R4multidata project

An user and developer community of multidimensional data analysis tools with R software.

Innovative Project 2025 supported by DipSO

Alyssa Imbert, Cati SysMics

Challenges of multidimensional methods with R

Context: Multivariate analyses as Principal Component Analysis (PCA), Partial Least Square Regression (PLS, PLSR) or multi-blocks approaches are essential for analyzing complex data. However, the large number of R packages applying the same method raises questions in terms of:

- Results and reproducibility
- Settings
- Maintenance

Aim: establish a common reference framework for comparing R packages according to:

- Their approach (algorithm, results)
- Their application (input data, parameters, outputs).

$\Rightarrow$ reasoned choice of tool

Methodology

1. **Rigorous benchmark:**
   - Tests on simulated and real standardized datasets
   - Evaluation of results and performance (compliance/conformity of results, stability, execution speed, management of limit cases, user-friendliness,…)
     - First packages considered: mixOmics and RGCCA
     - First methods studied: PLS, PLS-DA, sparse-PLS, sparse-PLS-DA, multi-block PLS, multi-block PLS-DA, multi-block sparse-PLS, multi-block sparse-PLS-DA
2. **Gap analysis:**
   - Differences related to algorithms
   - Differences related to configuration options and applications

Participants: 11 statisticians and (bio-)computer scientists of 5 different CATIs (eMPrEInTE, SysMics, CODEX, BOOM, Bios4Biol)

Practical organization, deliverables

Project phases	Delivrables
preparation of comparative study of R package functions *5 month (January-May 25)*	- method theory - comparison plan - comparison criteria - collaborative work tools - real and simulated datasets
Comparison work *hackathon 2 days (May 25)*	- R scripts for applying methods - Rdata derived from the application to the data - Summary of the obtained results
Analysis, formatting and dissemination of results (in progress)	

Nextcloud space

r4multidata

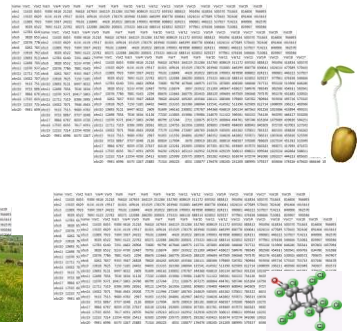
r4multidataresultats

r4multidata

Internal training within the group

PDF

3 datasets



hackathon

Developed R functions

Configure & build models

findKeepx(),
run_mixomics_rgcca()

validate, assess their performance

run_perf_mixrgcca()
rend_perf_table()

rmse or cross-validated classification errors, explained variances of the different blocks… estimated by package functions and recalculated

compare_criteres()
compare_rmse()
compare_selected_vars()
compare_stability()

Correlations between scores or between loadings of the two packages, comparison of rmse, variables selected by sparse methods, stability of feature selection estimated by package functions

compare models

First results. Datasets from BestDairy project

Aim of BestDairy project: demonstrate the variations in milk composition linked to rearing practices and the technologies implemented

- Many datasets,
  - **Individual milks** (6 datasets)
  - Mixed milks (n=12, 9 datasets)
  - Fermented milks (11 datasets)
  - Cheese (n=12, 13 datasets)
  - Two categories: System AE or IN

Dataset	Nrows	Ncols
GG protein	20	693
Lipids	**76**	**92**
Major proteins	**37**	**26**
Oligosac	**38**	**11**
Physicochemical	78	11
Zootechnical	20	15

- Heterogeneity: different types of omics (protein, lipids, etc.), physicochemical, zootechnic, microbiome, sensory analyses, etc.
- Not the same number of samples between datasets
- Number of variables: depending of dataset, p = 7 to several hundred variables

https://syalsa.hub.inrae.fr/thematiques/qualites-des-aliments/bestdairy-projet-exploratoire-2021-2023

Functions and applied methods

Methods tested on Bestdairy datasets:

- dataset restricted to the same samples (n=37)

Methods	X/X1	X2	Y
Partial Least Square (PLS canonical & regression mode)	Lipids	∅	Major proteins
Sparse PLS (sPLS canonical & regression mode)	Lipids	∅	Major proteins
PLS-DA & sPLSDA	Lipids	∅	Category (AE/IN)
Multiblock (s)PLS	Lipids	Oligosac	Major proteins
Multiblock (s)PLS-DA	Lipids	Oligosac	Category (AE/IN)

Each method was therefore executed with {mixOmics} and {RGCCA}, and the outputs and performance were compared.

