Variable selection in multivariate linear models with high-dimensional covariance matrix estimation

Marie Perrot-Dockès*†[1], Céline Lévy-Leduc[1], Laure Sansonnet[1], and Julien Chiquet[1]

[1]UMR MIA-Paris – AgroParisTech, INRA - Université Paris-Saclay – France

Abstract

We propose a novel variable selection approach in the framework of multivariate linear models taking into account the dependence that may exist between the responses. It consists in estimating beforehand the covariance matrix Σ of the responses and to plug this estimator in a Lasso criterion, in order to obtain a sparse estimator of the coefficient matrix. The properties of our approach are investigated both from a theoretical and a numerical point of view. More precisely, we give general conditions that the estimators of the covariance matrix and its inverse have to satisfy in order to recover the positions of the null and non null entries of the coefficient matrix when the size of Σ is not fixed and can tend to infinity. We prove that these conditions are satisfied in some particular case. Our approach is implemented in the R package MultiVarSel available from the Comprehensive R Archive Network (CRAN) and is very attractive since it benefits from a low computational load. We also assess the performance of our methodology using synthetic data and compare it with alternative approaches. Our numerical experiments show that including the estimation of the covariance matrix in the Lasso criterion dramatically improves the variable selection performance in many cases. Eventually we successfully applied our methodology to a LC-MS (Liquid Chromatography-Mass Spectrometry) data set made of 30 copals samples in which we measured 1019 metabolites. The samples were divided into three groups, according to their genera and their geographical provenance. Our methodology allowed us to identify the most important metabolites for distinguishing the different groups.

*Speaker

†Corresponding author: marie.perrot-dockes@agroparistech.fr