Handling unbalanced multifactorial designs with Rebalanced ASCA

M. de Figueiredo[1,2,3] S. Giannoukos[1] S. Rudaz[2,3] R. Zenobi[1] J. Boccard[2,3]

[1] Department of Chemistry and Applied Biosciences, ETH Zurich, Zurich, Switzerland

[2] School of Pharmaceutical Sciences, University of Geneva, 1211 Geneva, Switzerland

[3] Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, 1211 Geneva, Switzerland

Keywords: Rebalancing, ANOVA, ASCA, unbalanced designs.

1 Introduction

A novel chemometric approach is proposed to analyze high-dimensional data collected from unbalanced designs of experiments. It combines a rebalancing strategy with the ASCA method under the name Rebalanced ASCA (RASCA). The ability of RASCA to handle unbalanced designs was compared with classical ASCA, as well as its latest improvements, such as ASCA+ [1] and WE-ASCA [2]. Figure 1 briefly describes the RASCA algorithm.

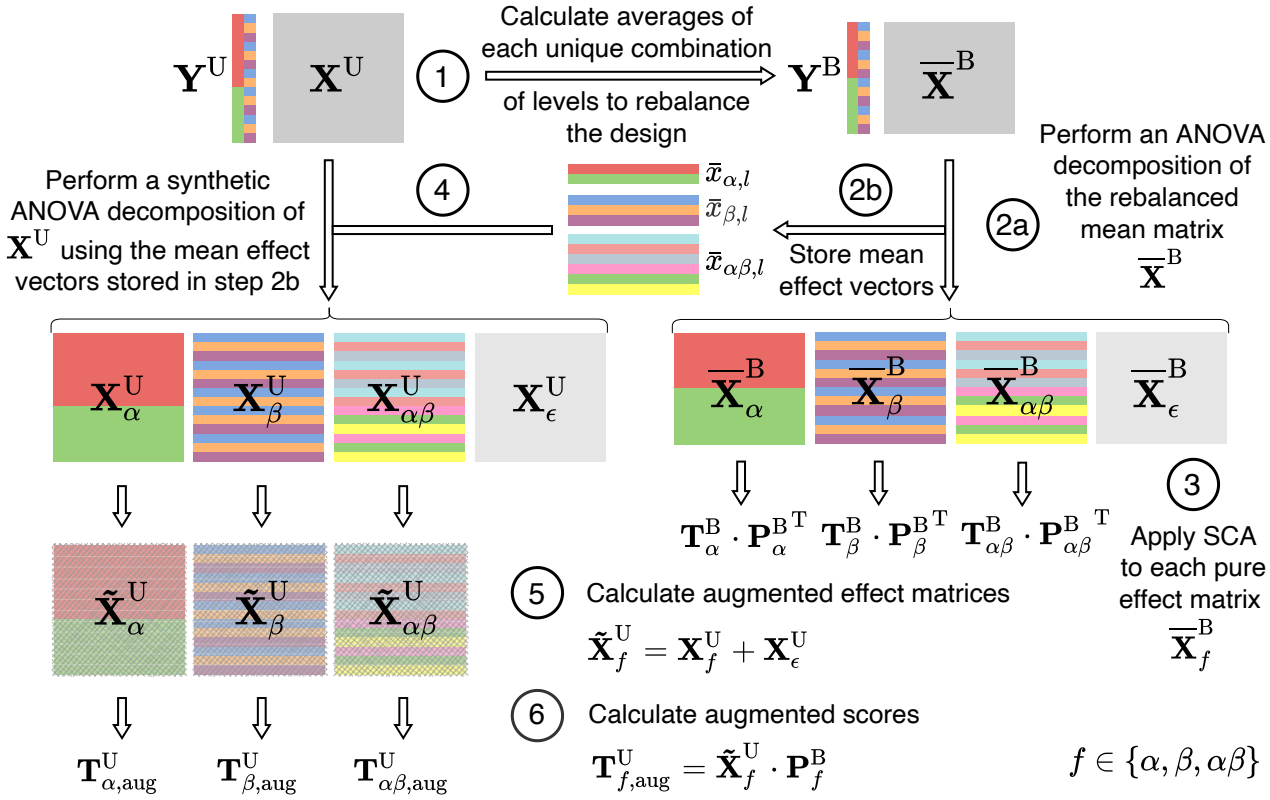

Figure 1 – RASCA algorithm and brief description of its mains steps (1 to 6)

