Accepted Manuscript

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

Marie Perrot-Dockès, Céline Lévy-Leduc, Laure Sansonnet, Julien Chiquet

PII: S0047-259X(17)30425-6
DOI: https://doi.org/10.1016/j.jmva.2018.02.006
Reference: YJMVA 4328

To appear in: *Journal of Multivariate Analysis*

Received date : 13 July 2017

Please cite this article as: M. Perrot-Dockès, C. Lévy-Leduc, L. Sansonnet, J. Chiquet, Variable selection in multivariate linear models with high-dimensional covariance matrix estimation, *Journal of Multivariate Analysis* (2018), https://doi.org/10.1016/j.jmva.2018.02.006

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

Marie Perrot-Dockès[a,*], Céline Lévy-Leduc[a], Laure Sansonnet[a], Julien Chiquet[a]

[a] *UMR MIA-Paris, AgroParisTech, INRA - Université Paris-Saclay, Paris 75005 France*

Abstract

In this paper, 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 the particular case of some Toeplitz matrices. 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.

Keywords: High-dimensional covariance matrix estimation, Lasso, Multivariate linear model, Variable selection.

1. Introduction

The multivariate linear model consists in generalizing the classical linear model, in which a single response is explained by p variables, to the case where the number q of responses is larger than 1. Such a general modeling can be used in a wide variety of applications ranging from econometrics [9] to bioinformatics [11]. In the latter field, for instance, multivariate models have been used to gain insight into complex biological mechanisms like metabolism or gene regulation. This has been made possible thanks to recently developed sequencing technologies. For further details, we refer the reader to [10]. However, the downside of such a technological expansion is to include irrelevant variables in the statistical models. To circumvent this, devising efficient variable selection approaches in the multivariate setting has become a growing concern.

A first naive approach to deal with the variable selection issue in the multivariate setting consists in applying classical univariate variable selection strategies to each response separately. Some well-known variable selection methods include the least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani [16] and the smoothly clipped absolute deviation (SCAD) approach devised by Fan and Li [5]. However, such a strategy does not take into account the dependence that may exist between the different responses.

In this paper, we shall consider the following multivariate linear model:

$$Y = XB + E, \tag{1}$$

*Corresponding author

Email address: `marie.perrot-dockes@agroparistech.fr` (Marie Perrot-Dockès)