Sparse multi-block PLS-DA regression results

```
> mbsplsda_run <- run_mixrgcca(parametres = list(method = "block.splsda", scale = TRUE,
                                                 sparsity = c(0.4, 0.5, 1)),
                               datalist = list("X“ = X_block, "Y_quali“ = Y_quali))
```

```
mbsplsdascale_run$times
```

```
Time differences in secs
        mix       rgcca
0.006419897 0.019752026
```

Sparsity: c(0.4,0.5,1) ➛ number of selected features per component & block

block	Comp1	comp2
X1	23	27
X2	6	4

Connection between blocks

```
mbsplsdascale_run$runmixomics$design
```

```
   X1 X2 Y
X1  0  0 1
X2  0  0 1
Y   1  1 0
```

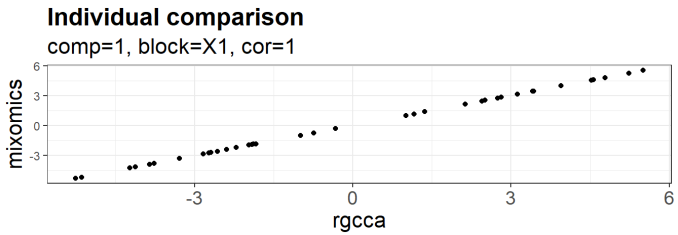
```
mbsplsdascale_run$runrgcca$call$connection
```

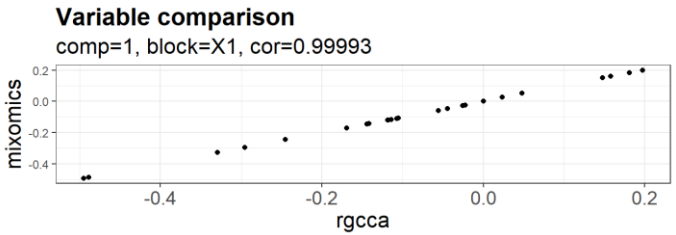
```
   X1 X2 Y
X1  0  0 1
X2  0  0 1
Y   1  1 0
```

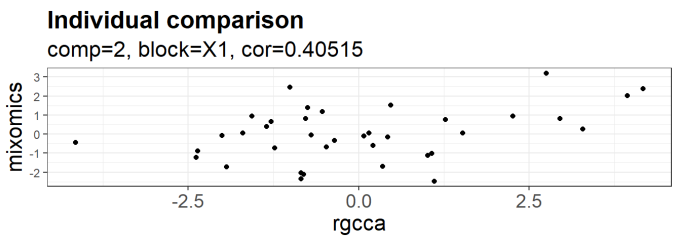
Observation scores and variable loadings

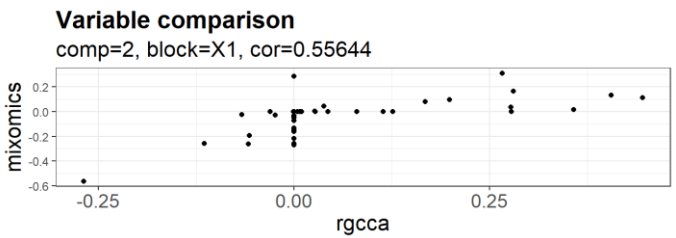
> rescc_mbsplsda <- compare_criteres(mbplsda_run$runmixomics, mbplsda_run$runrgcca, return_graph = TRUE)

