Plsda versus pca on barycenters applied to metabolomics in a context of discrimination

Marion Brandolini-Bunlon, Benoît Jaillais, Mohamed Hanafi

PLSDA VERSUS PCA ON BARYCENTERS APPLIED TO METABOLOMICS IN A CONTEXT OF DISCRIMINATION

Marion Brandolini-Bunlon[1], Benoît Jaillais[2], Mohamed Hanafi[2]

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

[2]Oniris, INRAE, StatSC, 44300 Nantes, France

E-mail: marion.brandolini-bunlon@inrae.fr

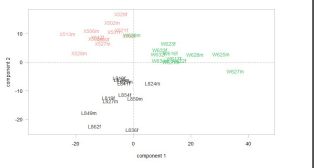
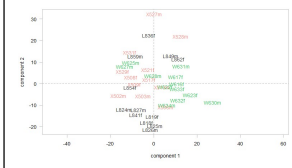
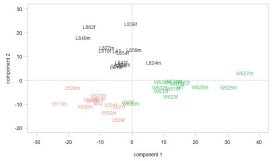
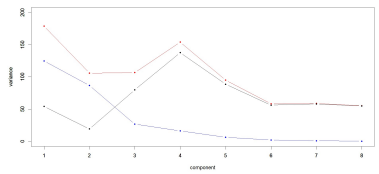
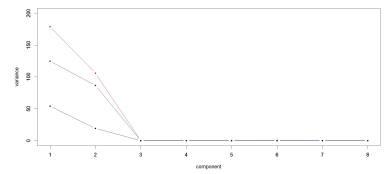
Abstract for **oral communication**

Introduction. Untargeted metabolomics is a powerful phenotyping tool to better understand the biological mechanisms involved in the physiopathological processes, and identify biomarkers of metabolic status. The complex data need dedicated preparation and treatments to extract meaningful information. The major specificity of metabolomics data is the large number of variables compared to the number of samples, as well as their high degree of correlation. The common analysis strategy consists in performing univariate and multivariate statistics to highlight variables of interest. In a discriminant context, partial least squares-discriminant analysis (PLSDA) is one of the most effective multivariate tools currently used, because of its ability to analyze collinear and noisy data. Another multivariate method that could be used is the Principal Component Analysis (PCA) of the matrix of barycenters of the observation groups (called here "PCA_C"). The objective of our study is to compare these approaches in terms of explained variances and important variables.

Material and methods. Published data from a project on the impact of genetic mutations in mice (ProMetIS) were used as a case study (Imbert, 2021). Mice (n=42), males and females, belonged to one of the three genotype groups (wild type, lacking the linker for activation of T cells gene, or lacking the MX dynamin-like GTPase 2 gene). The metabolomics dataset we used, was obtained from the analysis of plasma samples using a mass spectrometry-based untargeted approach (LC-MS), and contained 6104 variables after preparation. In the present work, data analysis was performed with the R-package "rchemo". Six atypical mice were removed to have a balanced experiment design, before Pareto scaling. Due to the sex effect,

without interaction with the genotype, the data were centered by sex before applying PLSDA, and PCA_C, to discriminate genotype groups. On one hand, the optimal number of PLS components was determined according to the error in repeated cross-validation (30 repetitions of 10-fold cross-validation) and application of the one-standard-error-rule. On the other hand, using PCA_C model, subjects were projected onto the components. For each PLSDA or PCA_C component, the total, inter- and intra-group variances were then calculated, and the group effect was assessed by ANOVA. ANOVA were also performed for metabolomics variables becoming important in the discrimination in the PLSDA model with the optimal number of components (called here "$PLSDA_{opt}$"), compared to the one with 2 components.

Results. The optimal number of PLS components was 3, and the number of PCA_C components was 2. Only these components had a significant p-value in the ANOVA. As expected, the 1st component of both methods were the same, and, as shown below, in our study, the 2nd components were closely similar. The 3rd component of the PLSDA model was also of interest because it still significantly explained intergroup variability and highlighted other important discriminant variables.

	PLSDA	PCA_C
Score scatter plots	 	
Component variances • total • between • within		

Discussion and conclusion. The PLSDA and the PCA_C components maximize the inter-group variability. When 3 groups are to be discriminated, the PCA_C finds 2 components while the PLSDA can find more. Presumably, each component of the $PLSDA_{opt}$ can discriminate one group from one or both others, and would thus allow a better discrimination.

References

Imbert A, *et al.* (**2021**) ProMetIS, deep phenotyping of mouse models by combined proteomics and metabolomics analysis. *Sci Data*, 8(1), 311.