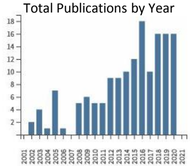
A small walk around multiblock and path modeling approaches in the scope of interactions between health and food

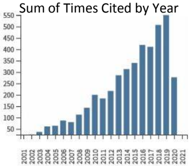
Véronique Cariou

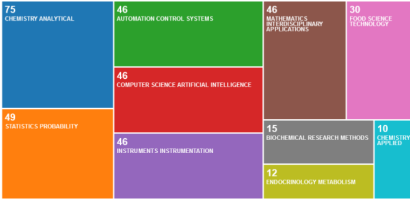
Statistics, Sensometrics, Chemometrics Research Unit
Oniris, INRAE, Nantes

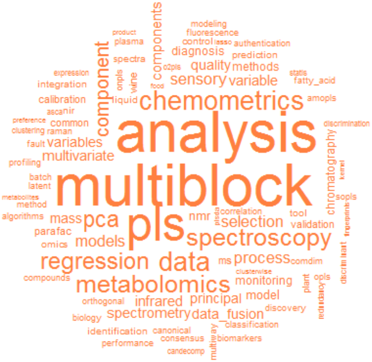
Overview

Multiblock in Web of Science



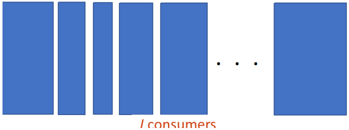


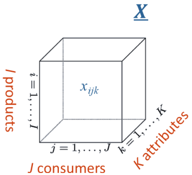


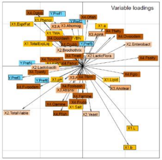


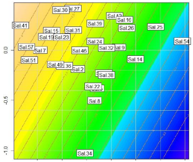
Approaches in sensometrics leading to higher order structures

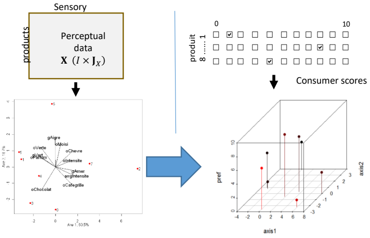
Developments in sensory evaluation and consumer studies (Free Choice Profiling, Free Sorting, Projective Mapping, CATA, ...) associated to higher-order data structures.





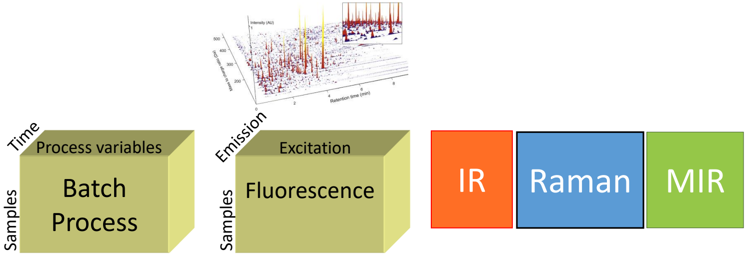






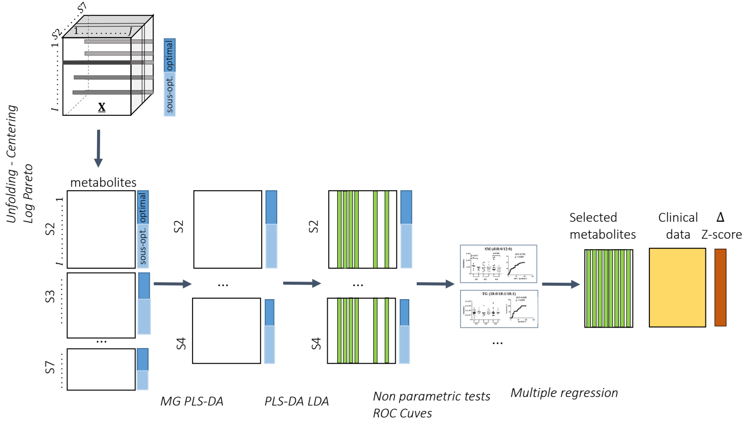
Handling omnifarious datasets from analytical platforms in chemometrics

In a data fusion framework to depict systems observed with different types of instrumental techniques (e.g. spectroscopic, chromatographic, imaging-based ones), at different time, in different conditions, or under varying experimental setups.



