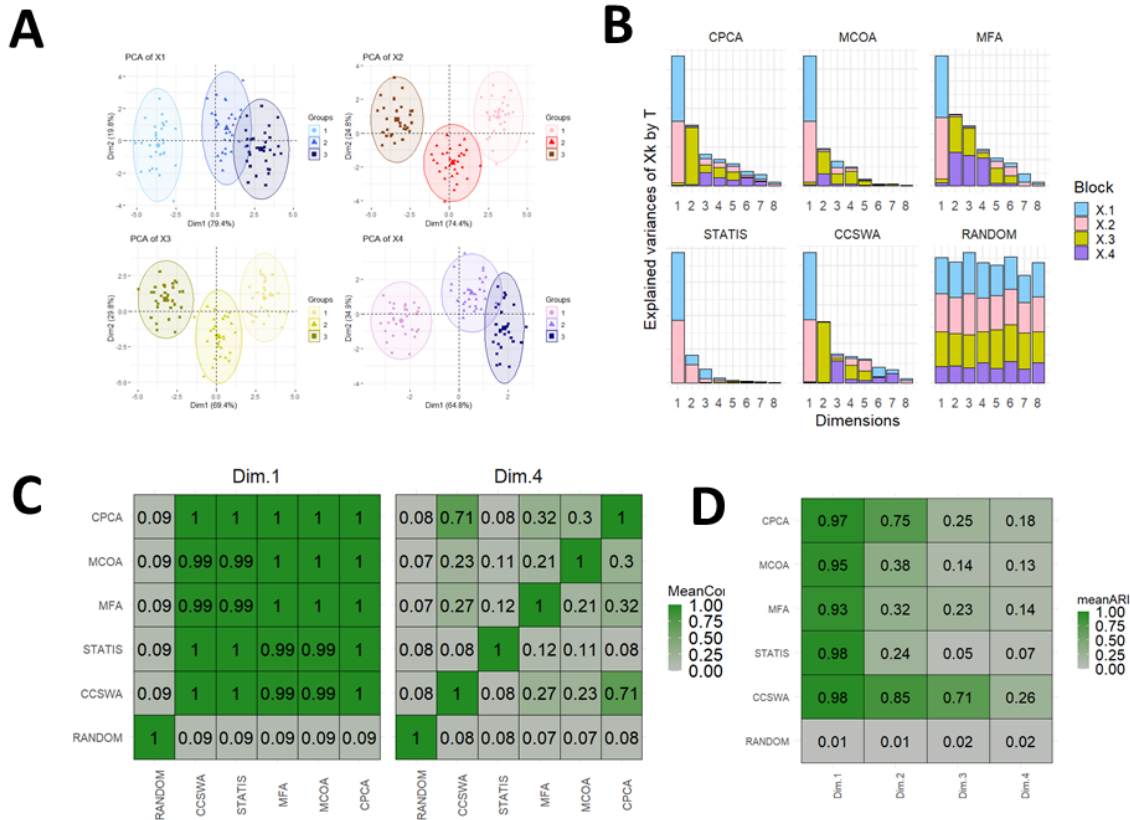


Figure 2: How do multiblock methods behave in case of blocks with both common and specific observation-structures (S_2)? **A**: Simulated blocks (block-PCA), **B**: Block-variances provided by components (D_k^2), **C**: Common-components' similarity ($\rho^{(h)}$), **D**: Ability to recover the structure (ARI).

328 X had 438 rows and 46 columns. Since variables were expressed in different units, they were
 329 standardized. For CPCA, MCOA and CCSWA, to avoid situations where blocks had a strong
 330 influence due to their size (=inertia), block-data were divided by square root of their number
 331 of variables. The aim was to explore relationships between mouth comfort (X_1), aroma release
 332 (X_2) and oral food processing (X_3). For this purpose, exploratory multiblock methods were
 333 used, the challenge being to select the most suitable one.

334 3.2.2 Block description

335 The preliminary step of any exploratory multiblock method was to capture block-structure
 336 across variables (variance distribution along dimensions) and observations (clustering struc-
 337 ture). PCA and K-means were applied to each block. The choice of optimal block-clustering was
 338 obtained from the most frequent results among 30 indices that evaluated clustering performance

