Metabolomics (2019)15:134
https://doi.org/10.1007/s11306-019-1598-y

SOFTWARE/DATABASE

Multi-block PLS discriminant analysis for the joint analysis of metabolomic and epidemiological data

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

Received: 11 June 2019 / Accepted: 25 September 2019

Abstract

Introduction Metabolomics is a powerful phenotyping tool in nutrition and health research, generating complex data that need dedicated treatments to enrich knowledge of biological systems. In particular, to investigate relations between environmental factors, phenotypes and metabolism, discriminant statistical analyses are generally performed separately on metabolomic datasets, complemented by associations with metadata. Another relevant strategy is to simultaneously analyse thematic data blocks by a multi-block partial least squares discriminant analysis (MBPLSDA) allowing determining the importance of variables and blocks in discriminating groups of subjects, taking into account data structure.

Objective The present objective was to develop a full open-source standalone tool, allowing all steps of MBPLSDA for the joint analysis of metabolomic and epidemiological data.

Methods This tool was based on the *mbpls* function of the ade4 R package, enriched with functionalities, including some dedicated to discriminant analysis. Provided indicators help to determine the optimal number of components, to check the MBPLSDA model validity, and to evaluate the variability of its parameters and predictions.

Results To illustrate the potential of this tool, MBPLSDA was applied to a real case study involving metabolomics, nutritional and clinical data from a human cohort. The availability of different functionalities in a single R package allowed optimizing parameters for an efficient joint analysis of metabolomics and epidemiological data to obtain new insights into multidimensional phenotypes.

Conclusion In particular, we highlighted the impact of filtering the metabolomic variables beforehand, and the relevance of a MBPLSDA approach in comparison to a standard PLS discriminant analysis method.

Keywords Multiblock PLS discriminant analysis · Metabolomics · Multi-block · Discrimination · Epidemiology

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11306-019-1598-y) contains supplementary material, which is available to authorized users.

✉ Marion Brandolini-Bunlon
marion.brandolini-bunlon@inra.fr

[1] Université Clermont Auvergne, INRA, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, 63000 Clermont-Ferrand, France

[2] Centre de Recherche du Centre hospitalier de l'Université de Montréal, Montréal, Canada

[3] Département de médecine, Université de Montréal, Montréal, Canada

[4] Université Clermont Auvergne, INRA, UNH, 63000 Clermont-Ferrand, France

[5] Anses, BP53, Technopole Saint Brieuc Armor, 22440 Ploufragan, France

1 Introduction

Untargeted metabolomics has been recognized as a powerful phenotyping tool to better understand the biological mechanisms involved in the physiopathological processes and identify biomarkers of metabolic status (Ackermann et al. 2006; Ramautar et al. 2013). By giving an integrated vision of these phenomena, such an approach allows characterizing the impact of key environmental factors on human health, and consequently generates complex data that need dedicated treatments to extract meaningful information. The common strategy currently consists in performing univariate and multivariate statistics to reveal variables of interest that will be further used for biological interpretation. The major specificity of metabolomics data is the large number of variables (ions) compared to the number of samples, as

Published online: 03 October 2019

well as their high degree of correlation related to both analytical redundancy and biological relationships. Moreover, when using mass spectrometry profiling, relative intensities are influenced by the high diversity of physico-chemical properties and abundances of small molecules, as well as analytical drifts. These data characteristics require careful preparation of metabolomic data and selection of appropriate discriminant statistical analysis method. Partial least squares-discriminant analysis (PLSDA) is one of the most effective multivariate tool used in metabolomics for variable selection and classification, because of its ability to analyze highly collinear and noisy data, but also because of its large availability in different software and packages (Barker and Rayens 2003; Gromski et al. 2015). However, to investigate in depth relationships through which environmental factors and phenotypes are linked to metabolism, this analysis performed on the main outcome is often complemented by associations with metadata (anthropometric and clinical data, parameters regarding nutrition and physical activity…).

To integrate these data organized into meaningful blocks, the most intuitive strategy, with the objective of discriminating groups of subjects, is to create a model per block of variables, and then either analyze the correlations or build a PLSDA model from the most discriminant variables (Saccenti et al. 2014). However, in metabolomics, the study of correlations between metabolites does not lead to a straightforward interpretation in terms of the underlying biochemical pathways partially because of the ubiquity of some metabolites (Steuer 2006). In addition, this strategy (i) does not assess the importance of blocks in group discrimination, (ii) can discard a block of variables at the first step if it does not provide a valid model. An alternative intuitive strategy could consist in performing PLSDA on concatenated data blocks. In this case, due to differences in the characteristics of the variables and data blocks, it is very likely that variables found to be important in the discrimination would mainly come from the largest metabolomics data block. On the opposite, variables from smaller blocks could not come out, in spite of a potential role known in the studied biological system. In contrast, the multi-block PLS discriminant analysis method (MBPLSDA) simultaneously analyses data from all available blocks on the same observations, and allows determining the importance of variables and variable blocks in discriminating groups of subjects, taking into account blocks characteristics, and in particular their dimensions and their covariances with variables indicating groups to be explained. The blocks' pre-treatment step allows giving them similar weight in the analysis. This method is consequently of particular interest to integrate untargeted metabolomic data consisting in thousands of molecular features with clinical and/or nutritional data, where the number of variables per block is often less than a few hundred. However, this method requires several steps to be properly applied, and there is actually no open access standalone tool for a full implementation; while there are some to perform multi-block PLS analyses whose variables to be explained are quantitative (Bougeard and Dray 2018), or extensions of canonical correlation analysis for discriminating purposes (Gunther et al. 2014; Rohart et al. 2017; Singh et al. 2016).

In order to propose a full open-source standalone tool, the present objective was to develop an R package allowing all steps of MBPLSDA for the joint analysis of metabolomic and epidemiological data. This tool (the *packMBPLSDA* R package) was based on the *mbpls* function of the ade4 R package, enriched with different functionalities, including some dedicated to discriminant analysis, as well as indicators and visualization of results.

The different steps required to conduct a relevant MBPLSDA were described along with the corresponding functions provided in the R package developed. Its use was illustrated for the joint analysis of metabolomic, clinical and nutritional data from a human cohort study. A first aim was to investigate the interest of filtering the metabolomic variables beforehand using a criterion independent of the groups to be discriminated. A second aim was to highlight the relevance of the MBPLSDA method with block weighting by their inertia in comparison to a standard PLSDA method.

2 Material and methods

All steps of MBPLSDA, and the main inputs and outputs of the developed programs are shown in Figs. 1 and S1. After the calculation of a "basal model" (where the number of components is arbitrarily set by the user), the first step is dedicated to the selection of the optimal model, generally the one with a number of components that minimizes the cross-validated prediction error, without overfitting. Then, the validity of this model is evaluated in a second step. If the model is validated, three steps can be performed: evaluation of parameters values and predicted observations categories, evaluation of parameter value variabilities, evaluation of prediction variabilities.

2.1 MBPLSDA algorithm

2.1.1 Components and parameters

A standard PLSDA regression model is built in such a way as 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 $k = 1,\ldots,K$). The calculation of the components and parameters of models are performed in all steps of the present tool by an adaptation of the *mbpls* function from

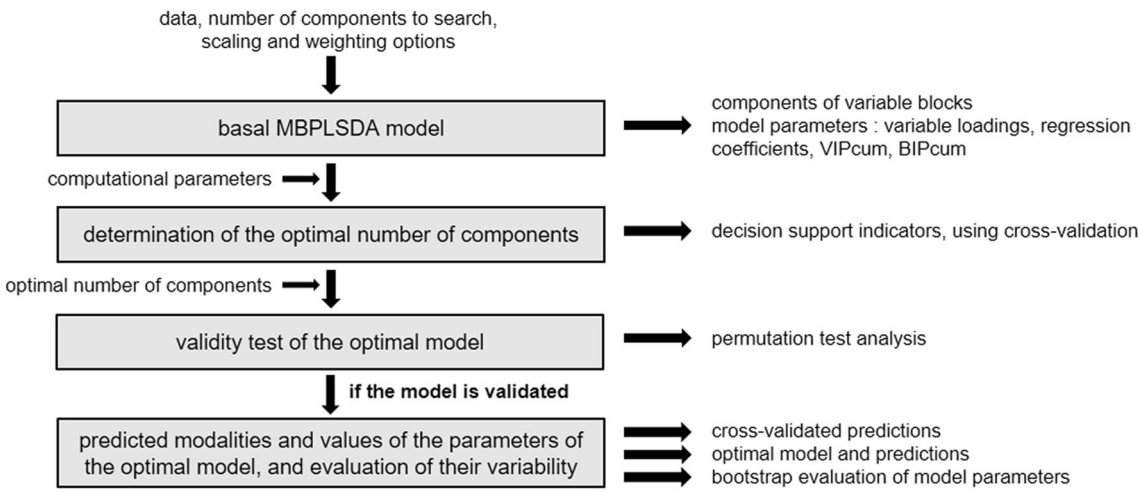

Fig. 1 Recommended steps for the application of the MBPLSDA method, and main inputs and outputs of the developed programs

the ade4 R package (Bougeard and Dray 2018; Dray and Dufour 2007) (Table S1). It takes as input: the Y-block, the Xk-blocks, the chosen variable standardization and block weighting (for a weighting by their inertia), and the number of components to search, limited to the maximum rank of the analysis. The algorithm, proposed by Wold (1984), 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 component. The cumulative importance of each Xk-block (BIPcum for Cumulated Block Importance in the Projection) and each explanatory variable (VIPcum for Cumulated Variable Importance in the Projection) in model with all the different numbers of components are calculated from the global components, the weights of the variables on the components, and the covariances between the components of the Xk and Y-blocks. They are expressed as a percentage in relation to the number of blocks of explanatory variables. In addition, the regression coefficients of Y-block on the explanatory variables, summed up with the global components, are estimated. The values of the model parameters are therefore weights of the variables on the components, regression coefficients of Y-block on the explanatory variables, VIPcum values, and BIPcum values.