Data integration in a multi-omic perspective

Integration of multi-omic information in a meaningful way to provide a more comprehensive analysis of a biological point of interest (Ritchie et al., 2015)



Breast milk lipidome is associated with early growth trajectory in preterm infants (Alexandre-Gouabau, et *al.* ,2018).

Multiblock data structures

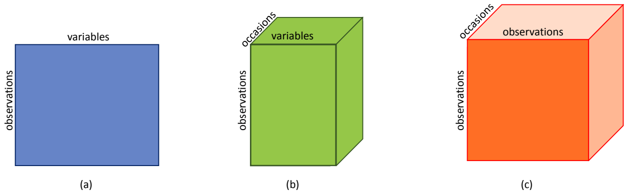


Figure 1 – Block with 2 modes (a), 3 modes (b) and 2 modes (c)

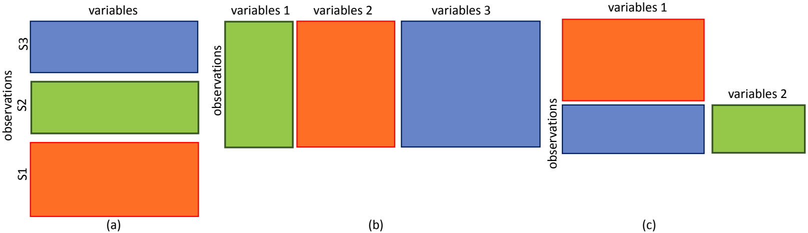
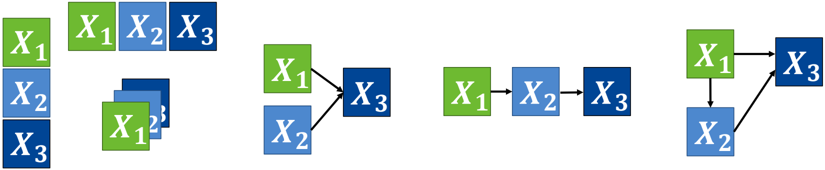


Figure 2 – Multiblock with a partitioning of rows (a) vs columns (b) and L-Shaped data (c)

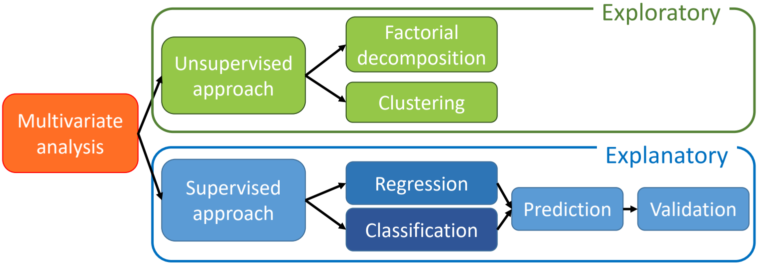
Multiblock data analysis

Data specificities

- Heterogeneous datasets
- Flat datasets $n \ll p$
- Missing values
- High multicolinearity



Multiblock data analysis



Main goals

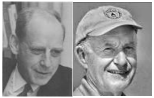
- Assess the commonalities and differences between the different data sets
- Take into account their linking relation
- Predict a phenotype or the outcome of an intervention
- Identify biomarkers / drivers of preference

Overview

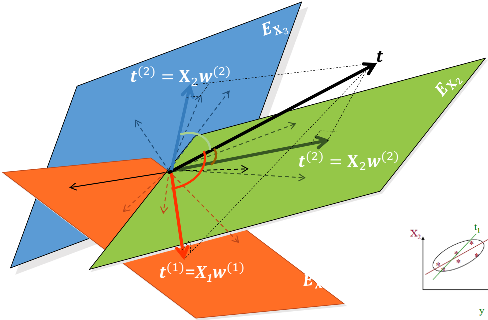
Key initial methods

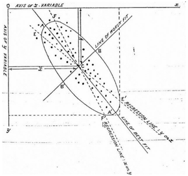
PCA
(Pearson, 1901)
Principal Component Analysis

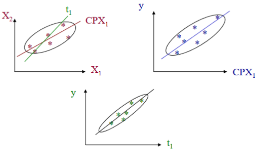
CCA / GCA
(Hotelling, 1936)
(Carroll, 1968)
(Generalized) Canonical Correlation Analysis



PLS
(Wold, 1966 ; 1975)
(Wold et al., 1984)
Non Linear Iterative Partial Least Squares





