Multifactorial designed data from multiple sources? Analyze them with a single model using AComDim

M. de Figueiredo[1] S. Giannoukos[1] C. Wüthrich[1] R. Zenobi[1] D. N. Rutledge[2,3]

[1] Department of Chemistry and Applied Biosciences, ETH Zurich, Zurich, Switzerland
[2] National Wine and Grape Industry Center, Charles Sturt University, Wagga Wagga, Australia
[3] ChemHouse Research Group, Montpellier, France

Keywords: AComDim, ANOVA, ComDim, multiblock, multifactorial designs.

1 Introduction

A novel chemometric approach is proposed to analyze high-dimensional data collected from multiple analytical platforms sharing the same multifactorial design of experiments. Although methods combining analysis of variance (ANOVA) with principal components analysis (PCA) or simultaneous components analysis (SCA) efficiently tackle multifactorial designed data, they require the construction and interpretation of one model for each experimental factor and interaction, and each data source. Taking advantage of the multiblock nature of the Common Dimensions (ComDim) method, we propose an extension of the original method ANOVA-ComDim (AComDim) [1] for the analysis of multifactorial designs from one source to multiple sources [2]. As an example, Figure 1 describes the algorithm implementation for two data sources and a 2-fixed effect factors experimental design with 2 and 3 levels, respectively.

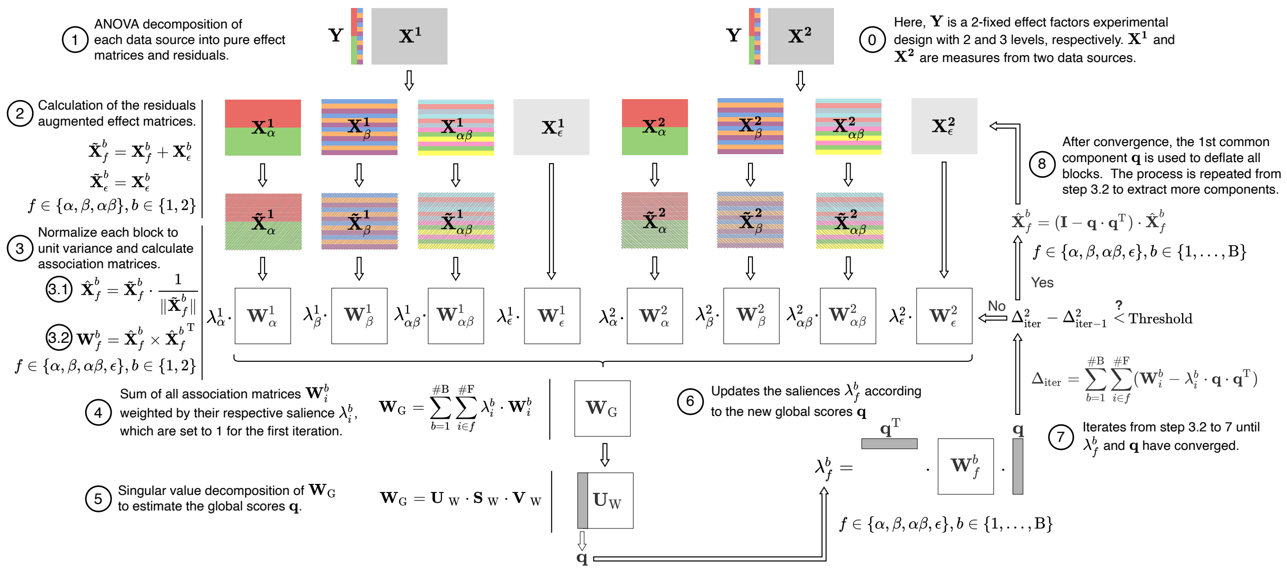

Figure 1 – AComDim algorithm implementation for the analysis of 2 data sources.

2 Application example: the lignin dataset with 8 blocks and 3-fixed effect factors

The lignin dataset consists in different nuclear magnetic resonance (NMR) relaxation curves of lignin-starch mixtures in blocks B1 to B8 from a design with Factors A (humidity), B (form) and C (lignin concentration) with 2, 2 and 5 levels, respectively. The extension of the AComDim method to multiple sources was applied simultaneously to all data blocks.