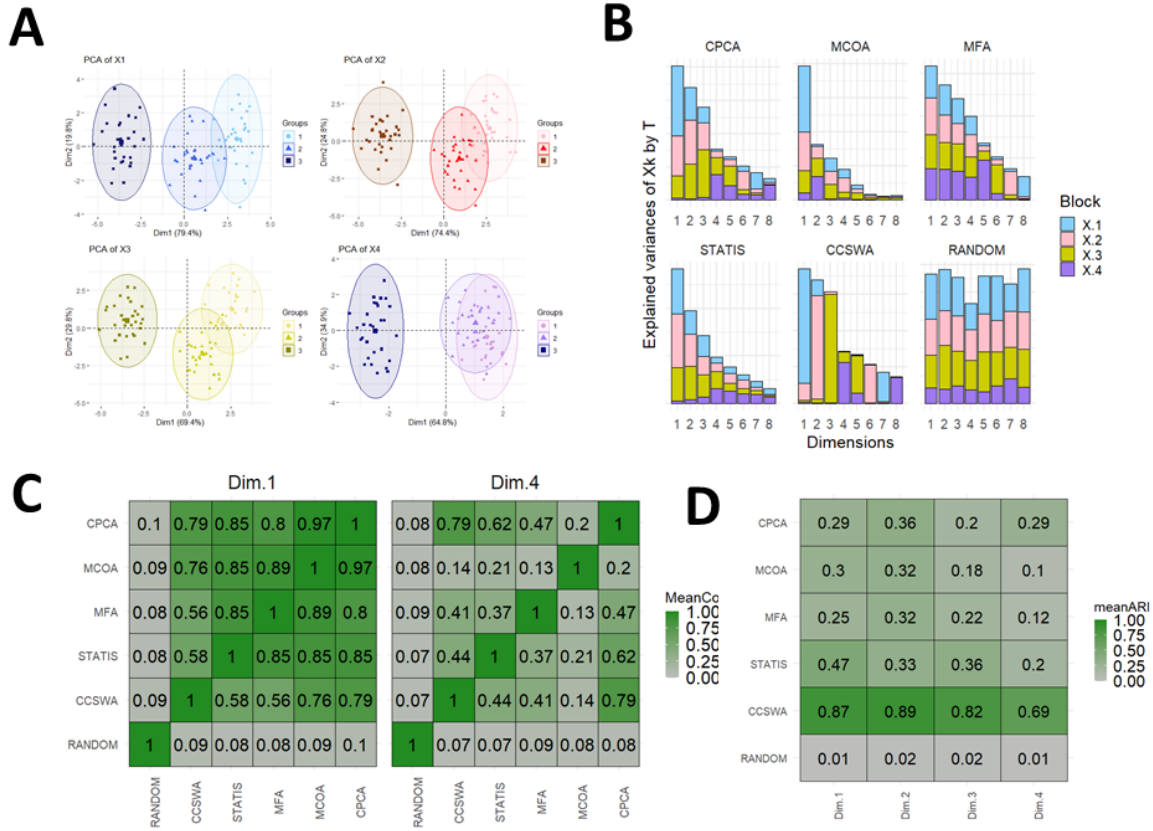


Figure 3: How do multiblock methods behave in case of blocks with all specific observation-structures (**S3**)? **A**: Simulated blocks (block-PCA), **B**: Block-variances provided by components (D_k^2), **C**: Common-components' similarity ($\rho^{(h)}$), **D**: Ability to recover the structure (ARI).

339 between 2 and 10 clusters. Results were presented in Sup. Mat. C. The 28 X_1 -variables (mouth
 340 comfort) were mainly clustered in the first dimension (23.6% of variance). The X_1 -observations
 341 were structured in 2 clusters of size (169, 269), separated from metadata-variables *Dentition* and
 342 *Age* and block-variables *times*, *swallow* and *incisive*. The 5 X_2 -variables (aroma release) were
 343 summarized on the first two dimensions (47.1% and 35.1% of variance). The X_2 -observations
 344 were structured in 3 clusters of size (124, 157, 157), separated from metadata-variable *Product*
 345 and block-variable *imax.Hex*. The 13 X_3 -variables (oral food processing) were summarized on
 346 the first component (33.5% of variance). The X_3 -observations were structured in 2 clusters
 347 of size (117, 321), separated from metadata-variables *Age* and *Dentition* and block-variables *area2*,
 348 *F2*, *area1*, *F1.hardness*, *area.neg.sticky*. These block-clusters were not related with each others:
 349 $ARI(\text{Clust.}X_1, \text{Clust.}X_2)=0.0145$ and $RV(X_1, X_2)=0.06$; $ARI(\text{Clust.}X_1, \text{Clust.}X_3)=0.000$ and $RV(X_1,$
 350 $X_3)=0.05$; $ARI(\text{Clust.}X_2, \text{Clust.}X_3)=0.006$ and $RV(X_2, X_3)=0.04$. Thus, each block had a specific
 351 structure (**S3** scenario).

3.2.3 Multiblock method comparison and strategies

From Fig. 4(B) and according to its block-scaling, MFA gave more importance to mouth comfort block (X_1) than CPCA or MCOA. STATIS tried to highlight a common structure that do not clearly exist. For CPCA, MCOA and CCSWA, the main information came from aroma release (X_2) (Dim. 1 and 2) then food oral processing (X_3) (Dim. 3). Indeed, X_2 observation-structure hold on two dimensions (82.2% of its block-variance; block-PCA) and X_3 on the third one (33.5% of its block-variance; block-PCA). CCSWA made different choices and highlighted aroma release (X_2) on the first dimension, mouth comfort (X_1) on the second dimension and oral food processing (X_3) on the third one. Since our goal was to study observation structure of blocks — with equal importance in analysis — but also relationships between variables, CPCA was chosen.