X1

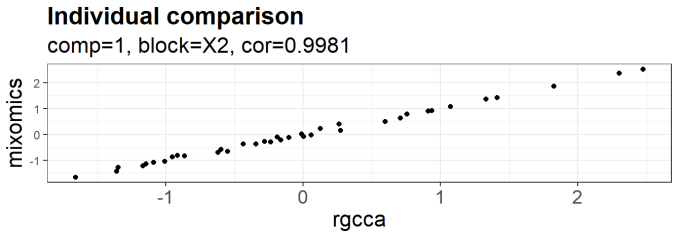


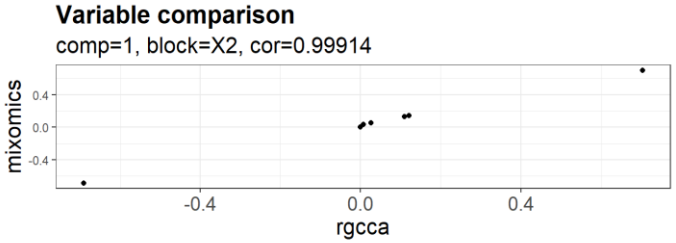


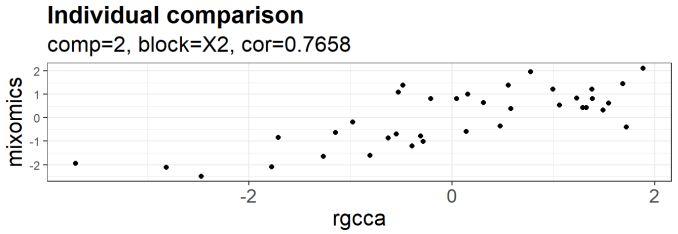


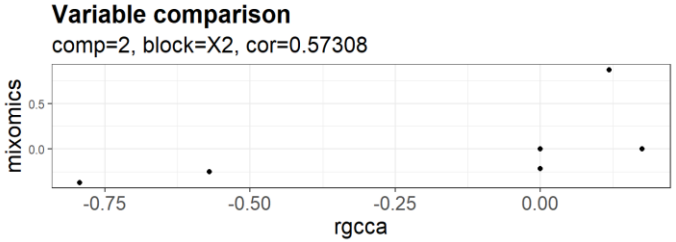


X2









Performance comparison with multiblock PLS-DA

```
> output_perf_mbsplsda <- run_perf_mixrgcca(models = mbsplsda_run, cvmethod = "kfold",
                                            nfolds = 5, nbpermut = 5, seed = 123)

> rend_perf_table(output_perf_mbsplsda)
```

package	Expl. Var. X1 comp1 (%)	Expl. Var. X1 comp2 (%)	Expl. Var. X2 comp1 (%)	Expl. Var. X2 comp2 (%)	Expl. Var. Y comp1 (%)	Expl. Var. Y comp2 (%)
mixOmics	40.99	5.17	12.02	18.44	100	59.42
RGCCA	40.97	6.91	11.37	20.33	NA	NA

package	Total error (%)	Total Err. CV (%)	Total Err. CV recalculated (%) mean (sd)
mixOmics	**35.14**	**5.59**	0.77 (1.58)
RGCCA	0	0.1	1.53 (1.69)

mixOmics performs a prediction per block, then a majority vote. In the event of a discrepancy between the predictions of the two blocks, 'NA' is used, which distorts the calculation of classification error rates.

Comparison of feature selection and stability selection

```
stab_mbsplsda <- compare_stability(output_run_rgcca = mbsplsda_run$runrgcca,
                                   output_run_mixomics = mbsplsda_run$runmixomics,
                                   method="block.splsda", nboot=100, folds=5, nrepeat=20, n_cores=1, ncomp=2)
```