2 Experimental: comparison framework and dataset

A framework was designed to provide a systematic comparison of the various approaches. For that purpose, a real dataset obtained from an initially balanced design was gradually unbalanced in a controlled fashion by removing observations belonging to specific combinations of factor levels. This procedure was repeated several times and ASCA results obtained with the complete dataset were

chosen as a reference case. This allowed an objective evaluation of the ability of the different methodologies to handle increasingly unequal group sizes. Scores and loadings obtained with each approach were compared with the reference case using the modified RV-coefficient (RV_2) [3], for which values close to 1 indicate high correlation between matrices.
This comparison framework was applied to metabolomic data collected from *Arabidopsis thaliana* after leaf wounding obtained from a full-factorial design with two fixed effect factors, namely plant type and time after wounding, with 2 and 4 levels, respectively, leading to a total of 72 observations characterized by 124 features (peak intensities). Autoscaling was applied to all features. Data are publicly available via https://www.metaboanalyst.ca/resources/data/cress_time.csv, 02.05.2022.

3 Results and discussion

Figure 2 shows the RV_2 curves between scores obtained in the unbalanced and balanced solutions. All methods considered led to identical solutions with the initial balanced design, while increasing differences appeared when the design was gradually unbalanced.

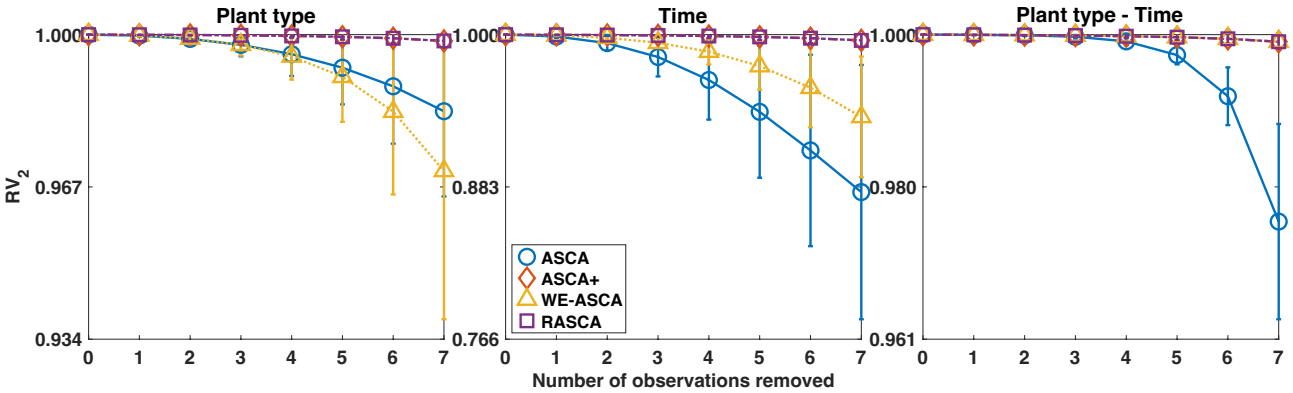

Figure 2 – RV_2 of the scores as a function of the number of observations removed.

The proposed benchmark showed that classical ASCA and WE-ASCA led to lower RV_2 for all effects, except for the interaction term with WE-ASCA, which may be related to the fact that the interaction term estimated by WE-ASCA is orthogonal to the main effects. On the other hand, RASCA and ASCA+ showed consistently high correlation with the balanced solution for the main and interaction effects. RASCA demonstrated to be suited to tackle unbalanced designs. Unlike ASCA+, RASCA also solves the issue of the non-orthogonality of the effect matrices in addition to unbiased parameter estimators, which may be of utmost importance for the interpretation of the models when facing unbalanced designs.

4 Conclusions

The comparison of RASCA with state-of-the-art methods demonstrated its adequacy to handle unbalanced designs and further investigations with other datasets will be conducted in this controlled framework.

5 References

[1] Thiel, M., Féraud, B., & Govaerts, B. ASCA+ and APCA+ : Extensions of ASCA and APCA in the analysis of unbalanced multifactorial designs. *Journal of Chemometrics*, 31(6), e2895, 2017.

[2] Ali, N., Jansen, J., van den Doel, A., Tinnevelt, G. H., & Bocklitz, T. WE-ASCA : The Weighted-Effect ASCA for Analyzing Unbalanced Multifactorial Designs—A Raman Spectra-Based Example. *Molecules*, 26(1), 2020.

[3] Smilde, A. K., Kiers, H. A. L., Bijlsma, S., Rubingh, C. M., & van Erk, M. J. Matrix correlations for high-dimensional data : The modified RV-coefficient. *Bioinformatics*, 25(3), 401-405, 2009.