Mathematically, MBPLSDA is equivalent to PLSDA on concatenated data after block weighting (Westerhuis et al. 1998). Therefore, PLSDA on concatenated data is equivalent to MBPLSDA method without prior weighting of the blocks (same components, weights of the variables on the components and regression coefficients). The weighting of a block modifies only the VIPcum values of its variables and consequently their ranks in the global model (Bougeard et al. 2011).

2.1.2 Prediction methods

The calculation of the Y-prediction of subject categories are performed in the present tool from the returned values of the model parameters and the subjects' scores on the components. The assignment of observations to categories can be done by different methods (Saporta 2006). In the present tool, for each Y-block variable, an observation takes either the category for which it has the highest predicted value, or the category for which it has a value greater than a fixed threshold, or the closest category when considering the categories' centers of gravity in the component space.

2.2 Detailed steps of MBPLSDA implementation

2.2.1 Selection of the optimal model

The determination of the number of components to be included in the MBPLSDA model is based on a repeated twofold cross-validation strategy known as "Monte Carlo cross-validation" (Kuhn and Johnson 2013). At each repetition, 2/3 of the observations are set in a calibration dataset and 1/3 in a validation one (Stone 1974). The number of repetitions is set by the user (at least thirty, usually about a 100). At each repetition of the cross-validation procedure, a MBPLSDA model is built only with calibration data. The validation data are weighted, centered and possibly reduced by respectively taking into account the inertia, means and standard deviations of the calibration data. Then, for a number of components ranging from 1 to the number of components of the basal model, the predicted categories of each variable for all observations are determined. The confusion matrices, classification error rates, and values of areas under receiver operating characteristic curves (AUC) in the case of binary Y-block variables, are then calculated, separately for

calibration and validation data, by Y-block variable category and taking into account all Y-block variables. At the end of the repetitions, it could be expected that the optimal number of components corresponds to one of the lowest overall average classification error rate or AUC obtained on the validation data. The graphs of error rates or AUC variations help selecting the optimal conditions while evaluating potential overfitting effect.

2.2.2 Validity test of the optimal model

The validity of the model is checked by a permutation test (Westerhuis et al. 2008). The user sets how many permutated Y-blocks, line permutations, and cross-validation repetitions should be computed. At the end of the MBPLSDA cross-validations with the permutated Y-blocks, the matrices of confusion, average errors in the classification, and average AUC for binary Y-block variables, are computed for the calibration and validation datasets. Then, if the number of line permutations of the Y-block has been set, a Student test is performed to compare the mean cross-validated classification errors or AUC on models with permutated Y-blocks, with the values obtained for the original Y-block, according to the same procedure. On an indicative basis, the cross-validated error rates, or AUC, based on the percentage of modified values in permutated Y-blocks, is plotted with a regression line (similarly to other commonly used software, as SIMCA® Umetrics AB), whose coefficient is compared to 0 by a Student test. The model is considered valid if the error rates increase with the percentage of modified Y-block values.

2.2.3 Predictions and parameter variabilities of the optimal model

Once the validity of the model is confirmed, the predicted categories for the observations are determined and graphically plotted by scatterplots. Then, the cross-validated predicted categories of the observations are calculated with the same cross-validation procedure as in the previous steps, with a set number of repetitions (at least ninety). After calculations, for each variable category, the most frequent value of the indicator is given for each observation, with the probability that it is equal to 1, and the associated confidence interval. These values are displayed with scatterplots.

Similarly, the values of the parameters of the optimal model—BIPcum, VIPcum, weight, regression coefficients—are determined and plotted. The variability of these parameters is commonly evaluated by a bootstrap method (Efron and Tibshirani 1994), with a set number of repetitions (at least thirty, a minimum of a hundred being recommended). Then for each parameter, the mean, median, standard deviation, 95% confidence interval and 0.025 and 0.975 percentiles are calculated. The graphical representation of all these values, or a chosen percentage of the most important values, is produced under the form of diagrams and histograms to support the model interpretation.

The tool is implemented under the R software [version 3.6.0, (R Development Core Team 2019)], with the possibility to parallelize the calculations of cross-validation and bootstrap repetitions.

2.3 Experimental

2.3.1 Variables and objective

The MBPLSDA method was applied to integrate metabolomic, nutritional and clinical datasets comparing case and control subjects from a human cohort to illustrate the use of the proposed tool and the associated procedure. This case–control study on Metabolic Syndrome (MetS) was designed within the Quebec Longitudinal Study on Nutrition and Successful Aging (NuAge) (Gaudreau et al. 2007). Included men were 61 cases and 62 controls of similar age (68–82 y.o.). This study has been approved by the Research Ethics Board of both the Geriatric University Institutes of Montreal and Sherbrooke. All the volunteers gave written and informed consent.

The Y-block contained the indicator variable of the subjects' status ("case" or "control"). The three explanatory data blocks—Xk-blocks—were derived from the subjects' monitoring, which includes metabolomics (X1) of serum samples collected at recruitment 2003–2005 and 3 years later and available in the NuAge biobank, and data collection through questionnaires and medical visits extracted from the NuAge database (supplemental material 1). The clinical data block (X2) included 18 quantitative variables, indirectly related to MetS (clinical data and physical activity scores). The nutritional data block (X3) included 87 quantitative variables, which corresponded to nutrient intakes or nutritional scores estimated from food surveys or questionnaires.

The block of metabolomics data (X1) was obtained from the analysis of serum samples using a mass spectrometry-based untargeted approach (Pujos-Guillot et al. 2017). Data were processed using the XCMS R-package (Tautenhahn et al. 2008) under a Galaxy web-based platform [Worflow-4Metabolomics, (Giacomoni et al. 2015)] to yield a data matrix containing retention times, accurate masses and processed peak intensities. This step included noise filtering, automatic peak detection and chromatographic alignment allowing the appropriate comparison of multiple samples by further processing methods. It constituted the first dataset of 1656 variables ("dataset 1", Fig. 2). Following these usual preprocessings steps, an analytical redundancy filter (removal of fragments, adducts and isotopes), and a more stringent background filter provided a second dataset of 1091

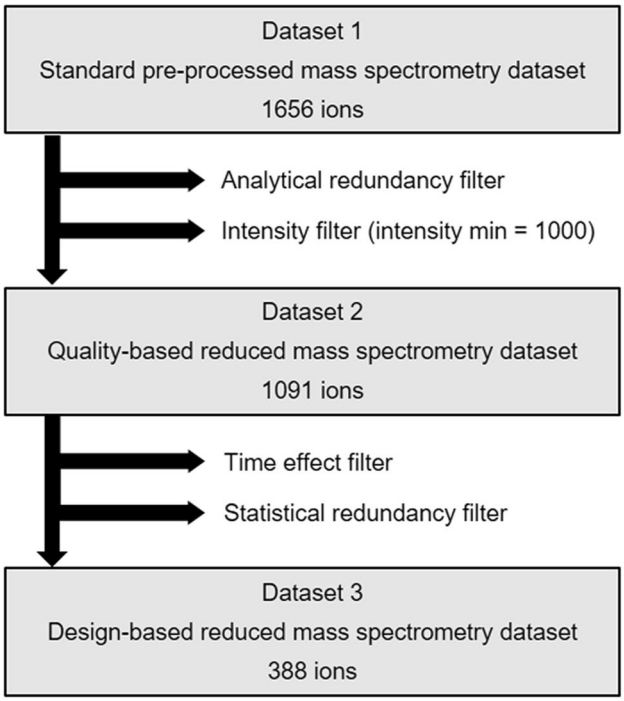

Fig. 2 Filtering of metabolomic data to obtain the three datasets

variables ("dataset 2"). Finally, following mixed models including MetS status and time cofactor, ions independent of time effect were selected. Then, this dataset was reduced by selecting the highest intensity features among groups of highly correlated ions ($r^2 > 0.8$), resulting in a third data set of 388 variables ("dataset 3"). Filtering details are available in supplemental material 1.

To illustrate the present tool, only baseline data were considered. Due to missing values in clinical and nutritional blocks, we chose to limit the data to the 118 individuals (58 cases, 60 controls) with complete observations at baseline.