3 Results

The ANOVA decomposition of each of the 8 blocks of the lignin dataset leads to 7 effect matrices, for a total of 56 matrices. With ANOVA-SCA or ANOVA-PCA, 56 models would have to be built but this new extension of AComDim provides a global picture of the results using a single model. Figure 2 shows the AComDim results for 8 data blocks, or sources. Main effects A, B and C are each described by common component (CC) 3, 4 and 5, respectively.

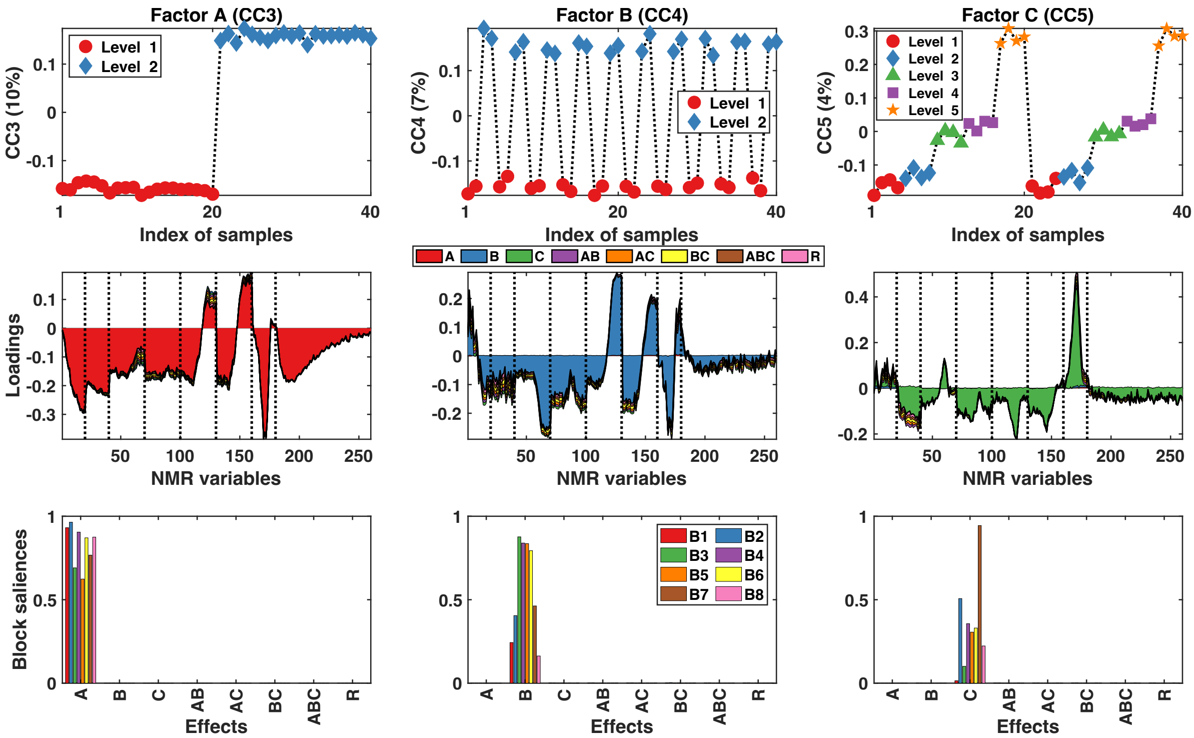

Figure 2 – AComDim results for scores (top), loadings (middle) and saliences (bottom).

Figure 2 row 1 shows the separation between the factor levels using the scores, row 2 shows the contributions (loadings) of the variables in each block to the construction of the associated CC and row 3 shows the saliences of each block for each of the effects and residuals.

4 Key takeaways

AComDim provides a global picture of the main sources of variation in complex data structures produced from multiple sources. To do so, AComDim builds a single model for all main effects and interactions instead of one for each one of them, thus helping the interpretation process. The saliences produced by AComDim are critical for the interpretation because they provide insights into the relationship between the data sources and their contribution to each of the effects.

5 References

[1] Jouan-Rimbaud Bouveresse, D., Pinto, R. C., Schmidtke, L. M., Locquet, N., & Rutledge, D. N. Identification of significant factors by an extension of ANOVA–PCA based on multi-block analysis. *Chemometrics and Intelligent Laboratory Systems*, 106(2), 173-182, 2011.

[2] de Figueiredo, M., Giannoukos, S., Wüthrich, C., Zenobi, R., & Rutledge, D. N. A tutorial on the analysis of multifactorial designs from one or more data sources using AComDim. *Journal of Chemometrics*, e3384, 2022.