A new tool for multi-block PLS discriminant analysis: application to metabolomic data in systems epidemiology

Marion Brandolini-Bunlon, Mélanie Pétéra, Pierrette Gaudreau, Blandine Comte, Stéphanie Bougeard, Estelle Pujos-Guillot

A new tool for multi-block PLS discriminant analysis: application to metabolomic data in systems epidemiology

Marion Brandolini-Bunlon[1] Mélanie Pétéra[2] Pierrette Gaudreau[3] Blandine Comte[4] Stéphanie Bougeard[5] Estelle Pujos-Guillot[6]

[1] Université Clermont Auvergne, INRA, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, F-63000 Clermont-Ferrand, France, marion.brandolini-bunlon@inra.fr

[2] Université Clermont Auvergne, INRA, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, F-63000 Clermont-Ferrand, France, melanie.petera@inra.fr

[3] Centre de Recherche du Centre hospitalier de l'Université de Montréal & Département de médecine, Université de Montréal, Montréal, Canada , pierrette.gaudreau@umontreal.ca

[4] Université Clermont Auvergne, INRA, UNH, F-63000 Clermont-Ferrand, France, blandine.comte@inra.fr

[5] Anses, BP53, Technopole Saint Brieuc Armor, 22440, Ploufragan, France, stephanie.bougeard@anses.fr

[6] Université Clermont Auvergne, INRA, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, F-63000 Clermont-Ferrand, France, estelle.pujos-guillot@inra.fr

Keywords: Multi-block PLS discriminant analysis, metabolomics, epidemiology

1 Introduction

Metabolomics is a powerful phenotyping tool in nutrition and health research. Massive and complex data are generated, and consequently dedicated treatments to enrich our knowledge of biological systems are needed. To deeper investigate relations between environmental factors, phenotypes and metabolism, statistical analyses performed separately on metabolomic datasets are often complemented by associations with metadata (anthropometric, clinical, nutritional, and physical activity data…). Another relevant strategy, in order to discriminate observation groups, is to perform a multi-block partial least squares discriminant analysis (MBPLSDA) that simultaneously processes data available from different sources. This method allows determining the importance of variables and variable blocks in discriminating groups of subjects, taking into account data structure in thematic blocks. In order to propose a full open-source standalone tool, an R package was developed, allowing all steps of MBPLSDA analysis for the joint analysis of metabolomic and epidemiological data

2 Theory

A standard PLSDA model is built to explain a block of variables ("Y-block"), corresponding to a matrix of indicators of categories related to observation groups, by K explanatory variable blocks ("Xk-blocks", with k=1,..,K). The algorithm [1] is based on maximizing a covariance criterion between the components from Xk-blocks, and the Y-block, under the constraint that the variable weight vectors are normalized to one. A global component is constructed using the weighted sum of the Xk-block components based on their (normalized) covariance with the Y-block components. The cumulative importance of each Xk-block (BIPcum: Cumulated Block Importance in the Projection) and each explanatory variable (VIPcum: Cumulated Variable Importance in the Projection) in model with all the different numbers of components are calculated from the global components, the variable weights on the components, and the covariances between the components

of the Xk and Y-blocks. In addition, the regression coefficients of Y-block on the explanatory variables, summed up with global components, are estimated. The model parameter values are therefore variable weights on the components, regression coefficients of Y-block on the explanatory variables, VIPcum values, and BIPcum values.

3 Material and methods

In this work, we propose a R package (named packMBPLSDA), based on the mbpls function of the ade4 R package [2], and enriched with different functionalities, including some dedicated to discriminant analysis. Indicators are provided to help to determine the optimal number of components, to check the MBPLSDA model validity, and to evaluate the variability of its parameters and predictions. To illustrate the potential of this package and the associated procedure, MBPLSDA was applied to a real case study involving metabolomics, nutritional and clinical data obtained from a case-control study (n=123) within a human cohort [3]. The block of metabolomics data was obtained from the analysis of serum samples using an untargeted approach. To study the impact of filtering the metabolomic variables beforehand, metabolomic data were either (i) only pre-processed (1656 features), (ii) pre-processed and decorrelated (1091 features), or (iii) pre-processed, decorrelated and filtered on a criterion independent of the outcome (388 features). The clinical data block included 18 quantitative variables and the nutritional data block included 87 quantitative variables. Then, MBPLSDA was compared to PLSDA on concatenated data.

4 Results and discussion

The availability of the different functionalities in a single R package allowed optimizing parameters for an efficient joint analysis of metabolomics and epidemiological data, in order to obtain new insights into multidimensional phenotypes. In our study, the cross-validated prediction error rates were used to 1) select the models whose number of components induces the best predictions without overfitting, and 2) verify the quality of the models. Using the comparison indicators provided, we found that MBPLSDA was improved by a relevant variable filtration within blocks. Compared to the application of PLSDA on the concatenated explanatory dataset, we highlighted that MBPLSDA, considering the blocks characteristics, allowed avoiding the predominance of the most important block and provided a more integrative subset of important variables.

5 Conclusion

MBPLSDA, with an appropriate data scaling and block weighting, is a discriminant method suited for the joint analysis of structured data in blocks of heterogeneous sizes, as metabolomics and epidemiological data. Based on our study, we highlighted that this approach is more relevant than a standard PLS-discriminant analysis method to obtain an integrative and global insight of a biological phenomenon. The benefit of the packMBPLSDA R package is to allow easy model evaluation, to provide indicators for model comparison and to facilitate the adaptation of statistical analysis to the experimental design. In particular, in our study the impact of filtering the metabolomic variables beforehand allowed to improve discriminant models.

6 References

[1] Wold, S. (1984). Three PLS algorithms according to SW. In *Report from the symposium MULTDAST (multivariate data analysis in science and technology)* (pp. 26–30). Umeå, Sweden.

[2] Bougeard, S., & Dray, S. (2018). Supervised multiblock analysis in R with the ade4 package. *Journal of Statistical Software*, *86*, 1–18.

[3] Brandolini-Bunlon, M., Pétéra, M., Gaudreau, P., Comte, B., Bougeard, S., Pujos-Guillot, E. (2019). Multi-block PLS discriminant analysis for the joint analysis of metabolomic and epidemiological data. *Metabolomics*, *15*, 134.