New Formulation for Multi-block-Partial Least Squares-Discriminant Analysis. Application to metabolomics data

Véronique Cariou[1], El Mostafa Qannari[1], Mohamed Soumah[1],
Marie-Cécile Alexandre-Gouabau[2] & Thomas Moyon[2]

[1] *LUNAM University, ONIRIS, USC "Sensometrics and Chemometrics Laboratory", Nantes, F-44322, France*
E-mail : *veronique.cariou@oniris-nantes.fr*

[2] *UMR PhAN, INRA, Nantes, France*

Abstract
In Chemometrics, the coupling of different kinds of measurements including genomics, proteomics and metabolomics generates a large amount of variables structured into meaningful blocks for the characterization of the same set of samples. Dealing with multi-blocks data in a discrimination scope, we propose, herein, to extend the PLS method to discrimination (PLS-DA), considering the decomposition of the between groups covariance matrix in the multi-block context. This leads to the simultaneous determination of global and block components. This method is illustrated on a case study pertaining to the LACATOL project (registered to the French Clinical Trial under N° NCT01493063) which aims at ensuring the optimal growth of preterm newborns through a personalized nutrition. A multi-block PLS-DA is performed to identify two phenotypes of milk, associated with a growth group (normal vs slow) of preterm newborns. The relationships between metabolomics, free amino acids and macronutriments are highlighted.

Keywords: Metabolomics ; Multiblock partial least squares ; Partial Least Squares Discriminant Analysis

1. Introduction

In Chemometrics, the coupling of different kinds of measurements including genomics, proteomics and metabolomics generates a large amount of variables structured into meaningful blocks for the characterization of the same set of samples. Therefore, the development of multivariate statistical methods taking account of the combination of the blocks has become a challenging task to provide a better understanding of the biological system (Skov et al, 2014). Notwithstanding, in the large literature covering multi-block analysis, to our knowledge, little attention has been paid to the development of an appropriate method for the discrimination problem in the multi-block framework (Boccard & Rutledge, 2013 ; Biancolillo, Måge & Næs, 2015). The usual approach, so-called multi-block PLSDA, consists in using a multi-block partial least squares regression (MB-PLS). As this model is dedicated to quantitative measurements, the categorical variable associated with the various groups is first dichotomized. Then, the MB-PLS regression is performed on the dummy matrix defined as the Y matrix.

We propose, herein, to extend a more natural PLS method for discrimination (PLS-DA), considering the decomposition of the between groups covariance matrix. The aim of this work is to extend the PLS-DA approach proposed both by Barker and Rayens (2003) and Nocairi & al. (2005) to the multi-block framework. Such an approach makes it possible to extract simultaneously block components and global components. From a user point of view, it provides graphical tools making it possible to check the discrimination ratio associated to the various blocks for a better interpretation.

2. New formulation of Multi-block PLS-DA

2.1 Recall of PLS-DA

Let us recall the formulation of PLS-DA proposed by Nocairi & al. (2005) and Barker and Rayens (2003). Let us define X the data matrix composed of n objects depicted by p continuous variables and one categorical variable, denoted C, whose q categories correspond to the various groups. We consider the binary matrix Y which corresponds to the dichotomization of the categorical variable C. PLS-DA consists in determining a latent variable t which is a linear combination of the X variables, such that:

$$\max var(P_y t) \text{ with } \mathrm{t} = \mathrm{Xu}$$

Where P_y is the projector associated to Y: $P_y = Y\left(Y^T Y\right)^{-1} Y^T$.

This leads to the maximization of the following criterion:

$$\max \frac{1}{n} u^T X^T Y\left(Y^T Y\right)^{-1} Y^T X u$$

This new formulation provides a clear interpretation of the first PLS-DA component as the first principal component of the between groups covariance matrix: $\frac{1}{n} X^T Y\left(Y^T Y\right)^{-1} Y^T X$. The successive latent components are determined as in PLS regression analysis by deflating the X matrix on the basis of the latent components already obtained at the previous steps.

2.2 Extension of PLS-DA in a multi-block framework

Consider a set of quantitative variables measured on the same objects and structured into K meaningful blocks $(\mathrm{X_k})(\mathrm{k} = 1, \dots, \mathrm{K})$, where $\mathrm{X_k}$ corresponds to the k^th matrix composed of $\mathrm{p_k}$ variables. As in standard PLS-DA, we also consider a categorical variable C having q categories. A global latent component t is determined so that it maximizes the between groups variance. Since the set of variables are structured into blocks, local latent components (also called block components) associated to the various blocks are computed such that: t is a linear combination of these latter ones and each block component is a linear combination of the variables belonging to its block.

First, a global component t_1 is determined so that the between groups variance, $var(P_Y t_1)$, is maximized. Let us define by $t_1^{(k)}$ the block component associated to the block $\mathrm{X_k}$: $t_1^{(k)} = X_k w_1^{(k)}$ with $\left\|w_1^{(k)}\right\| = 1, \quad k = 1{:}K$.

Moreover, the global component t_1 is a linear combination of the block components: $t_1 = \sum_{k=1}^{K} a_{1k} t_1^{(k)}$, where a_1 is of length 1.

