The R4multidata project: Comparison of R tools for multidimensional data analysis. Example with RGCCA and mixOmics for supervised methods

Marion Brandolini-Bunlon, Elise Maigne, Sébastien Theil, Isabelle Sanchez, Virginie Rossard, Eric Latrille, Gwendal Cueff, Marie Tremblay-Franco, Nadia Bessoltane, Caroline Peltier, et al.

The R4multidata project: Comparison of R tools for multidimensional data analysis. Example with RGCCA and mixOmics for supervised methods

Marion BRANDOLINI-BUNLON[1] Elise MAIGNE[2] Sébastien THEIL[3] Isabelle SANCHEZ[4] Virginie ROSSARD[5] Eric LATRILLE[6] Gwendal CUEFF[7] Marie TREMBLAY-FRANCO[8] Nadia BESSOLTANE[9] Caroline PELTIER[10,11] Alyssa IMBERT[12]

[1] Université Clermont Auvergne, INRAE, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, 63000 Clermont-Ferrand, France, marion.brandolini-bunlon@inrae.fr

[2] Université Fédérale de Toulouse, INRAE, MIAT, 31326 Castanet Tolosan, France, elise.maigne@inrae.fr

[3] Université Clermont Auvergne, INRAE, VetAgro Sup, UMR545 Fromage, 15000 Aurillac, France, sebastien.theil@inrae.fr

[4] INRAE, MISTEA, 2 place Pierre Viala, 34060 Montpellier, France, isabelle.sanchez@inrae.fr

[5] INRAE, Univ Montpellier, LBE, 102 Avenue des Etangs, F-11100 Narbonne, France, virginie.rossard@inrae.fr

[6] INRAE, Univ Montpellier, LBE, 102 Avenue des Etangs, F-11100 Narbonne, France, eric.latrille@inrae.fr

[7] Université Clermont Auvergne, INRAE, UNH, 63000 Clermont-Ferrand, France, gwendal.cueff@inrae.fr

[8] Toxalim, Université de Toulouse, UMR INRAE 1331, Metabohub-Metatoul-AXIOM, 31027 Toulouse cedex 3, France, marie.tremblay-franco@inrae.fr

[9] IJPB - Institut Jean-Pierre Bourgin - Sciences du végétal, nadia.bessoltane@inrae.fr

[10] Université Bourgogne Europe, Institut Agro, CNRS, INRAE, UMR CSGA, 21000 Dijon, France

[11] Probe Research Infrastructure, Chemosens facility, CNRS-INRAE, Dijon, France, caroline.peltier@inrae.fr

[12] INRAE, Université Clermont Auvergne, Vetagro Sup, UMRH, 63122 Saint-Genès-Champanelle, France, alyssa.imbert@inrae.fr

Keywords: multidimensional data analysis, R software, comparison.

1 Introduction

Multidimensional methods are essential for analyzing complex data (omics, spectral, etc.). Many R packages exist, but they have different philosophies. This leads users to question their differences in terms of functionality, maintenance, reproducibility, and results. The R4multidata project aims to create a standardized and collaborative environment for testing and comparing, with real and simulated data, the functions of these packages in terms of approach and application. In the event of algorithmic differences, the associated limitations are studied. The ultimate goal is to provide the necessary elements for making informed choices about tools.

2 Material and methods

In the statistical analysis and integration of complex heterogeneous data, multidimensional methods are mainly applied. Among the R packages, mixOmics [1] is one of the most widely used (2,500 to 3,000 downloads of the package per month). Functions in the initial mixOmics package were built based on methods developed by the authors of the RGCCA package [2], but in a way that simplifies their use by biologists. Besides, RGCCA was developed for including more methods in a unique and

general framework [3]. These two packages therefore share a number of basic statistical methods, such as partial least squares (PLS) regression and its discriminant ("DA") or variable selection ("sparse") versions, Canonical Correlation Analysis (CCA), and their derived multiblock methods considering more than two blocks of data. However, the two packages then evolved independently, still with two different philosophies.

In the R4multidata project, eight methods from these two packages were considered: PLS regression and discriminant, sparse, and/or multiblock variants (PLS, PLS-DA, sPLS, sPLS-DA, mbPLS, mbPLS-DA, mbsPLS, mbsPLS-DA). Initially, these methods were studied from a theoretical point of view (optimization problem, initialization, deflation, block weighting, regularization, variable selection method, missing values handling, prediction methods). Their implementations in the two packages were compared (inputs and outputs of the main functions, functions to tune or evaluate the models, plots). At the same time, comparison criteria were determined, and datasets were prepared using real data from research projects or available in the packages, before the application of the functions and the comparison of the results.

3 Results and discussion

The conceptual differences that have been identified between the packages, are due to the fact that different parameters have been set by the developers, or left to the user's choice. For example, four deflation modes are offered in mixOmics, whereas a single mode is offered in RGCCA. Conversely, in multiblock methods, the sum of covariances is always maximized in mixOmics, whereas it is possible to maximize the sum of covariances, the squares of covariances, or the absolute values of covariances in RGCCA. Moreover, data blocks can be weighted in RGCCA and not in mixOmics. There are also differences in the strategy applied, particularly for prediction in multiblock methods, which is done by block in mixOmics before averaging, and which is done from concatenated components in RGCCA.

Application to the datasets shows that, among other things, with equivalent settings, the first components obtained with the two packages are often comparable, and differences appear in the subsequent components.

There are also clear differences in terms of purpose and target users: The mixOmics R package is intended for use in regression and discriminant analysis (several classification methods and performance indicators dedicated to regression or discrimination), and/or by novice users. Conversely, the RGCCA R package is more intended for experienced users, as it allows for a more rigorous approach (more refined parameterization) and as it allows, through a judicious choice of parameters, a greater number of methods to be applied.

4 Conclusion

Although based on the same framework, the algorithms and functions of the two packages for applying PLS or PLS-DA regression with their sparse and multiblock variants have many differences that should be clearly identified when performing analyses to make the better choices.

5 References

[1] Rohart F., Gautier, B, Singh, A and Lê Cao, K. A. (2017) mixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol* 13(11): e1005752. doi: 10.1371/journal.pcbi.1005752.

[2] Tenenhaus, A., Tenenhaus, M. (2011) Regularized Generalized Canonical Correlation Analysis. *Psychometrika* 76(2), 257–284. doi: 10.1007/s11336-011-9206-8

[3] Tenenhaus M., Tenenhaus A. and Groenen P. J. (2017). Regularized generalized canonical correlation analysis: a framework for sequential multiblock component methods. *Psychometrika*, 82(3), pp.737-777. doi: 10.1007/s11336-017-9573-x.