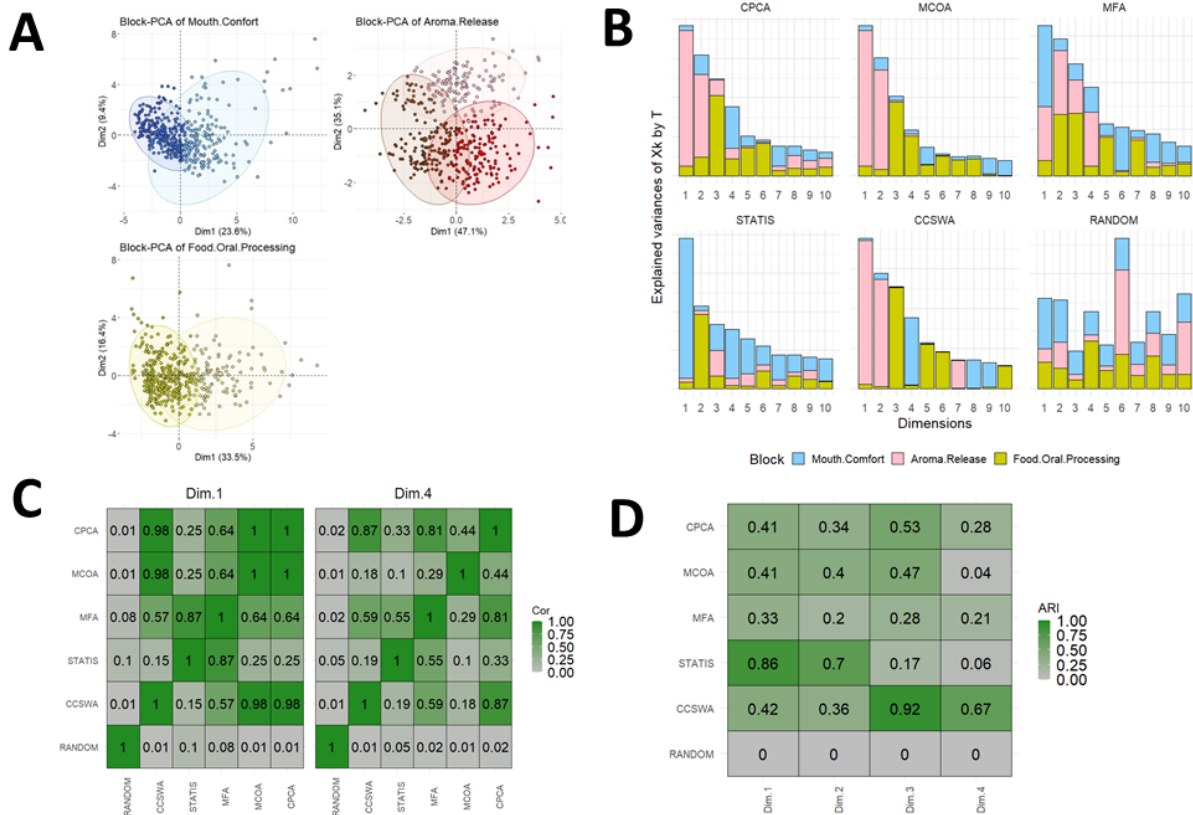


Figure 4: Meat data - Comparison of multiblock methods. **A**: Block-PCA, **B**: Block-variances provided by components (D_k^2), **C**: Common-components' similarity ($\rho^{(h)}$), **D**: Ability to recover the structure (ARI).

363 3.2.4 CPCA interpretation

364 The CPCA results were shown in Sup. Mat. D and E. The observation-structure of aroma release
365 (X_2) was supported by dimensions 1 and 2 (30.3% of variance), then that of oral food processing
366 (X_3) by dimension 3 (10.9% of variance) and the one of mouth comfort (X_1) by dimension 4
367 (7.8% of variance). (i) The 1-2 CPCA components highlighted aroma release (X_2) structure
368 related to products (X_2 -Cluster 1 with *minced chicken*; X_2 -Cluster 2 with *shredded beef*; X_2 -Cluster
369 3 with *shredded chicken*). These X_2 -clusters were explained by five aroma release (X_2) variables
370 *imax.Hex*, *imax.2But*, *imax.AcAld*, *imax.MeThiol*, *imax.MeBut*, and to a lesser extent by five oral
371 food processing (X_3) variables *chewing.time*, *nb.chew.cycle*, *chewing.efficiency*, *Area.2* and *F2*, and
372 15 mouth comfort (X_1) ones, *easy* and *comfort* among others. (ii) The third CPCA component
373 highlighted oral food processing (X_3) structure related to age and dentition (X_3 -Cluster 1 with
374 >80 years old and poor dental health; X_3 -Cluster 2 with <70 years old and good dental health).
375 This X_3 -structure was mainly explained by aroma release (X_2) variables (*imax.Hex*, *imax.MeThiol*
376 and *imax.2But*) and oral food processing (X_3) variables such as *F1.hardness*, *F2*, *Area1* and
377 *Area2*.(iii) The fourth CPCA component highlighted mouth comfort (X_1) structure related to
378 age and dentition (X_1 -Cluster 1 with 70-80 years old and poor dental health; X_1 -Cluster 2 with
379 <70 years old and good dental health). This X_1 -structure was explained by variables related
380 to all blocks: all the mouth comfort variables (X_1) in particular *times* and *swallow*, three aroma
381 release (X_2) variables in particular *imax.Hex* and *imax.MeBut*, and four oral food processing (X_3)
382 variables, in particular *nb.chew.cycle* and *chewing.time*.

383 4 Conclusion and perspectives

384 Our aim was to provide a comprehensive and unified framework — based on homogeneous
385 outputs similar to those of PCA — to compare and explain strategies of multiblock exploratory
386 methods. Many methods have been proposed independently and users have been lost in choos-
387 ing the appropriate one. The data decomposition by canonical factorization was introduced
388 and applied to widely-used methods CPCA, MCOA, MFA, STATIS and CCSWA (also known
389 as COMDIM, ACCPS, HPCA). This factorization extracts parameters that highlight strategy
390 adopted by methods. The methods have been compared on simulated (resp. real) data whose
391 structure is controlled (resp. known). Theoretical and practical results show that block-structure
392 must be explored beforehand, e.g., with K block-PCA, in order to answer two questions before
393 choosing method. The first question concerns the number of block-variables and block-variance
394 distribution along dimensions, which affects block-scaling (no block-scaling / block-scaling with
395 inertia / MFA-like block-scaling). The second issue affects observation-structure within and
396 between blocks, which impacts the choice of the method. In short, CPCA or MCOA mix com-
397 mon and specific information, STATIS emphasizes only common structure only while CCSWA
398 focuses on specific information. Methods and comparison tools were available on R (avail-
399 able code in Supplementary Material), Matlab/Octave ('MultiBlock toolbox for Chemometrics'
400 MB4Chem package; <https://forgemia.inra.fr/chemhouse/octave/mb4chem>) or with a Galaxy
401 web application based on the MB4Chem package (<https://vm-chemflow-francegrille.eu>).

