Computational optimisation for mixOmics, the R package dedicated to 'omics' data integration

F. Bartolo[a] and **S. Déjean**[a] and **B. Gautier**[b] and **I. González**[a] and **F. Rohart**[c] and **K. Lê Cao**[b]

[a]Institut de Mathématiques, UMR5219 CNRS
A-Université Toulouse 3 Paul Sabatier
A-118, route de Narbonne, 31062, Toulouse, Cedex 9, France
mixomics-devel@math.univ-toulouse.fr

[b]The University of Queensland Diamantina Institute
B-QLD 4102 Woolloongabba, Brisbane, Australia
mixomics-devel@math.univ-toulouse.fr

[c]Australian Institute for Bioengineering and Nanotechnology
The University of Queensland
QLD St Lucia, Brisbane, Australia
mixomics-devel@math.univ-toulouse.fr

Mots clefs : integrative analysis, Generalised and Sparse Canonical Correlation Analysis, computational optimisation, memory management, parallel computation.

We have recently implemented novel methodologies in mixOmics to integrate several 'omics data sets simulaneously. These novel developments require intensive computations which can be eased through efficient optimisation and memory management.
mixOmics is an R package dedicated to the exploration and integration of 'omics datasets. Its first release to the CRAN in 2009 proposed statistical methodologies to integrate two 'omics data sets. Since then numerous methodologies and variants have been implemented, and amongst those Generalised and Sparse Canonical Correlation Analysis (GCCA) to integrate more than two datasets. These latest developments require effective computational optimisation and memory management. Indeed, some functions could use one CPU for a full on a standard desk computer on large biological biological studies.
We investigated three ways to address these computational challenges via 1/ sequential optimisation (pre-compilation of functions) 2/ parallel computation (using the parallel package) and 3/ enhancement of memory management (using the bigmemory package). Our first results obtained on a micro-benchmark showed computation times divided by at least 4.
The poster will present a global overview of the computational improvements made with these enhancements on real biological datasets.

Références

[1]Tenenhaus A., Phillipe C., Guillemot V., Lê Cao K-A. , Grill J. , Frouin V. (2014), Variable selection for generalized canonical correlation analysis, *Biostatistics, doi: 10.1093/biostatistics.* PMID: 24550197. .
[2] González I., Lê Cao K.-A., Davis, M.D. and Déjean S. (2013) Insightful graphical outputs to explore relationships between two omics data sets. *BioData Mining* 5:19
[3] Yao F., Coquery J., Lê Cao K.-A. (2012) Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets, *BMC Bioinformatics* 13:24.
[4] Lê Cao K.-A., Boitard S. and Besse P. (2011) Sparse PLS Discriminant Analysis: biologically

relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, 22:253.

[5] Lê Cao K.-A., González I. and Déjean S. (2009) integrOmics: an R package to unravel relationships between two omics data sets. *Bioinformatics*, 25(21):2855-2856.