Since P_Y is symmetric and idempotent, the criterion can be rewritten: $cov(P_Y t_1,\ t_1)$, which in turn leads to:

$$\sum_{k=1}^{K} a_{1k} cov(P_Y t_1\ , t_1^{(k)}) = \frac{1}{n} \sum_{k=1}^{K} a_{1k} {w_1^{(k)}}^{\mathrm{T}} X_k^{\mathrm{T}} P_Y t_1$$

For a global component t_1 being fixed, the vector of loadings $w_1^{(k)} = \left(w_{11}^{(k)},\ w_{1j}^{(k)}, \dots, w_{1p_j}^{(k)}\right)^{\mathrm{T}}$ is determined by the following equation:

$$w_1^{(k)} = \frac{X_k^{\mathrm{T}} P_Y t_1}{\left\| X_k^{\mathrm{T}} P_Y t_1 \right\|}$$

Having set $w_1^{(k)}$, k=1, …, K, the vector a_1 is updated with: $\mathrm{a}_{1k} = \frac{{w_1^{(k)}}^{\mathrm{T}} {X_k}^{\mathrm{T}} P_Y t_1}{\sqrt{\sum_{k=1}^{K} \left({t_1}^T P_Y X_{\mathrm{k}} w_1^{(k)} \right)^2}}$.

Finally, the global component is updated: $t_1 = \sum_{k=1}^{K} a_{1k} t_1^{(k)}$.

The algorithm is depicted below:

a) initialisation of t_1

b) $w_1^{(k)} = \frac{X_k^{\mathrm{T}} P_Y t_1}{\left\| X_k^{\mathrm{T}} P_Y t_1 \right\|}$

c) $\mathrm{a}_{1k} = \frac{{w_1^{(k)}}^{\mathrm{T}} {X_k}^{\mathrm{T}} P_Y t_1}{\sqrt{\sum_{k=1}^{K} \left({t_1}^T P_Y X_{\mathrm{k}} w_1^{(k)} \right)^2}}$.

d) $t_1 = \sum_{k=1}^{K} a_{1k} t_1^{(k)}$

e) update of the criterion: $var(P_Y t_1)$,

Steps (b) to (e) are iterated until a negligible variation of the criterion is observed between two iterations. The subsequent components are determined in the same way after a deflation step.

3. Application to metabolomics

This method is illustrated on a case study pertaining to the LACATOL project (registered to the French Clinical Trial under N° NCT01493063) which aims at ensuring the optimal growth of preterm newborns through a personalized nutrition under the concept of "nutritional footprint" developed by Barker (2004). Specifically, it aims at a better understanding of the impact of perinatal nutrition on the growth of premature babies. The analysis of the composition of breast milk data (Metabolomics, free amino acids, macronutrients) has been performed on a sample of 18 pre-term newborns (gestational ages before 34 weeks of amenorrhea) characterized by an optimal growth (n1=9 newborns) and suboptimal (n2=9 newborns). These two groups have been defined on the basis of the difference between the newborn weight, on their return home, and their expected weight obtained by the intrauterine weight growth curve (Olsen et al, 2010).

During hospitalization, a representative breast milk sample of the last 24 hours has been taken each week. The lipidomic fraction has been analyzed through metabolomics analytical tools (namely HRLC-MS). In parallel, free amino acids and macronutrients have also been measured leading to three blocks of data associated to the samples. The metabolomics data consisting in 3280 metabolites have been first processed in order to select the most discriminant metabolites on the basis of the two groups (optimal growth / suboptimal growth). 141 out of the 3280 metabolites have been selected according to their baby's growth discrimination ratio.
The first two global components of MB-PLS-DA provide a clear separation of the two groups while we can notice a relatively poor associations between the different variables belonging to the three blocks. This is confirmed by the analysis of the contributions $\left(\mathrm{a}_{1k}^2, \mathrm{a}_{2k}^2\right)$ of each block component to the global components t_1 and t_2. Dealing with the first global component, the contribution of the metabolomics block (contribution equaling 74%) outscores the contribution of the free amino acids and macronutrients while for the second component, we can notice a major contribution of the macronutrients block (contribution equaling 55%).

4. Conclusion

The results, obtained herein, can be considered as a synthesis of PLS-DA performed separately on each data block. An interesting aspect of the method lies in the determination of weights associated to each block reflecting its contribution to the separation of the groups. The biological interpretation of the results was performed leading to new prospects for nutritional interventions of breastfeeding mother in a context of prematurity

References

Barker, D. J. P. (2004). The developmental origins of adult disease. *Journal of the American College of Nutrition*, *23*(sup6), 588S-595S.

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of chemometrics*, *17*(3), 166-173.

Biancolillo, A., Måge, I., & Næs, T. (2015). Combining SO-PLS and linear discriminant analysis for multi-block classification. *Chemometrics and Intelligent Laboratory Systems*, *141*, 58-67.

Boccard, J., & Rutledge, D. N. (2013). A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock Omics data fusion. *Analytica chimica acta*, *769*, 30-39.

Nocairi, H., Qannari, E. M., Vigneau, E., & Bertrand, D. (2005). Discrimination on latent components with respect to patterns. Application to multicollinear data. *Computational statistics & data analysis*, *48*(1), 139-147.

Olsen, I. E., Groveman, S. A., Lawson, M. L., Clark, R. H., & Zemel, B. S. (2010). New intrauterine growth curves based on United States data. *Pediatrics*, peds-2009.

Skov, T., Honoré, A. H., Jensen, H. M., Næs, T., & Engelsen, S. B. (2014). Chemometrics in foodomics: handling data structures from multiple analytical platforms. *TrAC Trends in Analytical Chemistry*, *60*, 71-79