402 Although solution proposed by canonical factorization was original and appropriate, further
403 theoretical and empirical work need to be done. For instance, the proposed matrix decompo-

404 sition could be improved by constraints related to B-factorization of Simultaneous Component
405 Analysis [34, 36, 38]. The effect of block-scaling was considered here as a block-weighted PCA,
406 but could be theoretically investigated with a generalized eigenvalue problem [10]. Canonical
407 factorization could also be applied to multiblock supervised methods, insofar as their criterion
408 can be written as a PCA-like one (e.g., multiblock PLS, multiblock PCAIV as special cases
409 of rGCCA). From a practical point of view, more complex block-clusters could be simulated
410 (e.g., different direction and/or shapes across clusters, clusters visible only on higher order
411 dimensions). The common observation-structure between blocks could be investigated with
412 consensus clustering [24]. The proposed approach allows us to integrate other multiblock
413 methods, such as JIVE [22] devoted to exploration of common and specific block-structures.
414 This researches together with the increase in data volume and complexity will help to make use
415 of exploratory multiblock methods more popular.

416 **CRedit authorship contribution statement**

417 SB, MH and BJ drafted the manuscript. SB, CP, MH and JCB programmed the R, Matlab codes
418 and the implementation on Galaxy. SB and CP analysed data. MH developed methodology
419 and supervised researches. All authors read, reviewed and approved the final manuscript.

420 **Declaration of competing interest**

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

423 **Acknowledgments**

424 The work was performed within a multidisciplinary consortium (MIMS) gathering more than
425 60 researchers whose objective is to examine the analysis and exploitation of multi-source data
426 (<https://eng-digitbio.hub.inrae.fr/themes/understand/consortium-mims-2022-2023>). This
427 consortium is a part of the DIGIT-BIO (Digital biology to understand and predict biological sys-
428 tems) INRAE metaprogramme (<https://eng-digitbio.hub.inrae.fr/themes/understand/consortium-mims-2022-2023>). The data come from the 'AlimaSSenS' project funded and
429 supported by French National Research Agency (ANR-14-CE20-0003; <https://www2.dijon.inrae.fr/senior-et-sens/alima1.php>).

432 **References**

- 433 [1] AlimaSSenS project. <https://www2.dijon.inrae.fr/senior-et-sens/alima1.php>, 2024 (accessed March
434 2024).
435 [2] Abdi, H. Valentin, D. The STATIS Method, 2006. https://www.researchgate.net/publication/239547120_The_STATIS_Method.
436