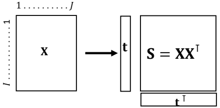


From M. Tenenhaus

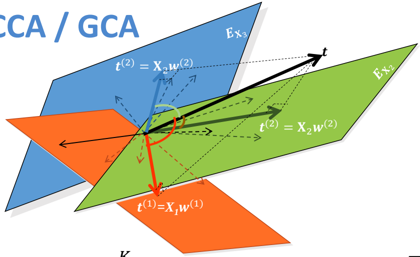
GCA from a criterion perspective

PCA

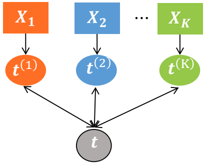
$$\max \sum_{j=1}^{J} cov^2(x_j, t)$$
with $t = \mathbf{X}w$ s.t. $\|t\| = 1$



CCA / GCA

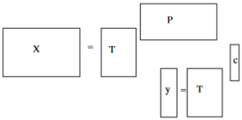


$$\max \sum_{k=1}^{K} cor^2(t^{(k)}, t)$$
with $t^{(k)} = \mathbf{X}_k w^{(k)}$ s.t. $\|t\| = 1$



PLS

$$\max \sum_{j=1}^{J} cov^2(y, t)$$
with $t = \mathbf{X}w$ s.t. $\|w\| = 1$



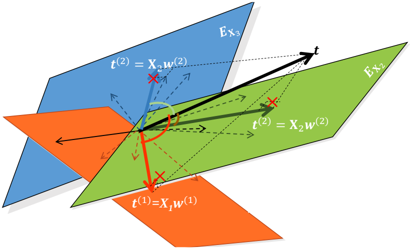
Unsupervised multiblock analysis

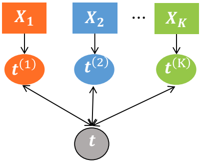
From GCCA (Carroll, 1968) :

- Common component : $\mathbf{t} \propto \sum_k \mathbf{t}^{(k)}$
- Block component : $\mathbf{t}^{(k)} = \mathbf{X}_k(\mathbf{X}_k^\top \mathbf{X}_k)^{-1}\mathbf{X}_k^\top \mathbf{t}$

To ComDim (Qannari et al., 2000 ; Qannari et al., 2001) :

- Common component : $\mathbf{t} \propto \sum_k \lambda^{(k)}\mathbf{t}^{(k)}$
- Block component : $\mathbf{t}^{(k)} = \mathbf{X}_k\mathbf{X}_k^\top \mathbf{t}$
- $max \sum_{k=1}^{K} cov^2(\mathbf{t},\mathbf{t}^{(k)})$ s.t. $\|\mathbf{t}\| = 1$ / INDSCAL $min \sum_{k=1}^{K} \left\| \mathbf{X}_k\mathbf{X}_k^\top - \lambda^{(k)}\mathbf{t}\,\mathbf{t}^\top \right\|_F^2$
- Salience $\lambda^{(k)}$ shows the importance of $\mathbf{X}_k$ in the determination of $\mathbf{t}$
- Iterative determination of the successive common components by deflation





A brief overview of unsupervied multiblock approaches (Mangamana et al., 2019)

ComDim : Common Component and Specific Weights Analysis
CPCA : Consensus Principal Component Analysis
HPCA : Hierarchical Principal Component Analysis
MCOA : Multiple CO-inertia Analysis
MFA : Multiple Factor Analysis
SCA : Simultaneous Component Analysis

`ade4: Analysis of Ecological Data: Exploratory and Euclidean Methods in Environmental Sciences`

Tools for multivariate data analysis. Several methods are provided for the analysis (i.e., ordination) of one-table (e.g., principal component analysis, correspor two-table (e.g., coinertia analysis, redundancy analysis), three-table (e.g., RLQ analysis) and K-table (e.g., STATIS, multiple coinertia analysis). The philosop package is described in Dray and Dufour (2007) <doi:10.18637/jss.v022.i04>.