2.3.2 Models performed and settings

To investigate the interest of filtering the metabolomic variables beforehand using a criterion independent of the groups to be discriminated, the MBPLSDA method was applied on several combinations of datasets for comparison purpose. Each model contains the clinical data block (X2) and the nutritional one (X3), completed with one metabolomic block (X1). This metabolomic block could be either dataset 1, 2 or 3. The models obtained are further refered as "model 1", "model 2", and "model 3", respectively. Data were centered and reduced, to cope with the differences in order of magnitude and variable units within the blocks. The blocks were weighted by their total inertia.

The MBPLSDA method was used with the most common assignment method (category of the higher predicted value). Although the Y-block variable was binary (and thus the use of AUC was possible), the determination of the optimal component number and the validity test of the model were based on the overall cross-validated classification error rates. To determine the optimal component number, models with 1 to 6 components were tested. For the optimal model validation by permutation test, 40 permutations of two observations were performed in 100 permutated Y-blocks. The variability of the classification error rates and parameters were respectively estimated with 30 cross-validation or 100 bootstrap repetitions. Concerning the variability of the predicted categories for observations, 120 cross-validation repetitions were performed. Finally, for better readability, in the present example, the graphical outputs of the values and confidence intervals of weights, coefficients, and VIPcum were limited to the 25 explanatory variables having the highest average absolute values. An example of function call, illustrated with "model 3", is available in Table S2.

To highlight the relevance of the MBPLSDA approach in comparison to a standard PLSDA method, we selected the best model configuration previously obtained, based on the cross-validated classification error rate. The model was compared to the optimal one obtained with the application of a PLSDA method on the corresponding blocks concatened into one (all explanatory variables in the same block) with the same settings. The comparison was based on the cross-validated classification error rate, the validity test results, the cross-validated predictions, and the cross-validated values of the models parameters.

3 Results and discussion

3.1 Steps of MBPLSDA and interest of metabolomic variable filtering

3.1.1 Selection of the optimal MBPLSDA models

In the three MBPLSDA models produced, the cross-validated classification error rates on the validation data were found to be stable according to the number of components. The example of model 3 is shown in Fig. 3. Similar results were obtained with models 1 and 2 (Fig. S2). The cross-validated error rates obtained on the calibration data decreased rapidly with the number of components to almost zero from 4 components; the optimal number of components of the three models was found to be one. In this condition, the cross-validated error rates on the validation data were 0.344 ± 0.081, 0.335 ± 0.078, and 0.315 ± 0.078, respectively for the three models. The error rate therefore decreases, but not significantly, with the size reduction of the metabolomic dataset.

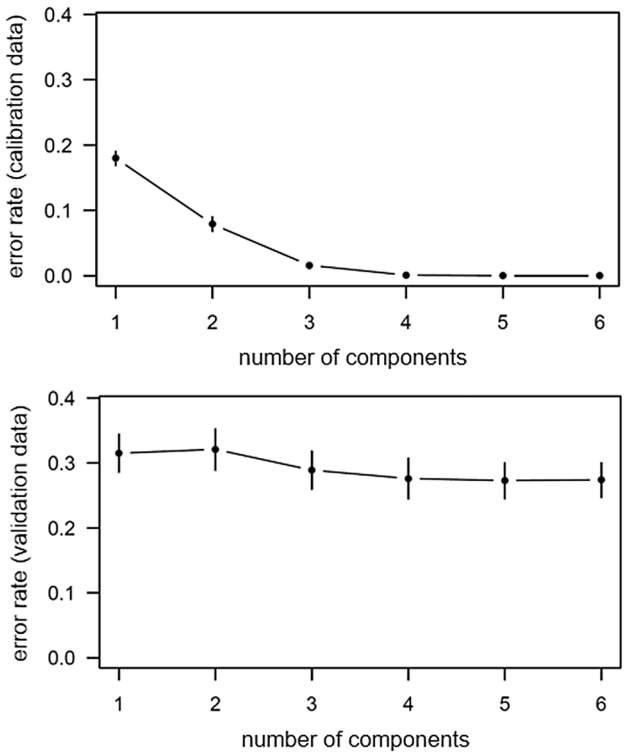

Fig. 3 Means and 95% confidence intervals of classification error rates according to the component number in model 3

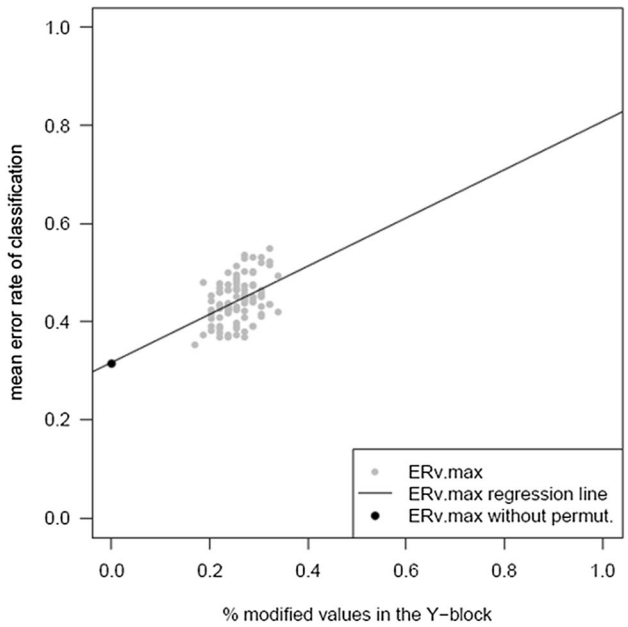

Fig. 4 Evolution of classification error rates for validation data (ERv.max) according to the percentage of modified values in the Y-block in model 3

3.1.2 Validity test of the optimal MBPLSDA models

The results of the permutation tests were also comparable between the three models (Fig. S3). The example of model 3 is given in Fig. 4. For each model, no cross-validated error rate on a permutated Y-block was lower than the one on the original Y-block. The average error rates obtained from cross-validation on original Y-blocks were unlikely to be reached when performing permutations of Y-blocks, means values being respectively 0.455 ± 0.041, 0.452 ± 0.042 and 0.443 ± 0.043 in models 1, 2 and 3 (p-values < 0.001). Similarly, the coefficients of the cross-validated error rate regression lines as a function of the percentage of modified values in the permutated Y-blocks were the highest, with our tested data, when the number of explanatory variables were the lowest (0.359 in model 1, 0.409 in model 2 and 0.491 in model 3) and were significantly different of zero (p-values < 0.01). Therefore, all three models were considered valid.

3.1.3 Cross-validated predicted observation categories by MBPLSDA models

The scatterplots of individuals, colored according to their true or cross-validated predicted status, were found highly similar between the three models (Fig. S4), and made it possible to visualize a good discrimination of the groups, and predictivity of the models.

3.1.4 Cross-validated values of the optimal MBPLSDA models parameters

The cross-validated BIPcum of the 3 models were found comparable: the BIPcum of the cross-validated clinical block was higher than the metabolomic block one, itself higher than the nutritional block one (Table 1). However, the cross-validated importance of the metabolomic block in the models slightly increased in model 3 compared to the others, at the cost of the cross-validated importance of the clinical block. This could be due to the decrease in the inertia of the metabolomic block, and therefore its lower weighting.

In addition, the variables with highest cross-validated weights on the model component, regression coefficients (not shown), or VIPcum (Figs. 5 and S5) were the clinical variables in the three models. Then, some metabolomic and nutritional variables came out at the 17th and 21th

Table 1 Bootstrap values of BIPcum in MBPLSDA models (mean ± SD)

BIPcum	Model 1	Model 2	Model 3
Clinical	0.473 ± 0.072	0.470 ± 0.070	0.449 ± 0.067
Metabolomic	0.315 ± 0.063	0.319 ± 0.058	0.349 ± 0.055
Nutritional	0.213 ± 0.059	0.211 ± 0.059	0.202 ± 0.055

