A Unified Regularized Group PLS Algorithm Scalable to Big Data

Benoît Liquet[a] and **Pierre Lafaye de Micheaux**[b]

[a]Universite de Pau et Pays de L'Adour
Laboratoire de Mathématiqes et de leurs Applications
UMR CNRS 5142
benoit.liquet@univ-pau.fr

[b]CREST, ENSAI,
Campus de Ker-Lannt
Rue Blaise Pascal, BP 37203, 35172 Bruz cedex

Mots clefs : Omics, Multivariate, PLS, Sparsity, Big Data.

The association between two blocks of 'omics' data brings challenging issues in computational biology due to their size and complexity. Here, we focus on a class of multivariate statistical methods called partial least square (PLS). Four classical versions of PLS coexist in the litterature. Moreover, a sparse version of PLS (sPLS) enables integration of two data-sets while simultaneously selecting the contributing variables. However, these methods do not take into account the important structural or group effects due to the relationship between markers among biological pathways. Hence, considering the predefined groups of markers (e.g., gene-sets) could improve the relevance and the efficacy of PLS approaches.

We propose a unified algorithm to implement the four classical versions of PLS, with an extension to group PLS (gPLS) and sparse group PLS (sgPLS). Our algorithm enables to study the relationship between two different types of omics data (e.g., SNP and gene expression) or between an omics dataset and multivariate phenotypes (e.g., cytokine secretion). We demonstrate the good performance of gPLS and sgPLS compared to the sPLS in the context of grouped data. Then, these methods are compared through an HIV therapeutic vaccine trial. Our approaches provide parsimonious models to reveal the relationship between gene abundance and the immunological response to the vaccine. We also propose a simple modification of this algorithm that makes it usable for big data. This is illustrated through simulated data.

The algorithm is implemented in an R package called bigsgPLS that will be made available on the CRAN.

Références

[1] Liquet B., Lafaye de Micheaux P., Hejblum B., Thiébaut R. (2016). Group and sparse group partial least square approaches applied in genomics context. *Bioinformatics* **32**(1), 35-42.