`FactoMineR: Multivariate Exploratory Data Analysis and Data Mining`

Exploratory data analysis methods to summarize, visualize and describe datasets. The main principal component methods are available, those with the largest po terms of applications: principal component analysis (PCA) when variables are quantitative, correspondence analysis (CA) and multiple correspondence analysis variables are categorical, Multiple Factor Analysis when variables are structured in groups, etc. and hierarchical cluster analysis. F. Husson, S. Le and J. Pages (2

`MBAnalysis: Multiblock Exploratory and Predictive Data Analysis`

Exploratory and predictive methods for the analysis of several blocks of variables measured on the same individuals. The methods included are: Multiblock Principal Components Analysis (MB-PCA), Common Dimensions analysis (ComDim), Multiblock Partial Least Squares (MB-PLS) regression and Multiblock Weighted Covariate analysis (MB-WCov). E. Tchandao Mangamana, V. Cariou, E. Vigneau, R. Glèlè Kakaï, E.M. Qannari (2019) <doi:10.1016/j.chemolab.2019.103856>; E. Tchandao Mangamana, R. Glèlè Kakaï, E.M. Qannari (2021) <doi:10.1016/j.chemolab.2021.104388>.

`multiblock: Multiblock Data Fusion in Statistics and Machine Learning`

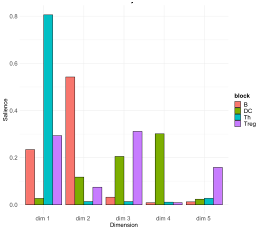
Functions and datasets to support Smilde, Næs and Liland (2021, ISBN: 978-1-119-60096-1) "Multiblock Data Fusion in Statistics and Machine Learning - Applications in the Natural and Life Sciences". This implements and imports a large collection of methods for multiblock data analysis with common interfaces, result- and plotting functions, several real data sets and six vignettes covering a range different applications.

`RGCCA: Regularized and Sparse Generalized Canonical Correlation Analysis for Multiblock Data`

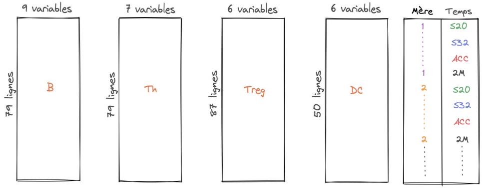
Multiblock data analysis concerns the analysis of several sets of variables (blocks) observed on the same group of individuals. The main aims of the RGCCA package are: (i) to study the relationships between blocks and (ii) to identify subsets of variables of each block which are active in their relationships with the other blocks.

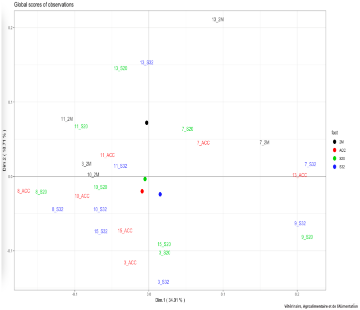
Illustration with immunology data in the framework of the ANR CIMMAP

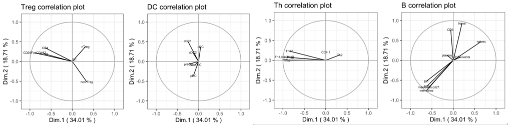
ComDim results



Characterising the effect of maternal prebiotic supplementation on perinatal Immune system maturation, Microbiota and breast Milk compositions for Allergy Prevention in high-risk children.

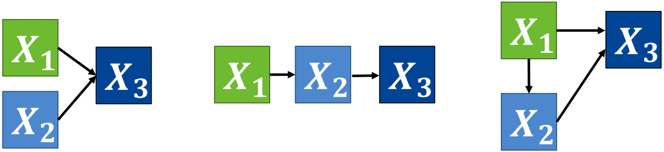






Overview

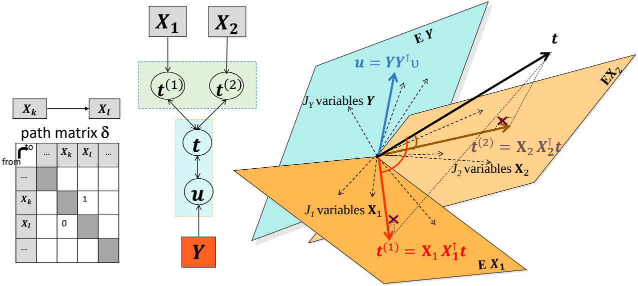
Supervised multiblock analysis



Integration of the relationships between blocks

P-ComDim, Path-ComDim (El Ghaziri et al., 2016 ; Cariou et al., 2018, 2019)