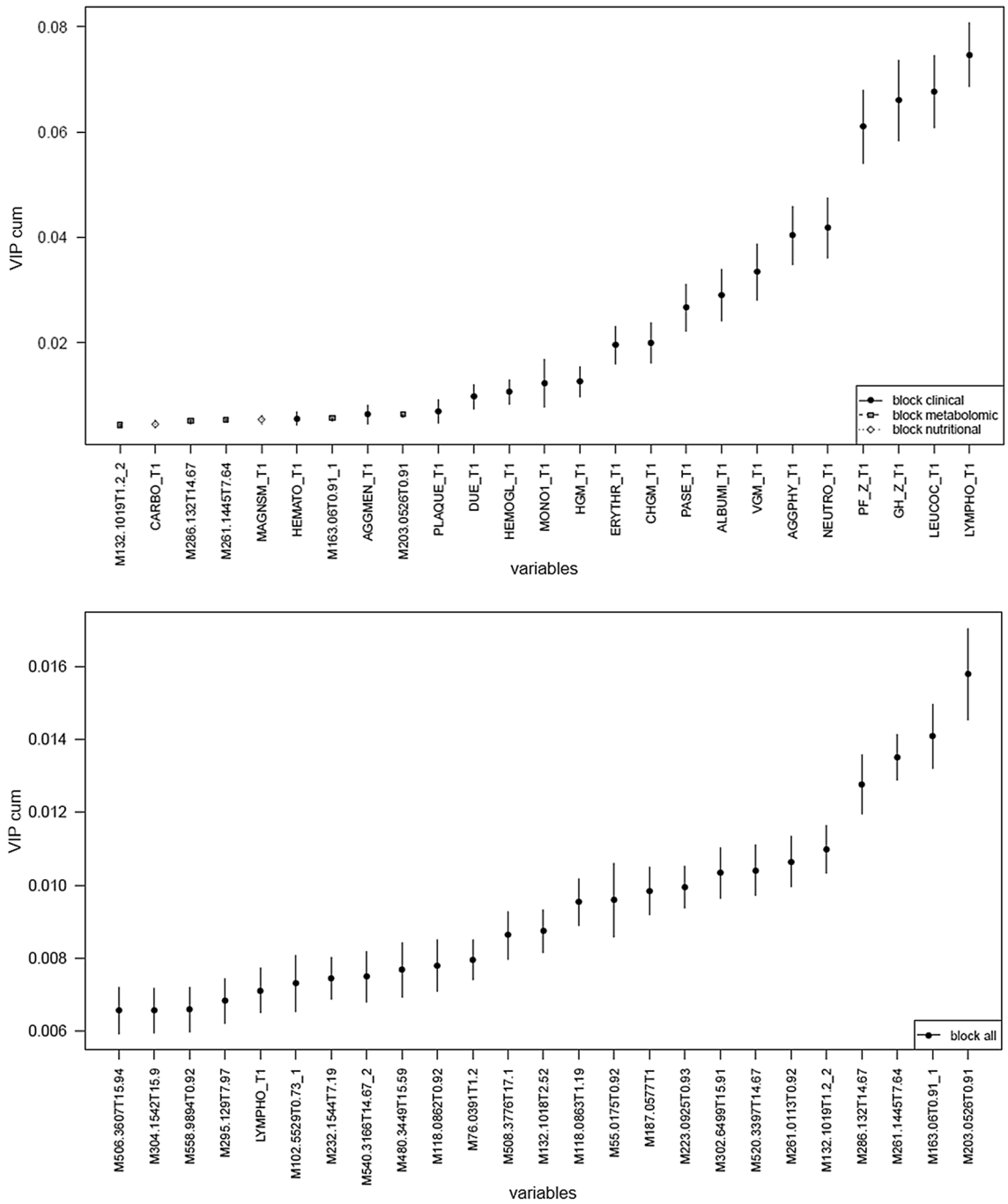

Fig. 5 Values and 95% confidence intervals of the 25 higher cumulative variable importances in the projection (VIPcum) in models 3 (top graph) and PLSDA (bottom graph)

rank on 493 variables respectively, in the model 3, and the same nutritional variables came out in models 1 and 2. In these two models, in descending order of VIPcum, the first metabolomic variables were at the 49th or 34th rank on 1656 or 1196 variables. Moreover, the variables having a VIPcum value higher than the mean VIPcum value were all the clinical variables and more than 66% of nutritional variables and 6% of metabolomic variables in models 1 or 2. They gathered all the clinical variables, and 17% and 9% of the nutritional and metabolomic variables respectively, in model 3. It could be explained by the lower weigthing of the metabolomic block in model 3 than in models 1 and 2.

3.1.5 Interest of metabolomic variable filtering

There was no change of information related to the case/control discrimination when filtering analytical redundancy, whereas there were some, by additionally filtering using co-factor effects.The quality and classification error rates were slightly improved in that case. The low importance of the metabolomic variables in models 1 and 2 could be due to a more important weighting of this block compared to the others, because of its larger size and information content. In particular, background noise, redundancy and co-factor effect tend to penalise the metabolomic block. Therefore, identifying the unwanted variables and filtering them could allow the model to better highlight the expected potential of the metabolomic block.

3.2 Comparison of MBPLSDA and standard PLSDA applied to concatenated blocks

3.2.1 Selection, validity test and cross-validated predictions of the optimal PLSDA model

The PLSDA comparison model included variables from the clinical and nutritional blocks, and the metabolomic dataset 3, concatenated in one single X-block. To avoid any risk of overfitting, the model considered as optimal model had only one component (Fig. S2), as in the MBPLSDA model. The cross-validated classification error rate on the validation data was found significantly lower than the one obtained by MBPLSDA for the same dataset, around 0.250 ± 0.070 (p-value = 0.001). Similarly, the permutation test supported the model validity (Fig. S3), and the scatterplots showed a good discrimination of the groups, and a good predictivity performance (Fig. S4).

3.2.2 Cross-validated values of the optimal PLSDA model parameters

The variables with the highest cross-validated weights on the model component, highest regression coefficients (not shown), or highest VIPcum (Figs. 5 and S5) were metabolomics features. In descending order of VIPcum, the first clinical and nutritional variables came out respectively at the 21st and 37th rank on 493 variables. Moreover, the variables having a VIPcum value higher than the mean VIPcum gathered 34% of the metabolomic features, 50% and 16% of the clinical and the nutritional ones, respectively. Because of their number, their residual correlation, and their probably higher diversity of variability compared to clinical and nutritional variables, it was expected that metabolomic variables were among the ones with the most significant weights. However, there is a higher risk for them to be related by chance to the outcome Y due to the dataset size. In addition, in a context of variable selection in clinical study, where the highest predictive power may not always be the main objective, allowing the model to give the same chance to epidemiological data as to the metabolomic ones may be of great interest.

Due to mathematical properties (part 2.1 and Table S1), when considered by block, the most important variables were found in the same VIPcum value order in the MBPLSDA and PLSDA models. When considered together, in our example, the MBPLSDA approach allowed us to obtain a more integrative subset of important variables.

4 Conclusion

MBPLSDA with data scaling and block weighting by their inertia is a discriminant method suited for the joint analysis of structured data in blocks of heterogeneous sizes (metabolomic and epidemiological data). The present work allowed performing all the steps of this method within a standalone tool. Although this was not illustrated in our example, the MBPLSDA method can be applied using the proposed package to explain a Y-block with several variables, each having 2 or more categories. The added value of this tool is to allow easy model evaluation, to provide indicators for model comparison and to facilitate the adaptation of statistical analysis to the experimental design. Using the comparison indicators provided, we found that MBPLSDA was improved by a relevant variable filtering within blocks. In terms of strategy, if the metabolomic variables have to be filtered, we recommend to eliminate analytical redundancy and, when appropriate, to filter using experimental cofactors. 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 more integrative results. Thus, this method can be interesting to obtain a global insight of a biological phenomenon.

Acknowledgements The authors would like to thank Charlotte Joly for the analyses by LC–MS and Stéphanie Monnerie for the provision of the data. All metabolomics analyses were performed within the metaboHUB French infrastructure (ANR-INBS-0010). The NuAge Study was supported by a research grant from the Canadian Institutes of Health Research (CIRH; MOP-62842). The NuAge Database and Biobank are supported by the Fonds de Recherche du Québec (FRQ; 2020-VICO-279753), the Quebec Network for Research on Aging funded by the Fonds de Recherche du Québec - Santé (FRQS) and by the Merck-Frosst Chair funded by La Fondation de l'Université de Sherbrooke.

Author contributions MP, MBB, BC and EPG conceived the study. MBB and SB designed the tool. EP, BC and PG provided the data. MBB analyzed the data. MBB, MP, BC and EPG wrote the manuscript. All authors read and approved the manuscript.

Data availability The tool is developed under the R software and is available via the *packMBPLSDA* CRAN R package.

Compliance with ethical standards

Conflict of interest All authors declare they have no conflict of interest.

Human and animal rights All procedures performed in the study involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study. The NuAge Study has been approved by the Research Ethics Board (REB) of both the Geriatric University Institutes of Montreal and Sherbrooke. The management framework of the NuAge Database and Biobank has been approved by the REB of the CIUSSS-de-l'Estrie-CHUS.

References

Ackermann, B. L., Hale, J. E., & Duffin, K. L. (2006). The role of mass spectrometry in biomarker discovery and measurement. *Current Drug Metabolism, 7,* 525–539.

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics, 17,* 166–173.

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

Bougeard, S., Qannari, E. M., & Rose, N. (2011). Multiblock redundancy analysis: Interpretation tools and application in epidemiology. *Journal of Chemometrics, 25,* 467–475.

Dray, S., & Dufour, A. (2007). The ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software, 22,* 1–20.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. New York: Chapman and Hall/CRC.

Gaudreau, P., Morais, J. A., Shatenstein, B., Gray-Donald, K., Khalil, A., Dionne, I., et al. (2007). Nutrition as a determinant of successful aging: Description of the Quebec longitudinal study Nuage and results from cross-sectional pilot studies. *Rejuvenation Research, 10*, 377–386.

Giacomoni, F., Le Corguille, G., Monsoor, M., Landi, M., Pericard, P., Petera, M., et al. (2015). Workflow4Metabolomics: A collaborative research infrastructure for computational metabolomics. *Bioinformatics, 31,* 1493–1495.

Gromski, P. S., Muhamadali, H., Ellis, D. I., Xu, Y., Correa, E., Turner, M. L., et al. (2015). A tutorial review: Metabolomics and partial least squares-discriminant analysis—A marriage of convenience or a shotgun wedding. *Analytica Chimica Acta, 879,* 10–23.

