A sparse method to handle two high dimensional symetric data sets

Kim-Anh Lê Cao[1,2], Pascal G.P. Martin[3], Christèle Robert-Granié[2] and Philippe Besse[1]

[1] Institut de Mathmatiques, UMR 5219, Université de Toulouse, France
[2] Institut National de la Recherche Agronomique, UR631 Toulouse, France
[3] Institut National de la Recherche Agronomique, UR 66, Toulouse, France

When dealing with high dimensional biological data, one important issue is to handle the $n << p$ problem, as most variables are irrelevant or noisy to explain the biological experiment. This issue is even more challenging when there are more than one group of variables and when the aim is to highlight the relationships between the different sets of variables. Here, we especially focus on the situation where there are two groups of variables measured on the same observations (symetric data sets).
Very few methods can deal with two (or more) data sets. Among them, Canonical Correlation Analysis (CCA, [2]) and Partial Least Squares regression (PLS, [3]) can answer the biological question. However, both approaches do not allow feature selection. To deal with the major drawback of CCA ($p < n$ and $q < n$), [1] proposed to regularize CCA with an L_2 penalization. But so far, no sparse method has been developped yet to handle this problem, as it is proposed by lasso in the context of regression.
We propose a sparse approach to select variables from each of the two data sets with two variants: either to model the relationships between the two sets of variables, or to predict one set of variables with respect to the other. We apply this method to some real world data sets with two types of measurements, and which purpose is to explain which group of variable imply the other group of variables (and/or vice versa).

References

[1] I. Gonzalez, S. Déjean, P.G.P. Martin, O. Gonçalves, P. Besse, A. Baccini (2007) Highlighting relationships through Regularized canonical Correlation analysis: application to high throughput biology data, *Journal of Biostatistics*, to appear.

[2] H. Hotelling (1936) Relations bewteen two sets of variates, *Biometrika*, **28**, 321-377

[3] H. Wold (1966) Estimation of principal components and related models by iterative least squares. In: Krishnaaiah, P.R. (editors), *Multivariate Analysis*. New York: Academic Press, pp.391-420