PLSDA versus PCA on barycenters, applied to metabolomics in a context of discrimination

Marion Brandolini-Bunlon, Benoit 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[3]

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

[2] Oniris, INRAE, StatSC, 44300 Nantes, France, benoit.jaillais@inrae.fr

3 Oniris, StatSC, 44300 Nantes, France, mohamed.hanafi@oniris-nantes.fr

Keywords: metabolomics, PCA on barycenters, PLSDA.

1 Introduction

Metabolomics is a powerful phenotyping tool in systems biology that provides a view of metabolic changes in the whole organism. In particular, the "global or untargeted" metabolomics approach notably defines a characteristic fingerprint of all metabolites present in the sample. The data are therefore massive and complex, and in particular, highly noisy, with many more variables than samples, and collinearities between variables of analytical or biological origin. They require appropriate statistical treatments to highlight the variables that characterize or predict the experimental groups. The common analysis strategy is to perform univariate and multivariate statistics to highlight the variables of interest. In a discriminant context, partial least squares discriminant analysis (PLSDA) is one of the most effective multivariate tools currently in use, 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 (here called "PCAc") whose objective, like PLSDA, is to find components that maximize the variance between groups. The aim of our study is to compare these approaches in terms of components and important variables.

2 Material and methods

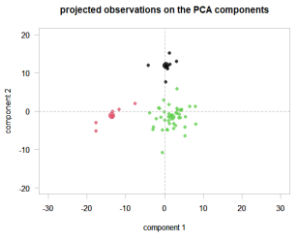
In the present work, the proposed method was tested on a set of real data published by Zhang et al [1], and data analysis was performed using the R package "rchemo" [2]. In their study, Zhang et al. investigated the effect of mouse loss-of-function mutations in 30 unique gene knockout lines on plasma metabolites. In our study, to better understand the differences between the methods, 3 groups of observations (a control group + 2 genotypes) were selected. The data matrix for subsequent analyses was therefore reduced to 52 observations (wild type (n=40), Atp6v0d1 gene knocked out (n=6), Iqgap1 gene knocked out (n=6)) and 823 variables before centering and scaling to unit variance. This matrix is labeled *Xinit*. Exploratory analysis of the data revealed a strong sex effect in the data, which was eliminated by centering the data by sex. The resulting data matrix was used for PCA of genotype group barycenters and PLSDA to discriminate genotype groups. This matrix is denoted *X* and the genotype group indicator matrix is denoted *Y*.

In PCAc, the barycenters correspond to an average observation weighted by the group size defined by *Y*. We noted $\bar{X}$ the corresponding barycenter matrix on which a centered PCA was performed.

Therefore, all initial subjects were therefore projected onto the components. For PLSDA, the optimal number of PLS components was determined according to the error in repeated cross-validation and application of the one-standard-error rule, and this model was validated with a permutation test. Kruskall-Wallis tests were also performed on the scores, and initial and predicted metabolomics variables became important for discrimination in the PLSDA model with the optimal number of components (here called "PLSDAopt") compared to the 2-component model.

3 Results and discussion

The number of components in PLSDAopt and PCAc were 3 and 2, respectively. All these components had a significant p-value in the Kruskall-Wallis tests. As expected, the 1st and 2nd components of both methods were very similar, and would have been exactly the same in the case of balanced plan. The 3rd component of the PLSDA model was also of interest because the scores of the genotypic groups were significantly different, and it highlighted other important discriminant variables.

	PLSDA		PCA_C
Score scatter plots			

4 Conclusion

Applying the one standard error rule, there are 3 components in PLSDAopt, while PCAc can't have more than 2 components when there are 3 group barycenters. The PLSDA model corrects for within-group error (a variable may have values that are significantly different between groups after projection). In addition, some variables, that are only important in the PLSDA model due to the additional components, provide an additional information to discriminate between the groups. In metabolomics, with the aim of understanding the biological mechanisms and metabolic pathways involved, and given the low rates of annotated variables, it may be interesting to highlight a larger number of variables involved in group discrimination. From a practical point of view for researchers, PCAc, which is easier to apply than PLSDA, can be used as a first approach to determine the discriminant variables.

References

[1] Zhang, Y., Barupal, D.K. Fan, S., Gao, B., Zhu, C., Flenniken, A.M., McKerlie, C., Nutter, L.M.J., Lloyd, K.C.K., Fiehn, O. Sexual Dimorphism of the Mouse Plasma Metabolome Is Associated with Phenotypes of 30 Gene Knockout Lines. *Metabolites* 13, 947, 2023. https://doi.org/10.3390/metabo13080947

[2] Brandolini-Bunlon M., Jallais B., Roger J.M. Lesnoff M., 2023 R package rchemo: Dimension Reduction, Regression and Discrimination for Chemometrics. https://github.com/ChemHouse-group/rchemo.