Feature selection

X1 component 1: same variables selected

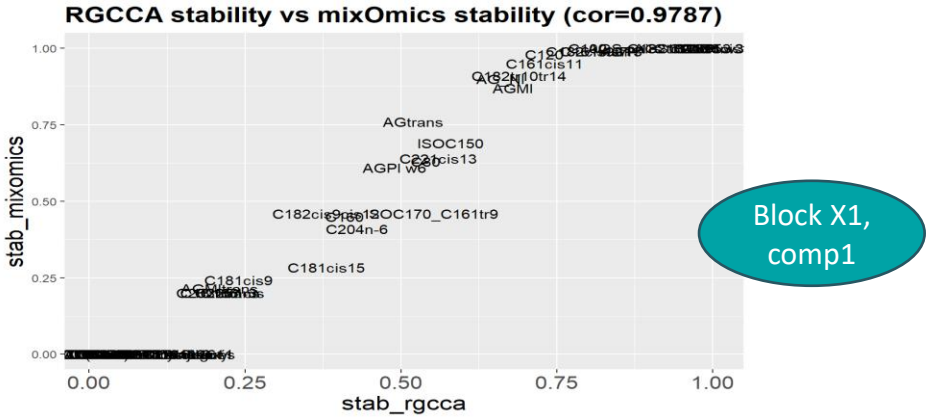
Composante 2, blocX1

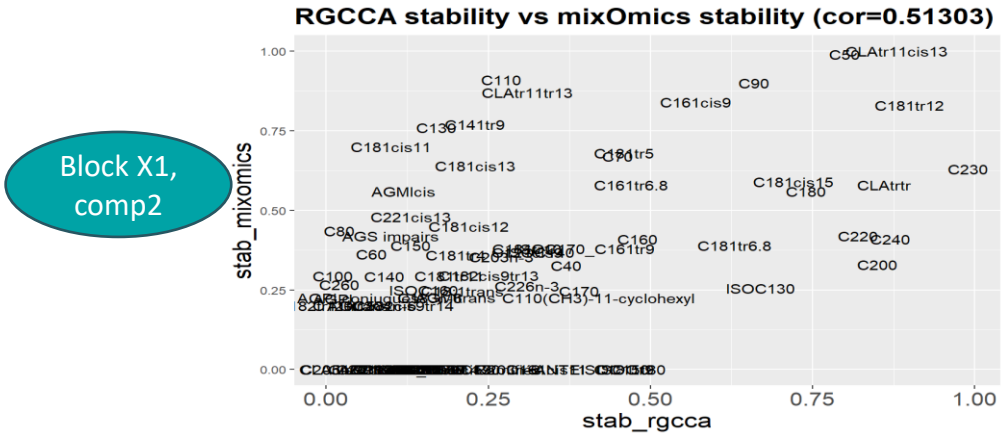
	C160	
	C161cis9	
	C161tr6.8	
	C180	
	C181cis15	
AGMIcis	C181tr12	AGMItrans
C110	C220	ANTEISOC150
C130	C230	C121cis9
C141tr9	C240	C170
C181cis11	C50	C18:1trans
C181cis13	C70	C181tr6.8
C181tr11	C90	C181tr9
C181tr4	CLAtr11cis13	C200
C181tr5	CLAtrtr	C40
C221cis13	ISOC140	ISOC130
CLAtr11tr13	ISOC170_C161tr9	ISOC180
MIXOMICS 11	BOTH 16	RGCCA 11

Same trend with the block X2

Feature stabilityselection

mixOmics: based on cross-validation ➛ % of selection across all folds
RGCCA: bootstrap resampling procedure, based on VIPs (higher VIPs averaged over the different models are kept) ➛ reworked to compare with mixOmics





Conclusion

Comparison

- Differences often seen from component 2 (not the same deflation method?)
- Some difference observed on feature selection
- Difference on stability selection, particularly on multiblock sPLSDA
- **Results dependent on datasets**

Difficulties encountered

- comparison of certain functions when they do not operate on the same principle (ex: selection, stability)
- Performance ➛ rewriting functions for better comparison

Perspectives

- Report some observed errors (and requirements) to authors of mixOmics & RGCCA
- Website (under construction):
  - Project description
  - Reports on analyses of the 3 datasets +/- modified using all methods
  - Comparison of package functions and conclusion
- Maintain the community by testing/comparing other functions from other R packages

Thank you for your attention.