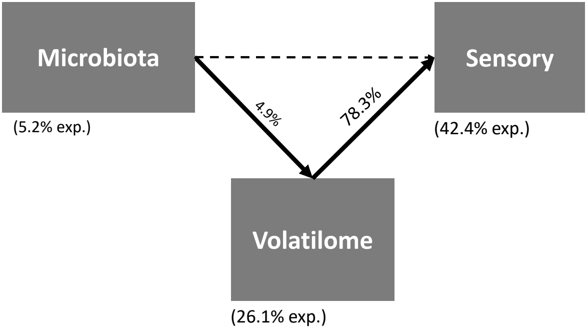
- Common components : $\mathbf{t} \propto \sum_{kl} \delta_{kl} \lambda^{(kl)} \mathbf{t}^{(k)}$ et $\mathbf{u} \propto \sum_{kl} \delta_{kl} \lambda^{(kl)} \mathbf{u}^{(l)}$
- Block components : $\mathbf{t}^{(k)} = \mathbf{X}_k \mathbf{X}_k^\top \mathbf{t}$ et $\mathbf{u}^{(l)} = \mathbf{X}_l \mathbf{X}_l^\top \mathbf{v}$
- Salience associated to each block $|\lambda^{(kl)}|$
- $max \sum_{k,l=1}^{K} \delta_{kl} cov^2(\mathbf{t}^{(k)}, \mathbf{u}^{(l)})$ s.t. $\|\mathbf{t}\| = 1$ / $min \sum_{k,l=1}^{K} \delta_{kl} \left\| \mathbf{X}_k \mathbf{X}_k^\top \mathbf{X}_l \mathbf{X}_l^\top - \lambda^{(kl)} \mathbf{t} \mathbf{v}^\top \right\|_F^2$



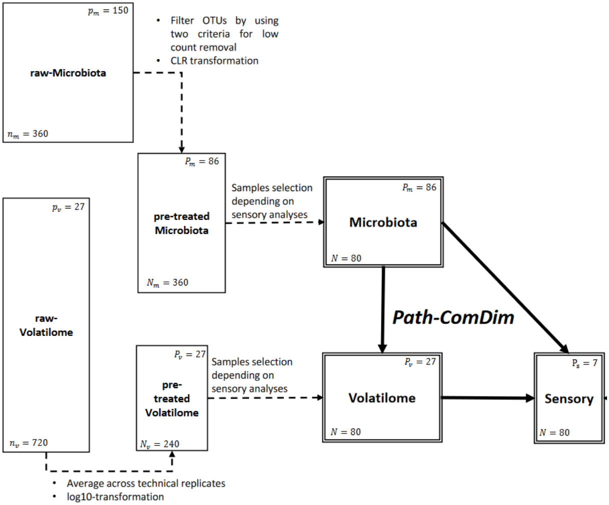
Application of a path-modeling approach within in the RedLosses project (Luong et al., 2020)

REDLOSSES
RÉDUCTION DES PERTES ALIMENTAIRES PAR LA PRÉDICTION DES ALTÉRATIONS MICROBIOLOGIQUES

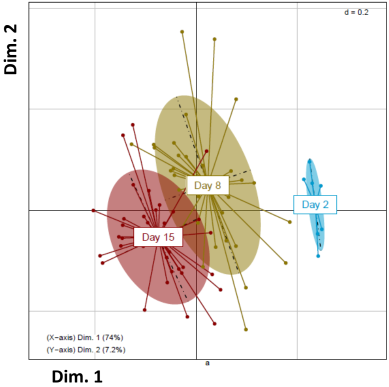
REDuction of food LOSSES by microbial spoilage prediction [French ANR project]

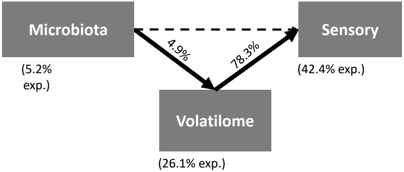


Application of a path-modeling approach for deciphering causality relationships between microbiota, volatile organic compounds and off-odour profiles during meat spoilage



Application of a path-modeling approach for deciphering causality relationships between microbiota, volatile organic compounds and off-odour profiles during meat spoilage



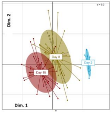


The first dimension structures the data according to storage time:

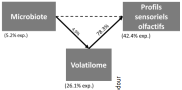
Dynamics of alteration characterized by the evolution of sensory profiles and the production of volatile compounds.

