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

C. Peltier[1,2] B. Jaillais[3] M. Hanafi[3]
[1]Centre des Sciences du Goût et de l'Alimentation, CNRS, INRAE, Institut Agro, University of Bourgogne Franche-Comté, F-21000 Dijon
caroline.peltier@inrae.fr
[2]CNRS, INRAE, PROBE research infrastructure, ChemoSens facility, F-21000 Dijon, France,
[2] StatSC, 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 $\mathbf{X}_k$ $(1 \leq k \leq K)$ are available and each data block $\mathbf{X}_k$ 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.

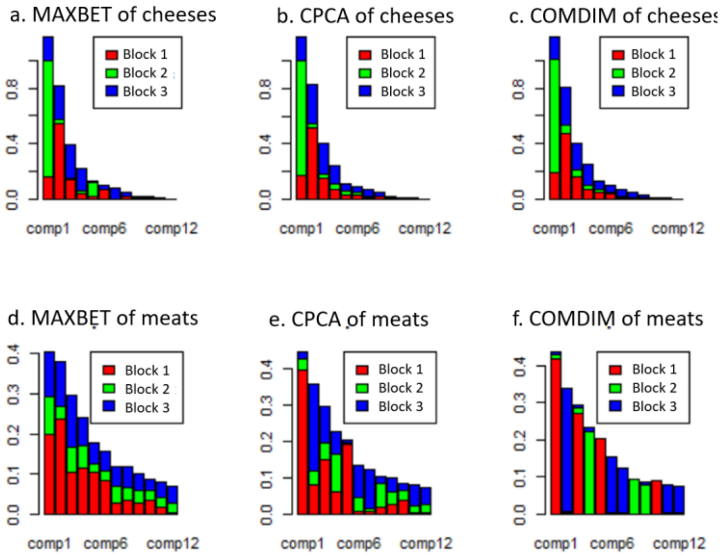

Figure 1: Explained variances according to the considered dimension for cheeses (a,b,c) and meats (d, e, f) depending on the chosen method (MAXBET, CPCA or COMDIM)

5 References

[1] Wold, S., Kettaneh, N. and Tjessem, K., (1996). Hierarchical multi-block PLS and PC models for easier interpretation and as an alternative to variable selection. *Journal of Chemometrics*, (10), 463-482.

[2] Westerhuis, J. A, Kourti, T., Macgregor J. F., (1998). Analysis of Multiblock and Hierarchical PCA and PLS Models. *Journal of Chemometrics*, (12), pp. 301-321.

[3] Qannari E. M,. Wakeling I., Courcoux Ph., MacFie M.F. (2000). Common Components and specific weights analysis performed on preference dataFood Quality and Preference, 11, 151-154.

[4] Mazerolles, G., Hanafi, M., Dufour, E., Qannari, E. M., Bertrand, D. (2006). Common Components and specific weights analysis: a chemometric method for dealing with complexity of food products. *Chemometrics and Intelligent Laboratory Systems*, (81), 41- 49.

[5] Hanafi, M., Mazerolles, G., Dufour, E., Qannari, E. M. (2006). Common components and specific weight analysis and multiple Co-inertia analysis applied to the coupling of several measurement techniques. J*ournal of Chemometrics*, (20), 5, 172-183.

[6]. Chessel, D., Hanafi, M. (1996). *Revue de Statistique Appliquée*, XLVI, (2), 35-60

[7]. S.Gourvenec, I. Stanimirova, C-A.Saby, C.Y.Airiau and D.L.Massart (2005). Monitoring batch processes with the STATIS approach. *Journal of Chemometrics*, 19: 288–300

[8] C. Lavit, Y. Escoufier , R. Sabatier, P Traissac (1994). The ACT (Statis method). *Computational Statistics & Data Analysis*, 18(1):97 119.

[9] Hanafi, Mohamed & Kiers, Henk A.L., 2006. Analysis of K sets of data, with differential emphasis on agreement between and within sets, *Computational Statistics & Data Analysis*, vol. 51(3), 1491-1508

[10] Hanafi, M., Canonical representation of exploratory multiblock methods, *Chimiométrie 2019 Montpellier*, https://chemom2019.sciencesconf.org/data/03_chemom2019_MohamedHanafi.pdf

[11] Jaillais B, Morrin V, Downey G, Image processing of outer-product matrices - A new way to classify samples. Examples using visible/NIR/MIR spectral data. Chemometrics and Intelligent Laboratory Systems, 2007, 86, 179-188.

[12] Mazerolles, G., Hanafi, M., Dufour, E., Qannari, E. M., Bertrand, D. (2006). Common Components and specific weights analysis: a chemometric method for dealing with complexity of food products. Chemometrics and Intelligent Laboratory Systems. (81), 41- 49.

[13] Hanafi, M., Mazerolles, G., Dufour, E., Qannari, E. M. (2006). Common components and specific weight analysis and multiple Co-inertia analysis applied to the coupling of several measurement techniques. Journal of Chemometrics. (20), 5, 172-183.

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

[15] Adi project, http://plasticnet.grignon.inra.fr/portailbagatel

[16] Tenenhaus, A. and Tenenhaus, M. (2011). Regularized generalized canonical correlation analysis. Psychometrika 76, 257–284. https://github.com/rgcca-factory/RGCCA