Gunther, O. P., Shin, H., Ng, R. T., McMaster, W. R., McManus, B. M., Keown, P. A., et al. (2014). Novel multivariate methods for integration of genomics and proteomics data: Applications in a kidney transplant rejection study. *OMICS: A Journal of Integrative Biology, 18,* 682–695.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer Nature.

Pujos-Guillot, E., Brandolini, M., Petera, M., Grissa, D., Joly, C., Lyan, B., et al. (2017). Systems metabolomics for prediction of metabolic syndrome. *Journal of Proteome Research, 16,* 2262–2272.

R Development Core Team (2019) R: A language and environment for statistical computing. *R Foundation for Statistical Computing,* Vienna, Austria.

Ramautar, R., Berger, R., van der Greef, J., & Hankemeier, T. (2013). Human metabolomics: Strategies to understand biology. *Current Opinion in Chemical Biology, 17,* 841–846.

Rohart, F., Gautier, B., Singh, A., & Le Cao, K. A. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Computational Biology, 13,* e1005752.

Saccenti, E., Hoefsloot, H. C. J., Smilde, A. K., Westerhuis, J. A., & Hendriks, M. M. W. (2014). Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics, 10,* 361–374.

Saporta, G. (2006). *Probabilités, analyse de données et statistiques*. Paris: Editions Technip.

Singh, A., Gautier, B., Shannon, C., Vacher, M., Rohart, F., Tebbutt, S., & Lê Cao, K. A. (2016). DIABLO: From multi-omics assays to biomarker discovery, an integrative approach. https://doi.org/10.1101/067611.

Steuer, R. (2006). Review on the analysis and interpretation of correlations in metabolomic data. *Briefings in Bioinformatics, 7,* 151–158.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B, 36,* 111–147.

Tautenhahn, R., Bottcher, C., & Neumann, S. (2008). Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics, 9,* 504.

Westerhuis, J. A., Hoefsloot, H. C. J., Smit, S., Vis, D. J., Smilde, A. K., van Velzen, E. J. J., et al. (2008). Assessment of PLSDA cross validation. *Metabolomics, 4,* 81–89.

Westerhuis, J. A., Kourti, T., & Macgregor, J. F. (1998). Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics, 12,* 301–321.

Wold, S. (Ed.). (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.

SUPPLEMENTAL INFORMATION

Multi-block PLS discriminant analysis for the joint analysis of metabolomic and epidemiological data

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

1. Université Clermont Auvergne, INRA, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, F-63000 Clermont-Ferrand, France
2. Centre de Recherche du Centre hospitalier de l'Université de Montréal, Montréal, Canada
3. Département de médecine, Université de Montréal, Montréal, Canada
4. Université Clermont Auvergne, INRA, UNH, F-63000 Clermont-Ferrand, France
5. Anses, BP53, Technopole Saint Brieuc Armor, 22440, Ploufragan, France

***Corresponding author:** Marion Brandolini-Bunlon ; marion.brandolini-bunlon@inra.fr ; tel +33473624676

Journal: Metabolomics

Supplemental information: Supplemental material 1, Figures S1-S5, Tables S1-S2

Table of contents of Supplemental information

Supplemental material 1

Data collection

The present study was designed within the 5-year longitudinal observation study of NuAge (Quebec Longitudinal Study on Nutrition and Successful Aging) constituted of 1,793 healthy men and women (Gaudreau et al., 2007). French- or English-speaking community-dwelling participants were committed to give fasting blood annually and to answer questionnaires related to food and health biannually. The NuAge database comprises large qualitative and quantitative data related to nutrition/dietary intakes, physical activity, and numerous markers of physical and cognitive status, functional autonomy and social functioning. The designed case-control study on Metabolic Syndrome (MetS) included individuals selected stable regarding their MetS status.

Metabolomics processing

After data extraction and standard quality processing, a first dataset of 1656 variables was generated ("dataset 1").

Then, a second dataset of 1091 variables ("dataset 2") was generated by removing signals below a mean intensity of 1000, as well as the highest correlations. Features coming from the same metabolite were grouped, using the following criteria: pair-wise correlation coefficients of intensities > 0.9; retention time difference < 0.1 min; mass difference < 0.005 Da (using a reference list containing isotopes, adducts and fragments). The highest intensity feature of each group was then selected as representative (parent ion).

The third reduced dataset was generated by selecting ions without significant time effect detected by repeated measures ANOVA (also called "mixed models") on dataset 2. The *mixmodel* module of the Galaxy web-based platform Worflow4metabolomics, based on the *lmerTest*, *nlme* and *multtest* R-packages, was used. As no interaction effect was observed between status and time, repeated measures ANOVA were performed considering status (case/control) as a fixed effect, T1-T4 as a time effect, and subject as random effect. This method was considered appropriate due to the repeated design and the sufficient number of individuals to assume normality of data. A p-value threshold of 0.05 after Benjamini-Hochberg correction was considered to detect variables strongly affected by time. Finally, this dataset was reduced by selecting the highest intensity features among groups of highly correlated ions ($r^2 > 0.8$ without considering retention time nor mass information), resulting in a third data set ("dataset 3").

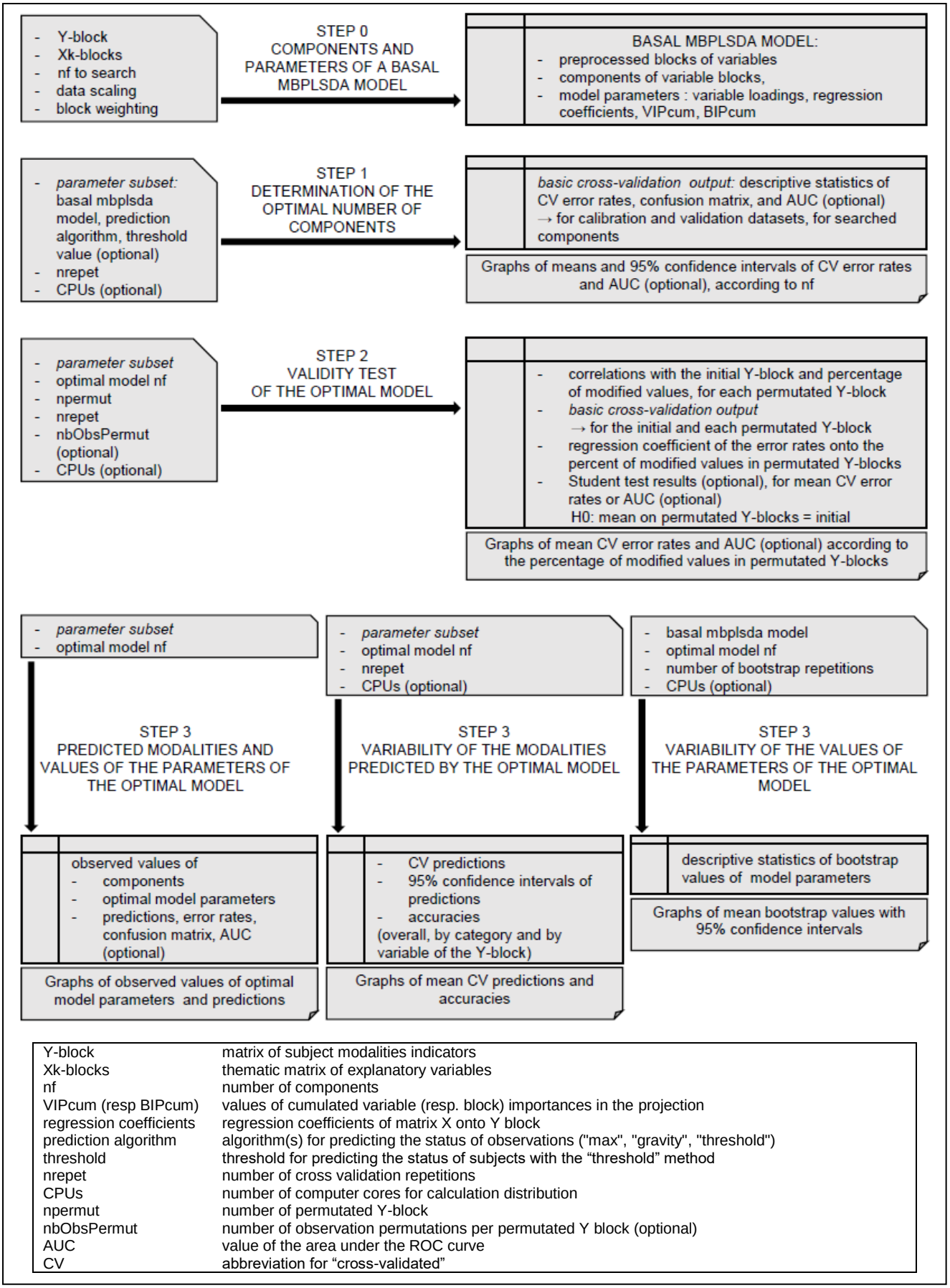

Figure S1 Steps for the application of the MBPLSDA method, and main inputs and outputs of the developed programs.