- 437 [3] Abdi, H., Williams, L.J., Valentin, D., Bannani-Dosse, M. STATIS and DISTATIS: Optimum Multitable Principal
438 Component Analysis and Three Way Metric Multidimensional Scaling. Wiley Interdisciplinary Reviews:
439 Computational Statistics 4 (2012) 124-67. <https://doi.org/10.1002/wics.198>.
- 440 [4] Cariou, V. Qannari, E.M. Rutledge, D.N. Vigneau, E. ComDim: From multiblock data analysis to path model-
441 ing. *Food Qual. Prefer.* 67 (2018) 27-34, <https://doi.org/10.1016/j.foodqual.2017.02.012>.
- 442 [5] Carroll, J.D. Generalization of Canonical Correlation Analysis to Three or More Sets of Variables. Proceedings
443 of the 76th Annual Convention of the American Psychological Association, Washington DC, 227-228, 1968.
- 444 [6] Charrad, M. Ghazzali, N. Boiteau, V. Niknafs, A. NbClust: An R Package for Determining the Relevant
445 Number of Clusters in a Data Set. *J. Stat. Softw.* 61 (2014) 1-36. <https://www.jstatsoft.org/v61/i06/>.
- 446 [7] Chessel, D. Hanafi, M. (1996) Analyse de la co-inertie de K nuages de points. *Revue de Statistique Appliquée.*
447 XLVI (2014) 35-60. <http://www.numdam.org/item?id=RSA_1996__44_2_35_0
- 448 [8] Dray, S. Dufour, A.B. Chessel, D. The ade4 Package - II: Two-Table and K-Table Methods. *R News.* 7 (2007)
449 47-52. <https://cran.r-project.org/doc/Rnews/>.
- 450 [9] Escoufier, B. Pagès, J. Multiple factor analysis (AFMULT package). *Comput. Stat. Data Anal.* 18 (1994) 121-140.
451 [https://doi.org/10.1016/0167-9473\(94\)90135-X](https://doi.org/10.1016/0167-9473(94)90135-X).
- 452 [10] Ghojogh, B. Karray, F. Crowley, M. Eigenvalue and Generalized Eigenvalue Problems: Tutorial, 2019 <https://arxiv.org/abs/1903.11240>.
- 453 [11] Girka, F. Camenen, E. Peltier, C. Gloaguen, A. Guillemot, V. Le Brusquet, L. Tenenhaus, A. RGCCA: Regu-
454 larized and Sparse Generalized Canonical Correlation Analysis for Multiblock Data. R package version 3.0.3,
455 2023 <https://CRAN.R-project.org/package=RGCCA>.
- 456 [12] Gower, J.C. Generalized procrustes analysis. *Psychometrika.* 40 (1975) 33-51. [https://doi.org/10.1007/](https://doi.org/10.1007/BF02291478)
457 [BF02291478](https://doi.org/10.1007/BF02291478).
- 458 [13] Hanafi, M. Mazerolles, G. Dufour, E. Qannari, E.M. Common components and specific weight analysis and
459 multiple Co-inertia analysis applied to the coupling of several measurement techniques. *J. Chemom.* 20 (2006)
460 172-183. <https://doi.org/10.1002/cem.988>.
- 461 [14] Hanafi, M. Qannari, E.M. Nouvelles propriétés de l'Analyse en Composantes Communes et Poids Spécifiques.
462 *Journal de la SFdS.* 149 (2008) 75-97. http://www.numdam.org/item/JSFS_2008__149_2_75_0/.
- 463 [15] Hanafi, M. Kohler, A. Qannari, E.M. Shedding new light on Hierarchical Principal Component Analysis. *J.*
464 *Chemom.* 24 (2010) 703-709. <https://doi.org/10.1002/cem.1334>.
- 465 [16] Hanafi, M. Kohler, A. Qannari, E.M. Connections between Multiple Co-inertia Analysis and Consensus Princi-
466 pal Component Analysis. *Chemometr. Intell. Lab.* 106 (2011) 37-40. [https://doi.org/10.1016/j.chemolab.](https://doi.org/10.1016/j.chemolab.2010.05.010)
467 [2010.05.010](https://doi.org/10.1016/j.chemolab.2010.05.010)
- 468 [17] Horst, P. Relations among m sets of variables. *Psychometrika.* 26 (1961) 126-149. [https://doi.org/10.1007/](https://doi.org/10.1007/BF02289710)
469 [BF02289710](https://doi.org/10.1007/BF02289710).
- 470 [18] Kassambara, A. Mundt, F. Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R
471 Package Version 1.0.7, 2020. <https://CRAN.R-project.org/package=factoextra>
- 472 [19] Lafosse, R. Hanafi, M. Concordance d'un tableau avec K tableaux : définition de K + 1-uples synthétiques.
473 *Revue de Statistique Appliquée* 45 (1997) 111-126. <http://eudml.org/doc/106424>.
- 474 [20] Lavit, C. Escoufier, Y. Sabatier, R. Traissac, P. The ACT (Statis method). *Comput. Stat. Data Anal.* 18 (1994)
475 97-119. [https://doi.org/10.1016/0167-9473\(94\)90134-1](https://doi.org/10.1016/0167-9473(94)90134-1).
- 476 [21] Lê, S. Josse, J. Husson, F. FactoMineR: A Package for Multivariate Analysis. *J. Stat. Softw.* 25 (2008) 1-18.
477 [doi:10.18637/jss.v025.i01](https://doi.org/10.18637/jss.v025.i01).
- 478 [22] Lock, E.F. Hoadley, K.A. Marron, J.S. Nobel, A.B. Joint and Individual Variation Explained (JIVE) for Integrated
479 Analysis of Multiple Data Types. *Ann. Appl. Stat.* 7 (2013) 523-42 <https://doi.org/10.1214/12-A0AS597>.
- 480 [23] Mazerolles, G. Hanafi, M. Dufour, E. Qannari, E.M. Bertrand, D. Common Components and specific weights
481 analysis: a chemometric method for dealing with complexity of food products. *Chemometr. Intell. Lab.* 81
482 (2006) 41-49. <https://doi.org/10.1016/j.chemolab.2005.09.004>.
- 483 [24] Niang, N. Ouattara, M. Weighted consensus clustering for multiblock data. *Proceedings of the SFC.* Paris, France.
484 hal-02471611, 2019.
- 485 [25] Pagès, J. Multiple Factor Analysis by Example Using R (1st ed.). Chapman and Hall/CRC. [https://doi.org/](https://doi.org/10.1201/b17700)
486 [10.1201/b17700](https://doi.org/10.1201/b17700), 2014
- 487 [26] Perron, O. Grundlagen für eine Theorie des Jacobischen Kettenbruchalgorithmus [Foundations for a theory
488 of the Jacobian continued fraction algorithm.] *Mathematische Annalen* 64 (1907) 11-76.
- 489

