Introduction to MultiVarSel

A. Marie Perrot-Dockes[a] and **B. Céline Lévy-Leduc**[a] and **C. Julien Chiquet**[a]

[a]MIA
A-AgroParisTEch
A-16 rue Claude Bernard 75005 Paris
marie.perrot-dockes@agroparistech.fr

Mots clefs : Statistics, Multivariate Regression, LASSO.

1 Introduction

This presentation explains basic usage of **MultiVarSel**, an R package to perform variable selection in the multivariate linear model taking into account the dependence that may exist between the responses. This package foccus on a model that can be described as follows :

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

where Y is a $n \times q$ matrix of responses, X is a $n \times p$ matrix of covariables, B is a $p \times q$ sparse matrix of coefficients and E is a random error matrix such that $\forall i \in \{1, \cdots, n\}$, $E_i = (E_{i,1}, \ldots, E_{i,q}) \overset{iid}{\sim} \mathcal{N}(0, \Sigma)$. The package consists in estimating Σ beforehand and to plug this estimator in a Lasso criterion, in order to obtain a sparse estimator of the coefficient matrix B.

2 Whitening test

To apply the methodology we start by estimating the matrix E which is obtained by fitting a linear model with the design matrix X to all the columns of Y as if they were independent:
We then use a Portmanteau test to check if each row of this matrix $\widehat{E}$ is a white noise or not. If the p-value is smaller than 0.05 we reject the hypothesis that each row of the residuals matrix is a white noise and we will try to estimate the dependance that exist among the collumns.

3 Whitening

We then try to remove the dependence among the columns of the residuals matrix by estimating the covariance matrix of the rows of E. To estimate it we try different structures for this covariance. The simplest assumption, proposed in the package **MultiVarSel** is that each row of E follows an AR(1) process, it also propose a modelling where each row is an ARMA(p,q) process and a nonparametric one where Σ is only assumed to be Toeplitz. To compare these different dependence modellings we perform a Portmanteau test on the "whitened" matrix $\widehat{E}\widehat{\Sigma}^{-1/2}$, where $\widehat{\Sigma}^{-1/2}$ is the square root of the inverse of the estimation of Σ.
We then select the simplest model that allows us to remove the dependence in the data. Using this model we compute the square root of the inverse of the estimator of the covariance matrix of each row of the residuals matrix.

In order to whiten the data (remove the dependence), we transform the data as follows:

$$\boldsymbol{Y}\widehat{\boldsymbol{\Sigma}}^{-1/2} = \boldsymbol{X}\boldsymbol{B}\widehat{\boldsymbol{\Sigma}}^{-1/2} + \boldsymbol{E}\widehat{\boldsymbol{\Sigma}}^{-1/2}. \tag{2}$$

The idea is then to use the Lasso criterion introduced by Tibshirani in 1996, and available in the R package `glmnet` on these whitened data. We recall that in the classical linear model

$$\mathcal{Y} = \mathcal{X}\mathcal{B} + \mathcal{E},$$

where $\mathcal{Y}$, $\mathcal{B}$ and $\mathcal{E}$ are vectors and $\mathcal{X}$ is a matrix, the Lasso estimator of $\mathcal{B}$ is defined by

$$\widehat{\mathcal{B}}(\lambda) = \text{Argmin}_{\mathcal{B}} \left\{ \|\mathcal{Y} - \mathcal{X}\mathcal{B}\|_2^2 + \lambda \|\mathcal{B}\|_1 \right\}.$$

In order to be able to use the Lasso criterion we will apply the *vec* operator to (2)

$$\begin{aligned}
\mathcal{Y} &= vec(\boldsymbol{Y}\widehat{\boldsymbol{\Sigma}}^{-1/2}) = vec(\boldsymbol{X}\boldsymbol{B}\widehat{\boldsymbol{\Sigma}}^{-1/2}) + vec(\boldsymbol{E}\widehat{\boldsymbol{\Sigma}}^{-1/2}) \\
&= ((\widehat{\boldsymbol{\Sigma}}^{-1/2})' \otimes \boldsymbol{X}) vec(\boldsymbol{B}) + vec(\boldsymbol{E}\widehat{\boldsymbol{\Sigma}}^{-1/2}) \\
&= \mathcal{X}\mathcal{B} + \mathcal{E}.
\end{aligned}$$

4 Variable selection

The Lasso criterion applied to $\mathcal{Y} = vec(\boldsymbol{Y}\widehat{\boldsymbol{\Sigma}}^{-1/2})$ will provide an estimation of the non null positions of $\mathcal{B} = vec(\boldsymbol{B})$ and hence the non null positions of B. In order to avoid false positive positions we add a stability selection step. These different steps (whitening, vectorization, Lasso, stability selection) are implemented in the function `variable_selection` of the R package `MultiVarSel`.

Références
[1] M. Perrot-DockÃ¨s et al. A multivariate variable selection approach for analyzing LC-MS metabolomics data,arXiv:1704.00076.