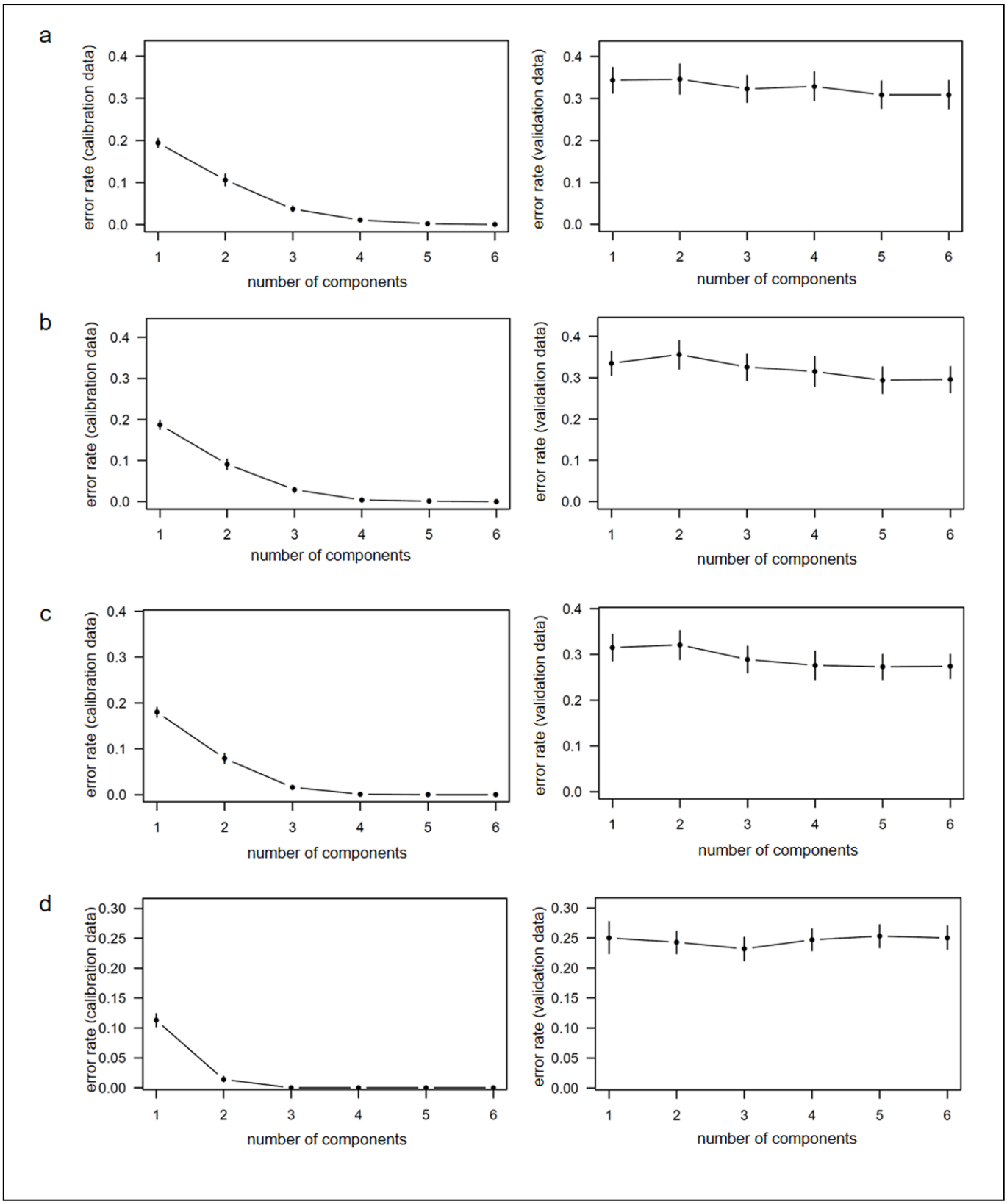

Figure S2 Means and 95% confidence intervals of classification error rates according to the number of components in MBPLSDA and PLSDA models **(a)** model 1, **(b)** model 2, **(c)** model 3, and **(d)** PLSDA model.

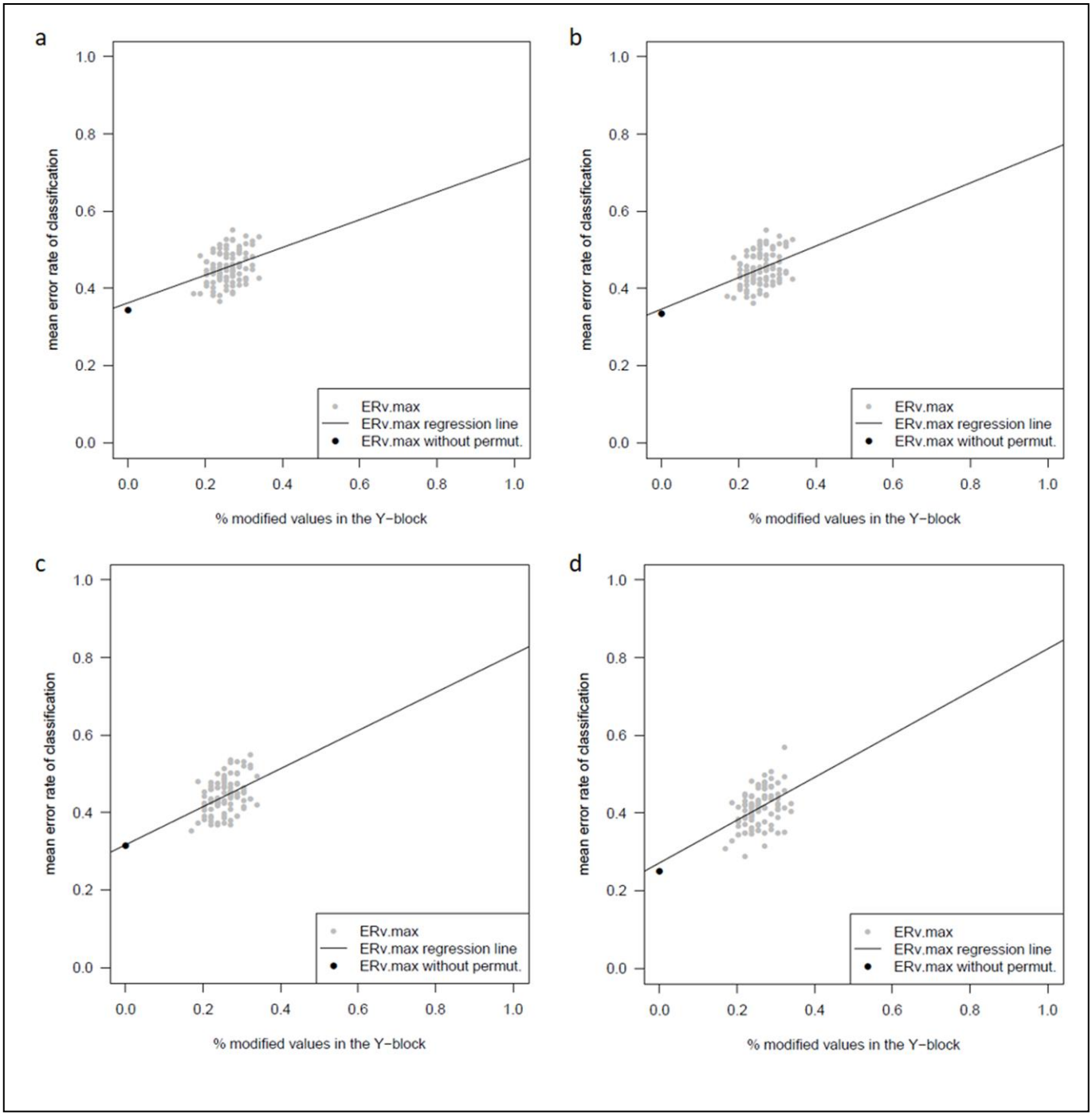

Figure S3 Evolution of classification error rates for validation data (ERv.max) according to the percentage of modified values in the Y-block in MBPLSDA and PLSDA models **(a)** model 1, **(b)** model 2, **(c)** model 3, and **(d)** PLSDA model.

