Exploratory multiblock methods: do they lead to the same results?

C. Peltier¹,² B. Jaillais³ M. Hanafi³

¹Centre des Sciences du Goût et de l’Alimentation, CNRS, INRAE, Institut Agro, University of Bourgogne Franche-Comté, F-21000 Dijon
caroline.peltier@inrae.fr
²CNRS, INRAE, PROBE research infrastructure, ChemoSens facility, F-21000 Dijon, France,
³StatSC, Oniris INRAE, Rue de la Géraudière, 44322, Nantes, France

Keywords: multiblock methods, canonical factorization

1 Introduction
Extracting relevant information from multiblock data by reducing dimensionality, summarizing the information in an understandable way or visualizing multiblock data for interpretation purposes, are challenges often raised in chemometrics. When K data blocks denoted X_k (1 ≤ k ≤ K) are available and each data block X_k reflects the measurements of p_k quantitative variables on n individuals, several multiblock methods are proposed in the literature such as:
(i) Hierarchical Principal Component Analysis (HPCA) [1,2] that is identical to Common Component and Specific Weights Analysis (CCSWA) [3,4,5],
(ii) Consensus Principal Component Analysis (CPCA) [1,2] that is identical to Multiple Co-inertia Analysis (MCOA) [5,6],
(iii) Structuration de Tableau A Trois Indices de la Statistique (STATIS) [7,8].
(iv) Other multiblock exploratory methods less used in chemometrics such as MAXBET [9].
Relating these methods to each other is an important issue, that was modestly studied in the literature.

2 Material and methods
Recently, the canonical representation of multiblock methods was introduced from a factorization lemma for partitioned matrices [10]. It highlights the strategy adopted by these methods for analyzing multiblock data. This strategy involves two analyzes: (i) a global analysis described by a factorization of the whole data matrix. (ii) a block analysis described by the factorization of each block. The interpretation and visualization of the results of this canonical representation are based on the same principle as the usual Principal Component Analysis and allow the different methods to be directly compared.
Four multiblock datasets were considered for comparing the results of the different methods: butters (5 blocks, [11]), bigCheeses (3 blocks, [12,13]), meats (3 blocks, [14]) and cheeses (3 blocks, [15]).
Several multiblock methods were computed on these datasets with R.4.0.2 using RGCCA package [16]. Results were compared with several indicators based on explained variance, but also correlations between the individual projections.

3 Results and discussion
Figure 1 shows the explained variance (per dimension and per block) for the cheeses and the meats datasets. A convergence between the results was observed for cheeses while a divergence between results was observed for meats. Other comparison indicators led to the same results. Same approach will be used for STATIS method.

4 Conclusion
Canonical representation of multiblock methods allows different multiblock methods to be compared. Conducting this comparison on 4 different datasets showed that divergence (partial or total) between methods can happen.
5 References


[14] AlimaSSenS project, ANR-14-CE20-0003
