Group and Sparse Group Partial Least Square Approaches Applied in Genomics Context

Benoît Liquet[a] and **Pierre Lafaye de Micheaux**[b] and **Boris P. Hejblum**[c,d] and **Rodolphe Thiébaut**[c,d]

[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

[c] Inria, SISTM, Talence
Bordeaux University

[d] Vaccine Research Institute
Creteil, France

Mots clefs : Omics, Multivariate, PLS, Sparsity.

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). Sparse version of PLS (sPLS) operate 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., genesets), this could improve the relevance and the efficacy of the PLS approach.

We propose two PLS extensions called 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.

The approach is implemented in a comprehensive R package called sgPLS 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.