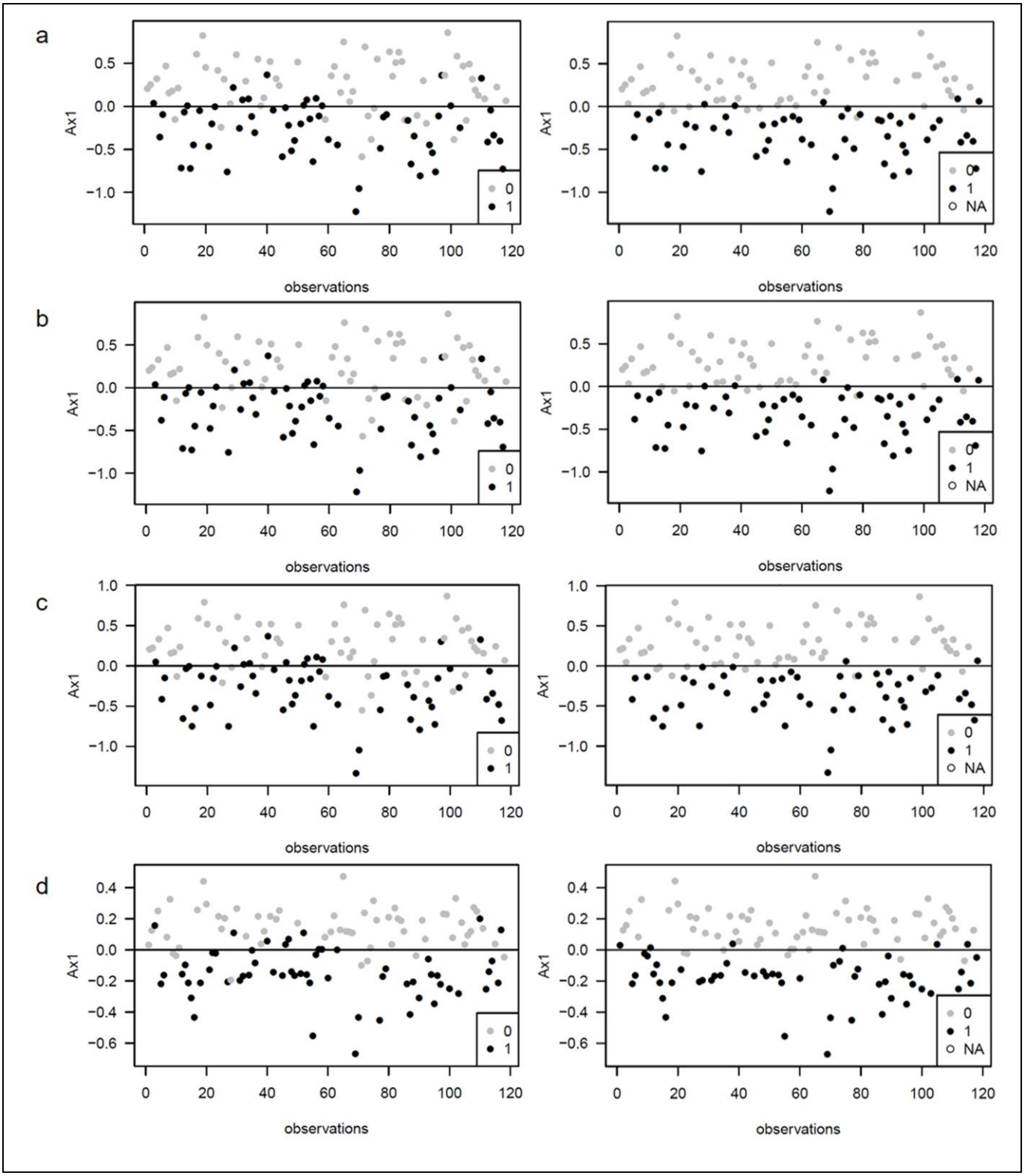

Figure S4 Scatterplots of observations with coloration according to (from left to right) their observed status, their cross-validated predicted status evaluated on the validation subsets in MBPLSDA and PLSDA models **(a)** model 1, **(b)** model 2, **(c)** model 3, and **(d)** PLSDA model.

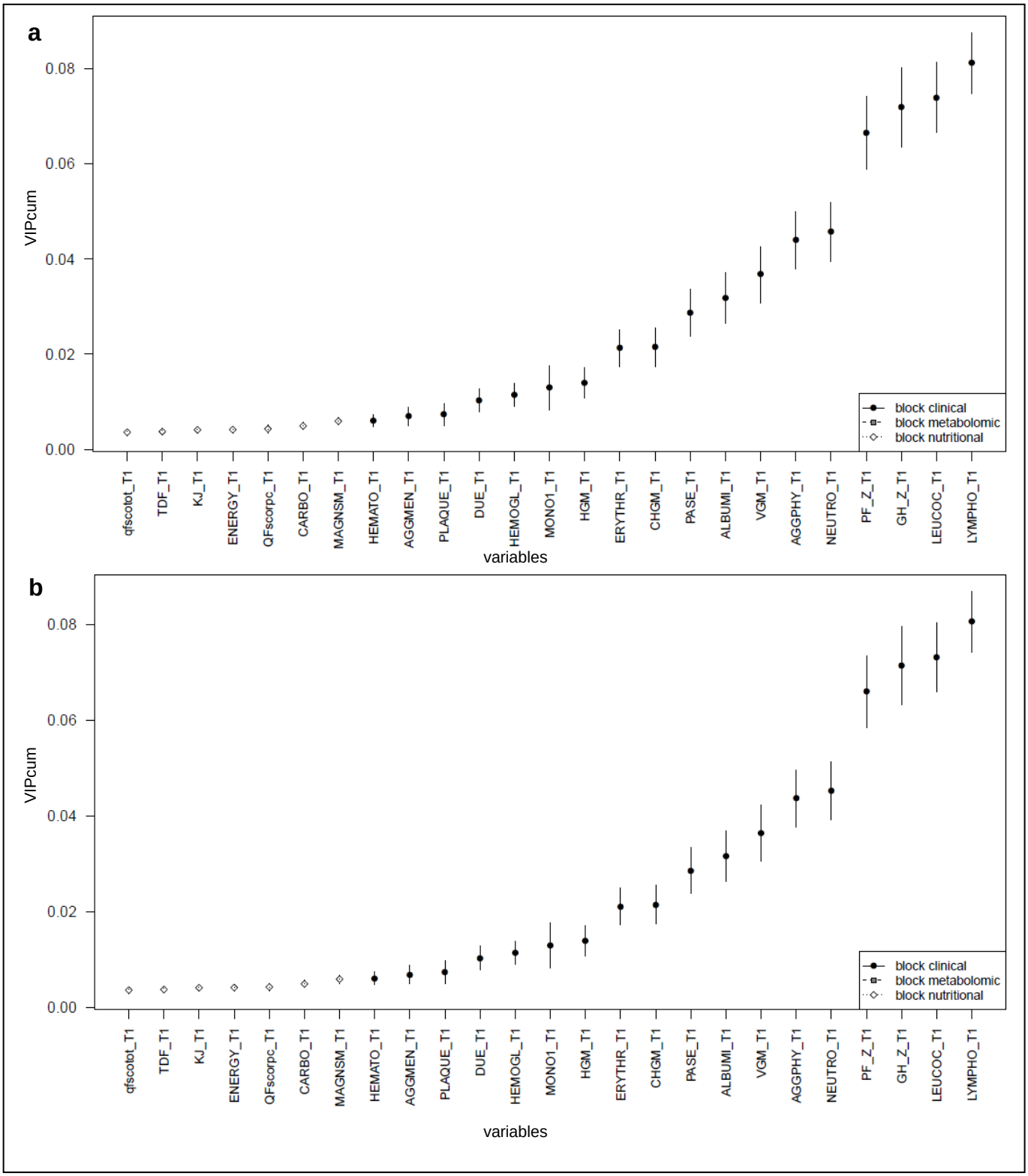

Figure S5 Values and 95% confidence intervals of the 25 higher cumulative variable importances in the projection (VIPcum) in MBPLSDA and PLSDA models **(a)** model 1, **(b)** model 2, **(c)** model 3, and **(d)** PLSDA model.