- 490 [27] Qannari, E.M. Courcoux, P. Vigneau, E. Common Components and specific weights analysis performed on
 491 preference data. *Food Qual. Prefer.* 11 (2000) 151-154. [10.1016/S0950-3293\(01\)00026-X](https://doi.org/10.1016/S0950-3293(01)00026-X).
- 492 [28] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Comput-
 493 ing, Vienna, Austria. <https://www.R-project.org/>, 2022.
- 494 [29] Rand, W.M. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* 66 (1971) 846-850.
 495 <https://doi.org/10.2307/2284239>.
- 496 [30] Robert, P. Escoufier, Y. A unifying tool for linear multivariate statistical methods: the RV-coefficient. *JSTOR*
 497 (*Applied Statistics*). 25 (1976) 257-265. <https://doi.org/10.2307/2347233>.
- 498 [31] Scrucca, L. Fraley, C. Murphy, T.B. Raftery, A.E. Model-Based Clustering, Classification and Density Estimation
 499 Using mclust, in R (1st ed.). Chapman and Hall/CRC (2023). <https://doi.org/10.1201/9781003277965>.
- 500 [32] Smilde, A.K. Westerhuis, J.A. de Jong, S. A framework for sequential multiblock component methods. *J.*
 501 *Chemometr.* 17 (2003) 323-337. <https://doi.org/10.1002/cem.811>.
- 502 [33] Tchanda Mangamana, E. Cariou, V. Vigneau, E. Glèlè Kakai, R.L. Qannari, E.M. Unsupervised multiblock
 503 data analysis: A unified approach and extensions. *Chemometr. Intell. Lab. Syst.* 194 (2019) 103856. <http://dx.doi.org/10.1016/j.chemolab.2019.103856>.
- 504 [34] Ten Berge, J.M. Kiers, H.A. Van der Stel, V. Simultaneous components analysis. *Stat. Appl.* 4 (1992) 277-392.
- 505 [35] A. Tenenhaus, M. Tenenhaus. Regularized generalized canonical correlation analysis, *Psychometrika*, 76
 506 (2011), 257-284. <https://doi.org/10.1007/s11336-011-9206-8>.
- 507 [36] Timmerman, M.E. Kiers, H.A.L. Four simultaneous component models for the analysis of multivariate time
 508 series from more than one subject to model intraindividual and interindividual differences. *Psychometrika* 68
 509 (2003) 105-121. <https://doi.org/10.1007/BF02296656>.
- 510 [37] Van de Geer, J.P. Linear relations among k sets of variables. *Psychometrika*. 49 (1984) 79-94. <https://doi.org/10.1007/BF02294207>.
- 511 [38] Van Deun, K. Smilde, A.K. van der Werf, M.J. Kiers, H.A.L. Van Mechelen, I. A structured overview of
 512 simultaneous component based data integration. *BMC Bioinformatics* 10, 246 (2009). <https://doi.org/10.1186/1471-2105-10-246>.
- 513 [39] Waller, N.G. Generating correlation matrices with specified eigenvalues using the method of alternating
 514 projections. *Am. Stat.* 74 (2020) 21-28. <https://doi.org/10.1080/00031305.2017.1401960>.
- 515 [40] Waller, N.G. fungible: Psychometric Functions from the Waller Lab. University of Minnesota, Minneapolis,
 516 Minnesota. R package 2.4.2 (2024) <https://CRAN.R-project.org/package=fungible>.
- 517 [41] Westerhuis, J.A. Kourtí, T. Macgregor, J.F. Analysis of Multiblock and Hierarchical PCA and PLS Models.
 518 *J. Chemometr.* 12 (1998) 301-321. [https://doi.org/10.1002/\(SICI\)1099-128X\(199809/10\)12:5%3C301::AID-CEM515%3E3.0.CO;2-S](https://doi.org/10.1002/(SICI)1099-128X(199809/10)12:5%3C301::AID-CEM515%3E3.0.CO;2-S)
- 519 [42] Wold, S. Hellberg, Y. Lundstedt M. Sjostrom H. Wold Proc. Symp. on PLS Model Building: Theory and
 520 Application, Frankfurt am Main (1987).
- 521 [43] Wold, S. Geladi, P. Esbensen, K. Ohman, J. Multi-way principal components-and PLS-analysis. *J. Chemometr.*
 522 1 (1987) 41-56. <https://doi.org/10.1002/cem.1180010107>.
- 523 [44] Wold, S. Kettaneh, N. Tjessem, K. Hierarchical multi-block PLS and PC models for easier interpretation and
 524 as an alternative to variable selection. *J. Chemometr.* 10 (1996) 463-482. [https://doi.org/10.1002/\(SICI\)1099-128X\(199609\)10:5/6%3C463::AID-CEM445%3E3.0.CO;2-L](https://doi.org/10.1002/(SICI)1099-128X(199609)10:5/6%3C463::AID-CEM445%3E3.0.CO;2-L).

530 **Appendices**

531 **A Multiblock methods as special cases of rGCCA**

Multiblock method	Criterion & constraints	Block-scaling	Package & option
CPCA [42]	$\sum_k \text{cov}^2(\mathbf{X}_k \mathbf{w}_k, \mathbf{X}_{k+1} \mathbf{w}_{k+1})$ with $\ \mathbf{t}\ = \ \mathbf{w}_k\ = 1$	$1/\sqrt{J_k}$	RGCCA, 'cpca-2'
MCOA [7]	$\sum_k \text{cov}^2(\mathbf{X}_k \mathbf{w}_k, \mathbf{X}_{k+1} \mathbf{w}_{k+1})$ with $\ \mathbf{t}\ = \ \mathbf{w}_k\ = 1$	$1/\sqrt{J_k}$	RGCCA, 'mcoa'
MFA [9]	$\sum_k \text{cov}^2(\mathbf{X}_k \mathbf{w}_k, \mathbf{X}_{k+1} \mathbf{w}_{k+1})$ with $\ \mathbf{t}\ = \ \mathbf{w}_k\ = 1$	$1/\sqrt{\lambda_k^{(1)}}$	RGCCA, 'mfa'
STATIS [20]	$\sum_k \text{cov}^2(\mathbf{X}_k \mathbf{w}_k, \mathbf{X}_{k+1} \mathbf{w}_{k+1})$ with $\ \mathbf{t}\ = \ \mathbf{w}_k\ = 1$	$\sqrt{\alpha_k}$	ade4, 'statis'
CCSWA [27]	$\sum_k \text{cov}^4(\mathbf{X}_k \mathbf{w}_k, \mathbf{X}_{k+1} \mathbf{w}_{k+1})$ with $\ \mathbf{t}\ = \ \mathbf{w}_k\ = 1$	$1/\sqrt{J_k}$	RGCCA, 'hpca'

Table 1: Exploratory multiblock methods as special cases of rGCCA with a super-block [11].

532 **B How do methods behave when all blocks share the same structure?**

533 All multiblock methods recover observation-structure on the first two dimensions and in the
534 same way, which is not the case for the RANDOM method.