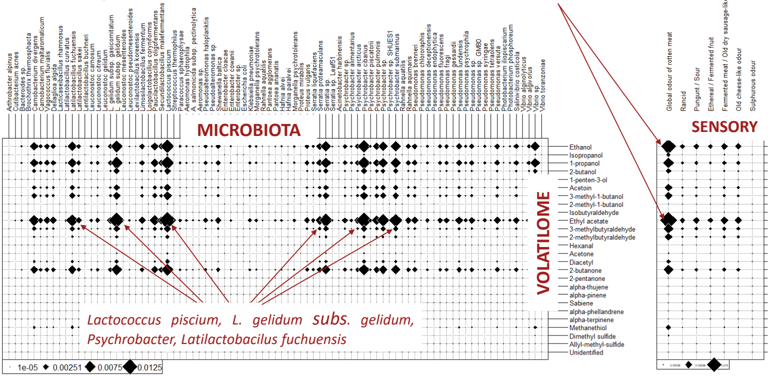
Microbiota, lower inertia:
Large number of species that do not all contribute to this dynamic.

Application of a path-modeling approach for deciphering causality relationships between microbiota, volatile organic compounds and off-odour profiles during meat spoilage



The smell of spoiled meat and the production of ethyl acetate and ethanol





Lactococcus piscium, L. gelidum subs. gelidum, Psychrobacter, Latilactobacilus fuchuensis

Overview

Conclusion

Multiblock approaches

- unsupervised and supervised methods mainly originated from psychometrics and chemometrics,
- genese from Canonical Correlation Analysis
- common issues between supervised multiblock approaches and path modeling
- increasing interest for Data Fusion and Data Integration in the study of complex systems toward holistic, data driven approach

Some challenges

- Predictive models in a path modeling context
- Introduction of non linearity with kernels
- take into account of a priori knowledge
- Partial couplings between blocks : Network PCA

References

Alexandre-Gouabau, et al. (2018). Breast milk lipidome is associated with early growth trajectory in preterm infants. Nutrients, 10(2), 164.

Cariou, V. et al. (2018). ComDim : From multiblock data analysis to path modeling. Food Quality and Preference, 67, 2734.

Cariou, V. et al. (2019). ComDim methods for the analysis of multiblock data in a data fusion perspective. In Data Handling in Science and Technology, vol. 31. Elsevier.

Carroll, J. D. (1968). Generalization of canonical correlation analysis to three or more sets of variables. Pages 227228 of Proceedings of the 76th annual convention of the American Psychological Association, vol. 3.

El Ghaziri et al. (2016). Analysis of multiblock datasets using ComDim : Overview and extension to the analysis of (K+ 1) datasets. Journal of Chemometrics, 30(8), 420429.

Hotelling, H. (1936). Relations between two sets of variates. Biometrika, 28(3/4), 321377.

Lahat, D. et al. (2015). Multimodal data fusion : an overview of methods, challenges, and prospects. Proceedings of the IEEE, 103(9), 1449-1477.

Luong, N. D. M. et al. (2021). Application of a path-modelling approach for deciphering causality relationships between microbiota, volatile organic compounds and off-odour profiles during meat spoilage. International Journal of Food Microbiology, 348, 109208.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin philosophical magazine and journal of science, 2(11), 559-572.

Ritchie, M. D. et al. (2015). Methods of integrating data to uncover genotypephenotype interactions. Nature Reviews Genetics, 16(2), 85-97.

Qannari, E. M. et al. (2000). Defining the underlying sensory dimensions. Food Quality and Preference, 11(1-2), 151154.

Qannari, E. M. et al. (2001). Common components and specific weights analysis performed on preference data. Food Quality and Preference, 12(5-7), 365368.

Mangamana, E.T. et al. (2019). Unsupervised multiblock data analysis : a unified approach and extensions. Chemometrics and Intelligent Laboratory Systems, 194, 103856.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. Multivariate analysis, 1, 391420.

Wold, H. (1973). Nonlinear iterative partial least squares (NIPALS) modelling : some current develop-ments. Pages 383407 of Multivariate AnalysisIII. Elsevier.

Wold H. (1975). Modelling in complex situations with soft information. In Third World Congress of Econometric Society, Toronto, Canada.

Wold, S. et al. (1984). The collinearity problem in linear regression, The partial least squares approach to generalized inverses. SIAM J. Sci. Stat. Comput. 52. 735743.

StatSC Research Unit

INRAE
la science pour la vie, l'humain, la terre

Oniris
ÉCOLE NATIONALE
VÉTÉRINAIRE, AGROALIMENTAIRE ET DE L'ALIMENTATION
SANTÉ ET ALIMENTATION AU CŒUR DE LA VIE

Thanks for your attention