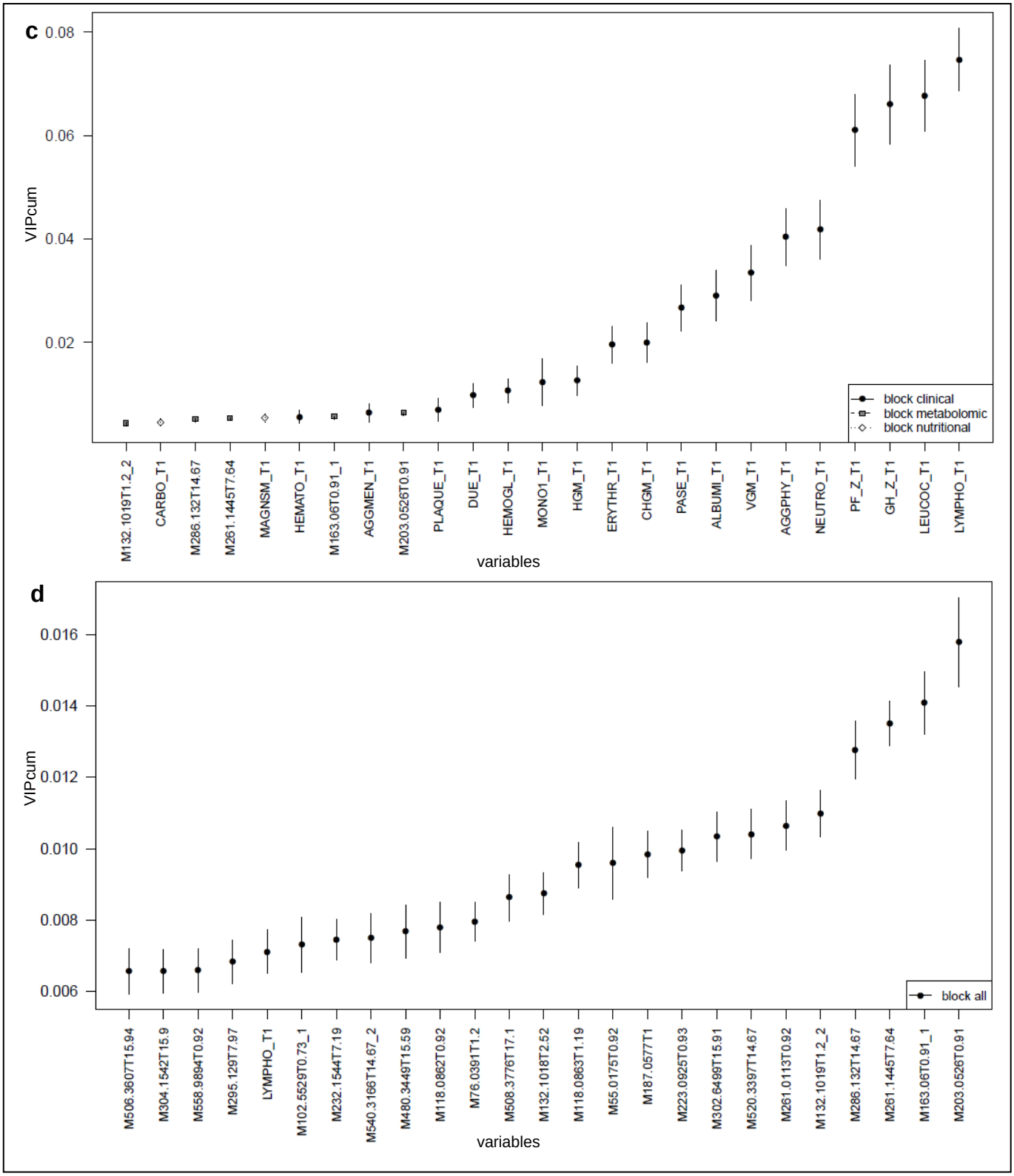

Figure S5 Values and 95% confidence intervals of the 25 higher cumulative variable importances in the projection (VIPcum) in MBPLSDA and PLSDA models **(a)** model 1, **(b)** model 2, **(c)** model 3, and **(d)** PLSDA model

Table S1 Calculations of a MBPLSDA model (Bougeard *et al*., 2011)

	Equations
NOTATIONS	K: the number of explanatory variable blocks, $k = 1, \ldots, K$ X_k: a block of explanatory variables Y: the block of variables to be explained N: the number of observations Θ: Hadamard element-wise product
Inertia applied for block weighting and X_k weighting	$inertia = \frac{\left(trace\left(X_k X_k'\right)\right)}{N}$ X_k weighting : $X_k \longmapsto \frac{X_k}{\sqrt{\frac{\left(trace\left(X_k X_k'\right)\right)}{N}}}$
Maximisation problem (first order solution)	$cov^2(u^{(1)}, t^{(1)})$ with $t^{(1)} = \sum_{k=1}^{K} a_k^{(1)} t_k^{(1)}$, $\sum_{k=1}^{K} (a_k^{(1)})^2 = 1$, $u^{(1)} = Y v^{(1)}$, $t_k^{(1)} = X_k w_k^{(1)}$, $\left\|t_k^{(1)}\right\| = \left\|v^{(1)}\right\| = 1$
Variable weights ($v^{(1)}$) (first order solution)	$v^{(1)}$ = Eigenvector of the matrix $\sum_{k=1}^{K} Y' X_k X_k' Y$ associated with the largest eigenvalue $\lambda^{(1)}$
Partial components ($t_k^{(1)}$) (first order solution)	$t_k^{(1)} = \frac{(P_{X_k} u^{(1)})}{\left\|P_{X_k} u^{(1)}\right\|}$ with $P_{X_k} = X_k (X_k' X_k)^{-1} X_k'$
Coefficients ($a_k^{(h)}$)	$a_k^{(h)} = \frac{cov\left(u^{(h)}, t_k^{(h)}\right)}{\sqrt{\sum_{k=1}^{K} cov^2\left(u^{(h)}, t_k^{(h)}\right)}} = \frac{\left\|P_{X_k} u^{(h)}\right\|}{\sqrt{\sum_{k=1}^{K} \left(\left\|P_{X_k} u^{(h)}\right\|\right)^2}}$
Global component ($t^{(h)}$)	$t^{(h)} = \sum_{k=1}^{K} a_k^{(h)} t_k^{(h)}$
Deflation step before search of the next component	$X_k^{(h+1)} = \left(I - \frac{t^{(h)} t^{(h)\prime}}{\left\|t^{(h)}\right\|^2}\right) X_k^{(h)}$
Regression coefficients of the Y-block onto the explanatory variables in the MBPLSA model including *h* components ($b^{(h)}$)	$b^{(h)} = \sum_{l=1}^{h} w^{(l)*} * c^{(l)}$ with $w^{(h)*} = \prod_{l=1}^{h-1} \left[I - \frac{w^{(l)} t^{(l)\prime}}{\left\|t^{(l)}\right\|^2}\right] w^{(h)}$, and $c^{(h)} = \frac{Y' t^{(h)}}{\left\|t^{(h)}\right\|^2}$
BIPcum of X_k in the MBPLSA model including *h* components ($BIPcum_k^h$)	$BIPcum_k^h = \frac{\sum_{l=1}^{h} \lambda^{(l)} (a_k^{(l)})^2}{\sum_{l=1}^{h} \lambda^{(l)}}$
VIPcum in the MBPLSA model including *h* components ($VIPcum^h$)	$VIPcum^h = \frac{\sum_{l=1}^{h} \lambda^{(l)} \frac{\left(\left(a_k^{(l)} \odot a_k^{(l)}\right) \odot \left(w^{(l)*} \odot w^{(l)*}\right)\right)}{\left\|\left(a_k^{(l)} \odot a_k^{(l)}\right) \odot \left(w^{(l)*} \odot w^{(l)*}\right)\right\|}}{\sum_{l=1}^{h} \lambda^{(l)}}$

Table S2 Example of function call, illustrated with "model 3"

Steps	R code
DATA	library(packMBPLSDA) ## 1.1. IMPORT AllData <- read.table("ClinNutMetabo.txt", header = TRUE, sep = "\t", rownames = 1) status <- data.frame(status= AllData["status"], row.names = row.names(AllData)) medical <- data.frame(AllData [,2:19]) nutrition <- data.frame(AllData [,20:106]) omics <- data.frame(AllData [,107:498) ## 1.2. X variables ktabX <- ktab.list.df(list(clinical = medical, nutritional = nutrition, metabolomic = omics)) ## 1.3. Y variable disjonctif <- (disjunctive(status)) dudiY <- dudi.pca(disjonctif , center = FALSE, scale = FALSE, scannf = FALSE) bloYobs <- 2
STEP 0. Components and parameters of a basal MBPLSDA model	mbplsdamodel <- mbplsda(dudiY, ktabX, scale = TRUE, option = "uniform", scannf = FALSE, nf = 6)
STEP 1. Determination of the optimal number of components	resdim <- testdim_mbplsda(object = mbplsdamodel, nrepet = 30, threshold = 0.5, bloY = bloYobs, cpus = 24, algo = "max") plot_testdim_mbplsda (obj = resdim, filename = "plotTDim_6nf_30rep") ncpopt <- 1
STEP 2. Validity test of the optimal model	rtsPermut <- permut_mbplsda (object = mbplsdamodel, nrepet = 30, npermut = 100, optdim = ncpopt, threshold = 0.5, bloY = bloYobs, nbObsPermut = 40, cpus = 24, algo = "max") plot_permut_mbplsda(obj = rtsPermut, filename = "plotPermut_1nf_30rep_100perm")
STEP 3. Predicted modalities and values of the parameters of the optimal model	predictions <- pred_mbplsda(object = mbplsdamodel, optdim = ncpopt, threshold = 0.5, bloY = bloYobs, algo = "max") plot_pred_mbplsda(obj = predictions, filename = "plotPred _1nf_25var", propbestvar = (25/497))
STEP 3. Variabilities of the modalities predicted by the optimal model	predCV <- cvpred_mbplsda (object = mbplsdamodel, nrepet = 120, optdim = ncpopt, threshold = 0.5, bloY = bloYobs, cpus = 24, algo = "max") plot_cvpred_mbplsda(obj = predCV, filename = "plotCVPred _1nf_120rep")
STEP 3. Variabilities of the values of the parameters of the optimal model	resboot <- boot_mbplsda(object = mbplsdamodel, optdim = ncpopt, nrepet = 100, cpus = 24) plot_boot_mbplsda(obj = resboot, filename = "plotBoot _1nf_100rep", propbestvar = (25/497))

References

Bougeard, S., Qannari, E.M. and Rose, N. (2011) Multiblock redundancy analysis: interpretation tools and application in epidemiology. *Journal of Chemometrics* **25,** 467-475.

Gaudreau, P., Morais, J.A., Shatenstein, B., Gray-Donald, K., Khalil, A., Dionne, I., Ferland, G., Fülöp, T., Jacques, D., Kergoat, M.J., Tessier, D., Wagner, R. and Payette, H. (2007) Nutrition as a determinant of successful aging: description of the Quebec longitudinal study NuAge and results from cross-sectional pilot studies. *Rejuvenation Res* **10,** 377-86.