The R4multidata project: comparison of multidimensional data analysis R tools. 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]
Alyssa IMBERT[11]

Summary

Multidimensional methods are widely used to analyze and integrate complex and heterogeneous data such as omics or spectral data. Although many R packages implement these methods, their differing philosophies and implementations make it challenging for users to choose the most suitable tools. The R4multidata project was created to provide a standardized and collaborative framework for testing and comparing multidimensional methods using real and simulated datasets, with the aim of supporting informed methodological choices.
This work focuses on two major R packages: ***mixOmics*** and ***RGCCA***. ***mixOmics*** is one of the most widely used packages and is designed to make advanced methods accessible to biologists, whereas ***RGCCA*** is developed to provide a general and flexible framework. Both packages share a common statistical base and share several methods, including Partial Least Squares (PLS), PLS Discriminant Analysis (PLS-DA), Canonical Correlation Analysis (CCA), and their sparse and multiblock extensions. However, as they have evolved independently, important methodological and practical differences have emerged.
Eight PLS-based methods from the two packages were examined, including regression, discriminant, sparse, and multiblock variants. In our project, the comparison covered both theoretical aspects (optimization, initialization, deflation, block weighting, regularization, handling of missing values, variable selection and prediction strategies) and implementation details (function inputs and outputs, tuning, evaluation, and visualization tools). Real datasets were used to compare results.
Interestingly, the two packages have different settings leading to different methods. ***mixOmics*** seems well suited for an initial approach with fewer parameters to fine-tune than ***RGCCA***, which offers a more flexible and customizable framework. Moreover, even with equivalent settings, the results slightly differ, partly due to the prediction and deflation algorithms.

Keywords (3 - 5) : multidimensional data analysis, R software, comparison.

[1] Université Clermont Auvergne, INRAE, marion.brandolini-bunlon@inrae.fr
[2] Université de Toulouse, INRAE, elise.maigne@inrae.fr
[3] Université Clermont Auvergne, INRAE, VetAgro Sup, sebastien.theil@inrae.fr
[4] INRAE, isabelle.sanchez@inrae.fr
[5] INRAE, Université Montpellier, virginie.rossard@inrae.fr
[6] INRAE, Université Montpellier, eric.latrille@inrae.fr
[7] Université Clermont Auvergne, INRAE, gwendal.cueff@inrae.fr
[8] Université de Toulouse, INRAE, marie.tremblay-franco@inrae.fr
[9] Institut Jean-Pierre Bourgin, nadia.bessoltane@inrae.fr
[10] Université Bourgogne Europe, Institut Agro, CNRS, INRAE, caroline.peltier@inrae.fr
[11] Université Clermont Auvergne, INRAE, VetAgro Sup, alyssa.imbert@inrae.fr

Development

Multidimensional methods are essential for analyzing and/or integrating complex data (omics, spectral…). Many R packages exist, but have different philosophies. This leads users to question their differences of functionalities, maintenance, reproducibility, results. The R4multidata project aims to create a community for testing and comparing the functions of these packages. The goal is to provide the elements for making informed choices about tools.

Among the R packages, ***mixOmics*** (Rohart et al., 2017) is one of the most widely used (2,500 to 3,000 downloads per month). Functions in the initial ***mixOmics*** package were built based on methods developed by the authors of the ***RGCCA*** package (Tenenhaus et Tenenhaus, 2011) in a way that simplifies their use. Besides, ***RGCCA*** was developed for including more methods within a general framework (Tenenhaus et al., 2017). These packages therefore share a number of basic statistical methods, such as partial least squares (PLS) regression and its discriminant ("DA") or variable selection ("sparse") or multiblock versions. These packages then evolved independently, still with different philosophies. In the R4multidata project, eight methods were considered: PLS regression and discriminant, sparse and/or multiblock variants ((multiblock-)(sparse-)PLS(-DA)). These methods were studied from a theoretical point of view. Their implementations in the packages were compared. At the same time, datasets from research projects or available in R packages were prepared, and functions were developed to compare the results from ***mixOmics*** and ***RGCCA*** on the same datasets, with the same settings.

The main conceptual and theoretical differences identified between the packages, are due to different parameters being set by developers, or left to user's choice. For instance, ***mixOmics*** provides up to four deflation modes for PLS, while ***RGCCA*** offers only one. Conversely, data blocks can be weighted in ***RGCCA*** but 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 from concatenated components in ***RGCCA***. In terms of variable selection, ***RGCCA*** uses a parsimony parameter for Lasso penalization. In contrast, ***mixOmics*** requires the user to specify the number of variables per block and component. 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 ones.

Although based on the same initial framework, the two packages have different settings leading to different methods. ***mixOmics*** seems well suited for an initial approach with fewer parameters to fine-tune than ***RGCCA***, which offers a more flexible and customizable framework. Moreover, even with equivalent settings, the results slightly differ, partly due to the prediction and deflation algorithms.

Références

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.

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

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.