Extensions of CCA and PLS to unravel relationships between two data sets

Sébastien Déjean[1*], Ignacio González[2], Kim-Anh Lê Cao [2]

1. Institut de Mathématiques de Toulouse, UMR 5219 Université de Toulouse et CNRS
2. Plateforme Biopuces, Genopôle Toulouse Midi-Pyrénées, Institut National des Sciences Appliquées
3. ARC Centre of Excellence in Bioinformatics, Institute for Molecular Bioscience, University of Queensland, Australia

* Contact author: sebastien.dejean@math.univ-toulouse.fr

Keywords: regularization, sparse methods, graphical display, gene expression data.

In the context of systems biology and in post-genomic studies, it becomes usual to analyze simultaneously transcriptomics, proteomics and/or metabolomics data. Several approaches have been proposed to understand and to highlight the mutual interactions between two different data sets. The main challenge of the proposed methodologies relies on their ability to handle very large data sets, where the number of variables is much greater than the number of observations. Lê Cao *et al.* (2008) proposed a sparse PLS method in a canonical correlation framework which includes variables selection while integrating data. González *et al.* (2009) developped a regularized version of Canonical Correlation Analysis to deal with singular matrices occurring in the classical CCA.

On the basis of the CCA package (González *et al.*, 2008), we developped the package `CCAsPLS` to implement these two approaches (Fig. 1).

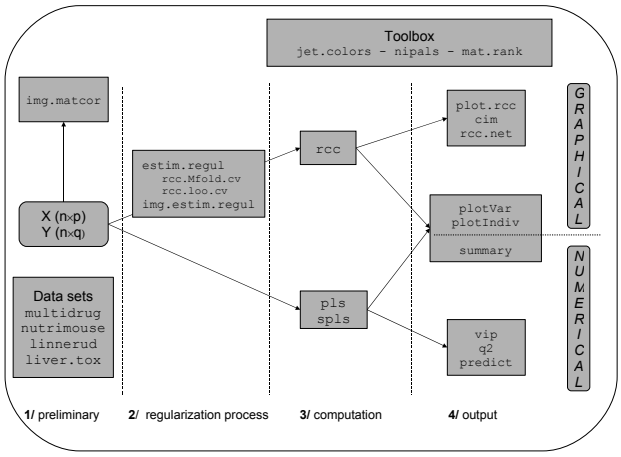

Figure 1: Schematic view of the CCAsPLS package.

The interpretation of the results is made easier with the use of various graphical display such as correlation loading plots as in factorial analysis (previously implemented in the CCA package). CCAsPLS also proposes heat maps and networks to better visualize the relationships between variables from the two data sets. These graphs are often used when dealing with high-throughput biological data.

References

I. González, S. Déjean, P. Martin, A. Baccini (2008). CCA: An R package to extend canonical correlation analysis. *Journal of Statistical Software*, 23(12).

K-A. Lê Cao, D. Rossouw, C. Robert-Granié, P. Besse (2008). A sparse PLS for variable selection when integrating Omics data. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 35.

I. González, S. Déjean, P.G.P. Martin, O. Gonalves, P. Besse, A. Baccini (2009). Highlighting Relationships Between Heteregeneous Biological Data Through Graphical Displays Based On Regularized Canonical Correlation Analysis. *Journal of Biological Systems*, to appear.