11 statisticians and (bio)informaticians of 5 CATIs

Marion BRANDOLINI-BUNLON

Caroline PELTIER

Gwendal CUEFF

Marie TREMBLAY FRANCO

Virginie ROSSARD

Eric Latrille

Isabelle SANCHEZ

Alyssa IMBERT

Nadia Bessoltane-Bentahar

Sébastien THEIL

Elise MAIGNE

6 supporters

Jean-Claude BOULET

Rémi SERVIEN

Julie AUBERT

Christelle HENNEQUET

Magali BERLAND

Jean-Michel ROGER

CATI eMPrEInTE

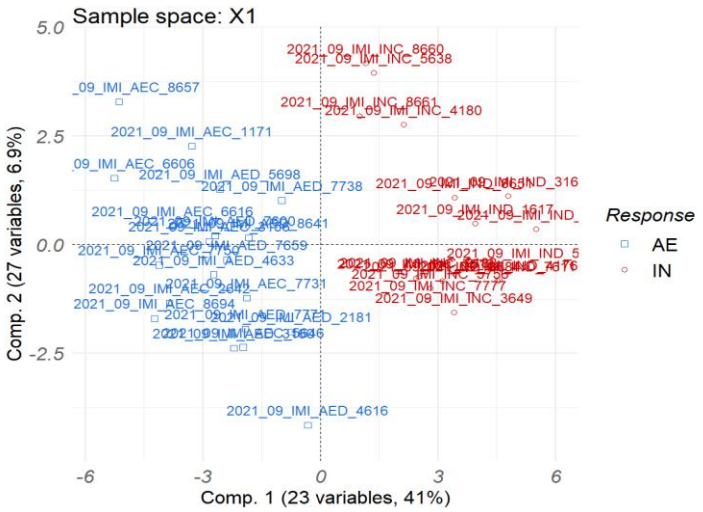
CATI Codex

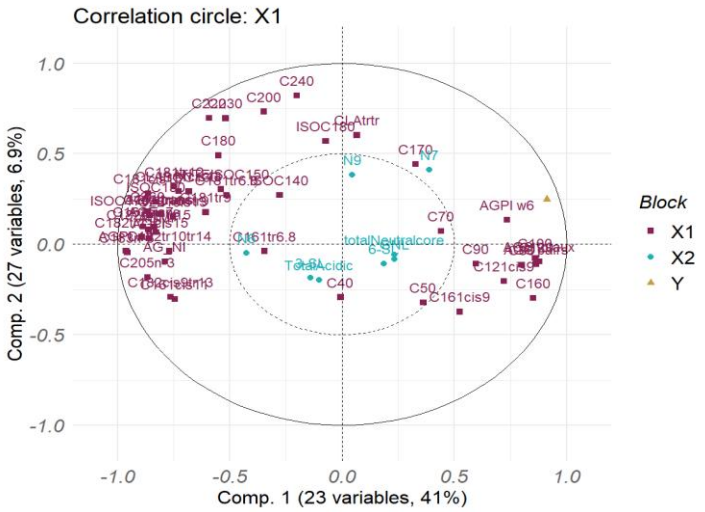
CATI SysMics

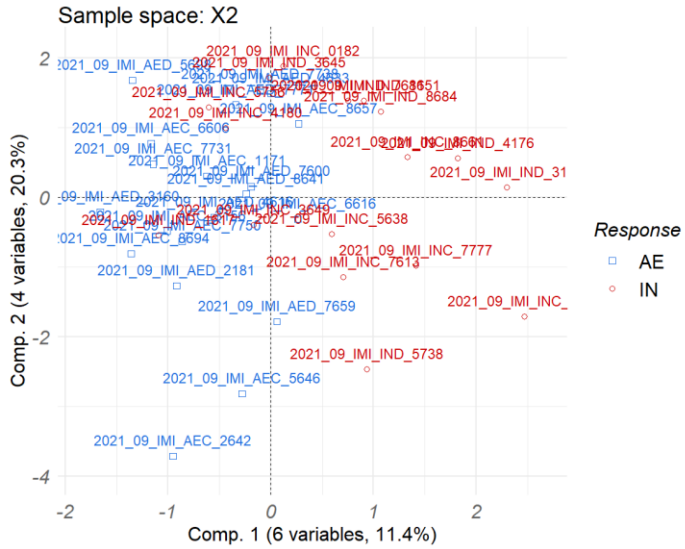
Cati BOOM

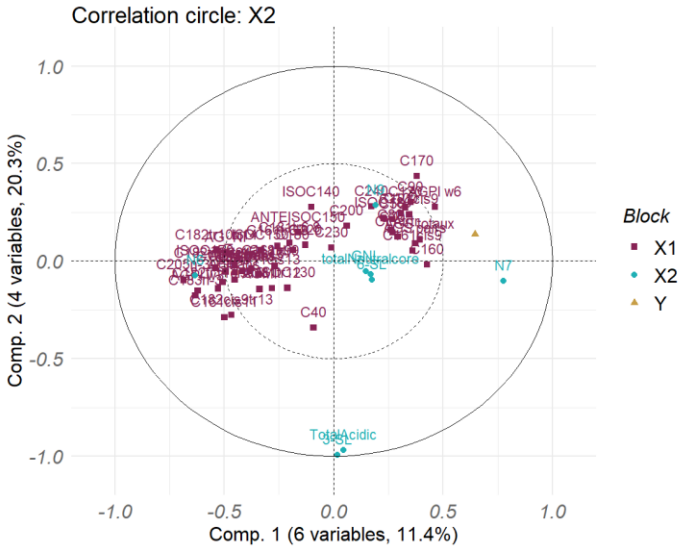
B4B

RGCCA, multiblock sPLSDA, plotIndiv and correlation circle









mixOmics: multiblock sPLSDA, plotIndiv and correlation circle