535 **C Meat data - Block-PCA biplots coloured according to their observation-structure**

536

537 **D Meat data - CPCA biplots coloured according to block observation-structure**

538 **E Meat data - CPCA weights of Meat data**

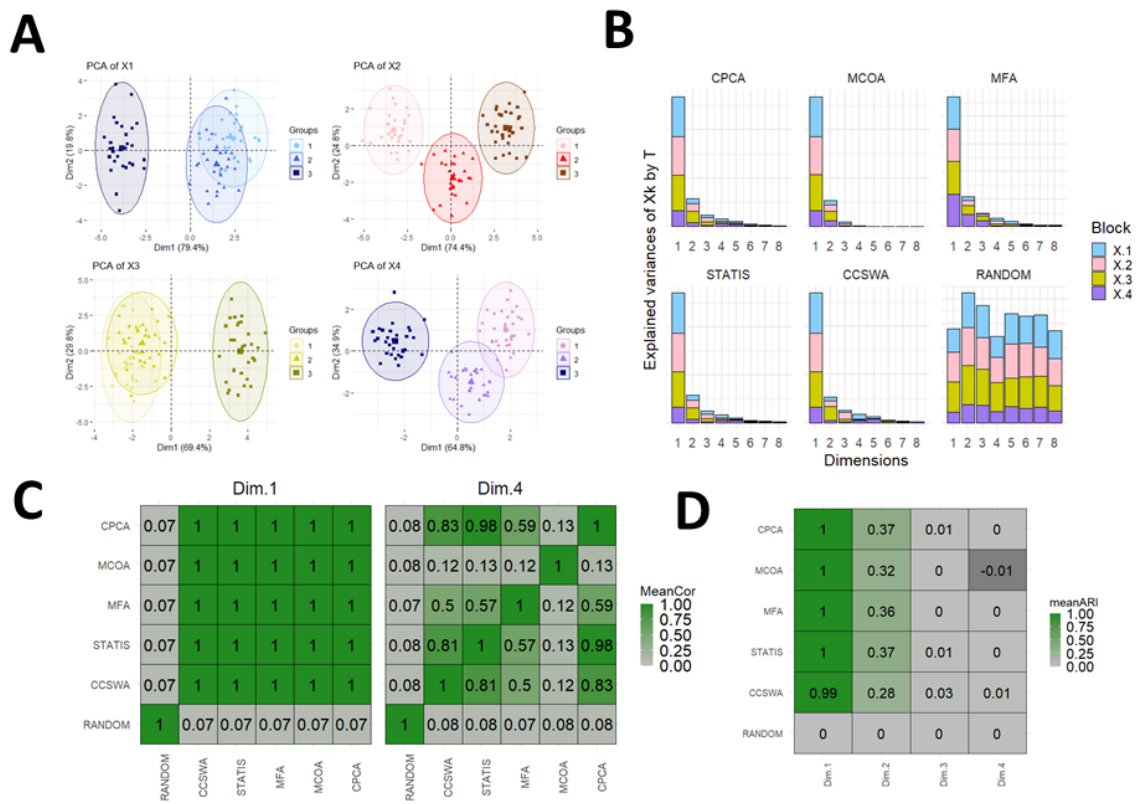


Figure 5: How do methods behave when all blocks have a common structure (S_1)? **A**: Simulated blocks (block-PCA), **B**: Block-variances provided by components (D_k^2), **C**: Common-components' similarity ($\rho^{(h)}$), **D**: Ability to recover the structure (ARI).

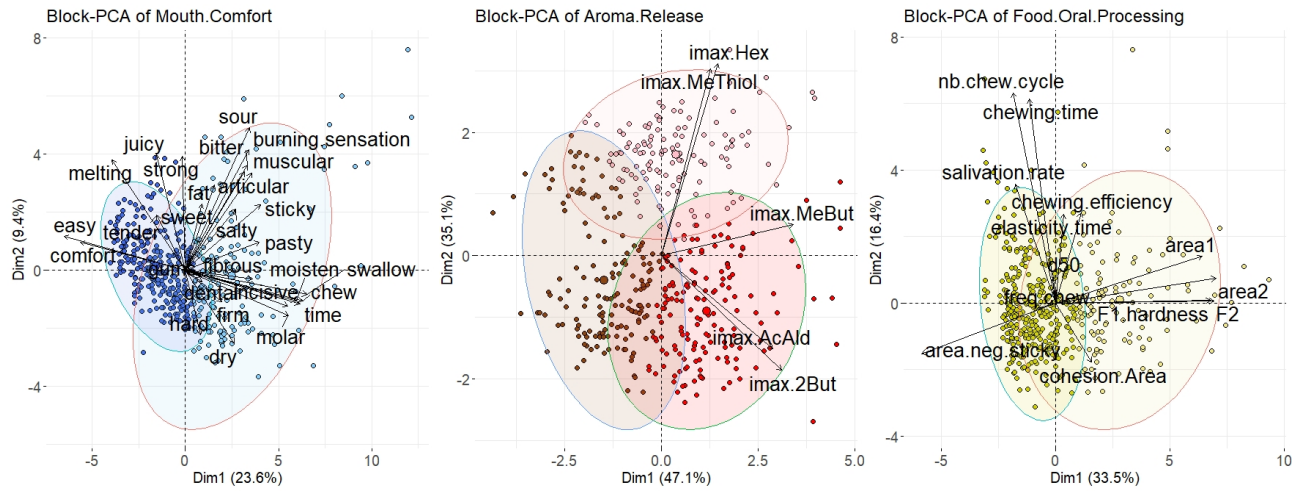


Figure 6: Meat data - Block-PCA biplots coloured according to their observation-structure.

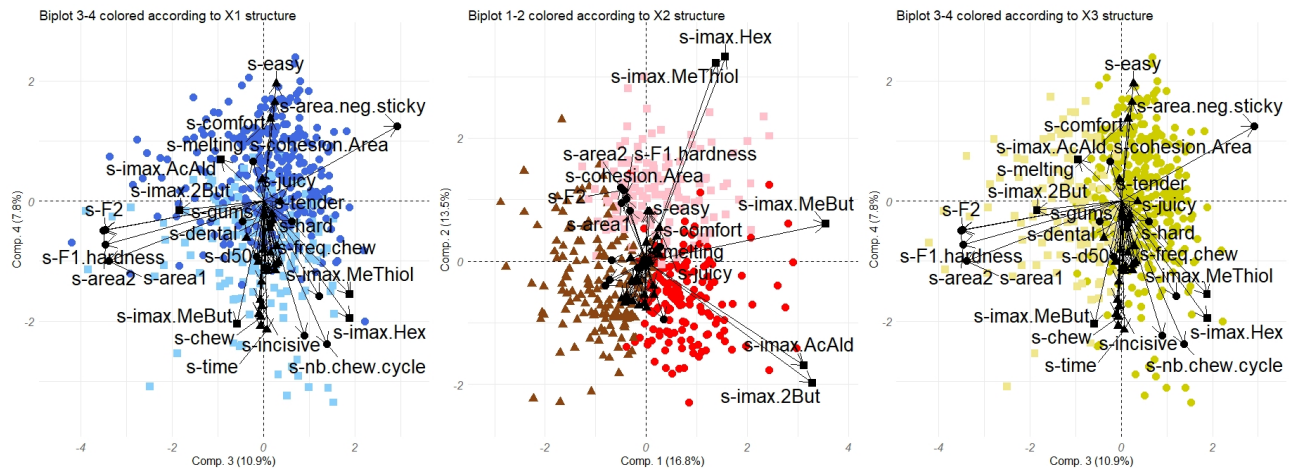


Figure 7: Meat data - CPCA biplots for dimensions (1,2) and (3,4) coloured according to block observation-structure.

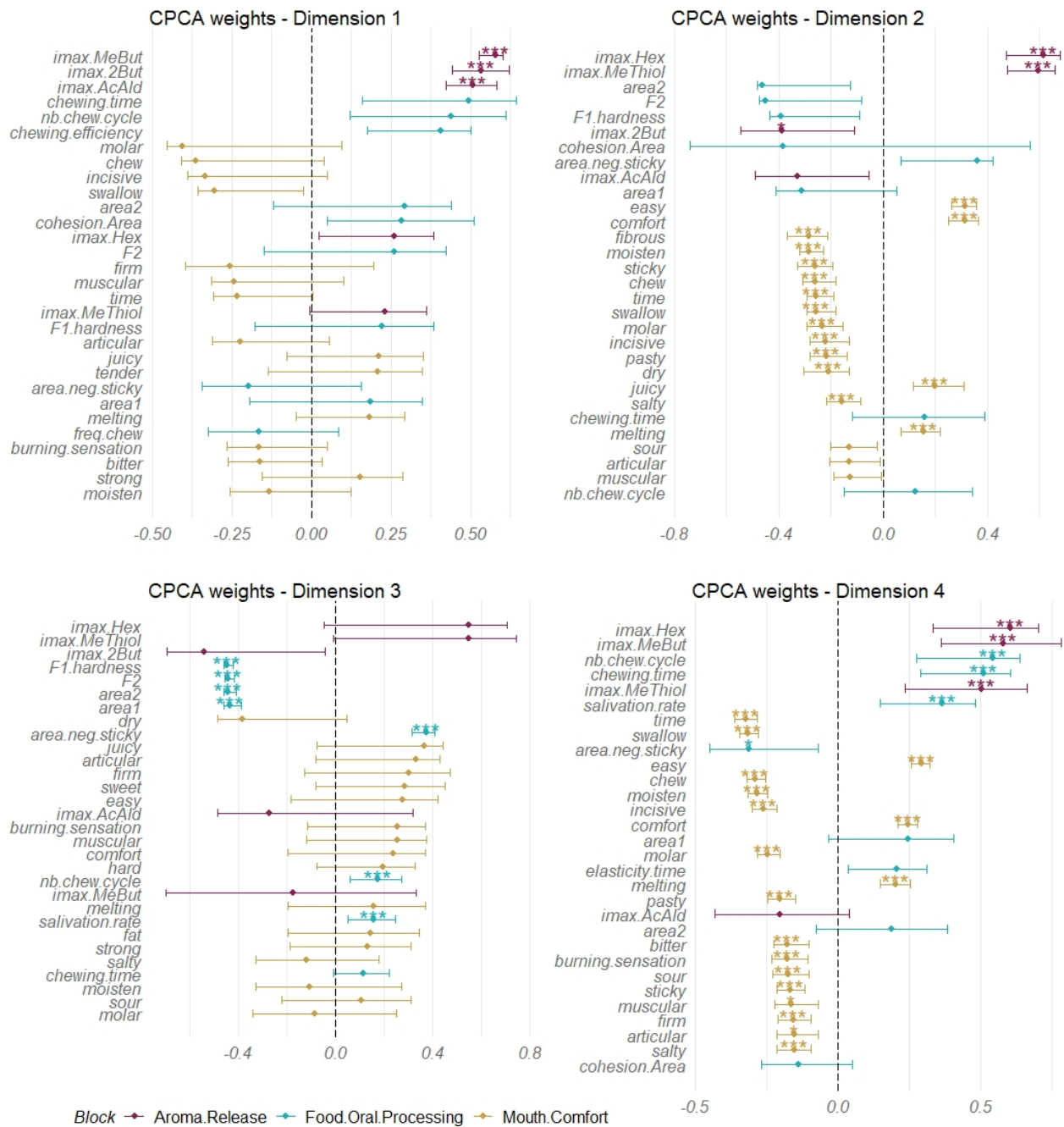


Figure 8: Meat data - CPCA weights with 95% confidence intervals and significance (500 bootstrap simulations) for dimensions 1 to 4.