# Automatic variable selection algorithms

## Statistical and Machine Learning on simulated Data sets

Abdoul-Aziz Berrada - Cécile Brissard
Morgane Caillosse - Hugo Hamon

master 1 econometrie-statistiques

May 2021

# Table of Contents

# The next section is...

# Introduction

- One major and key stake :
  ➢ parsimonious models (not too many variables with the best possible performance)
- Statistical Learning :
  ➢ Backward, Forward and Stepwise
- Machine Learning :
  ➢ LAR, LASSO and Elasticnet

- We built 3 Data Generating Processes (DGP)
- 1 000 data sets in each
- Each data set with 100 observations and 31 variables (one $Y$ and 30 $X$s)

# The next section is...

# DGP1

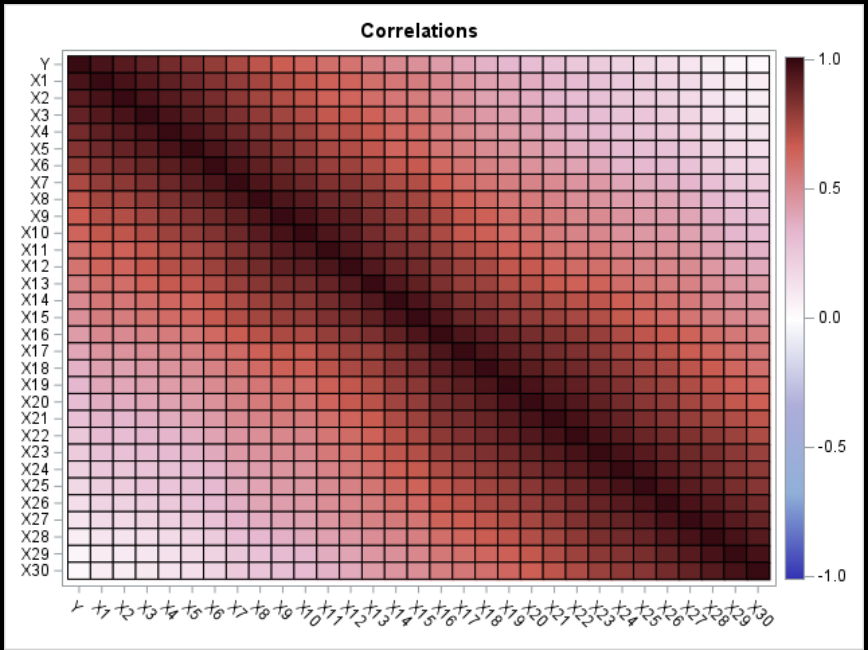
## Method (1/2)

- We need to know the structure of the correlation between variable to determine if the models selected the right variables.
- We draw graphics of the frequencies that variables have been selected, individually and jointly to determine the performance of each model (and on each DGP).

# DGP1

## Method (2/2)

### Correlation heatmap for DGP1



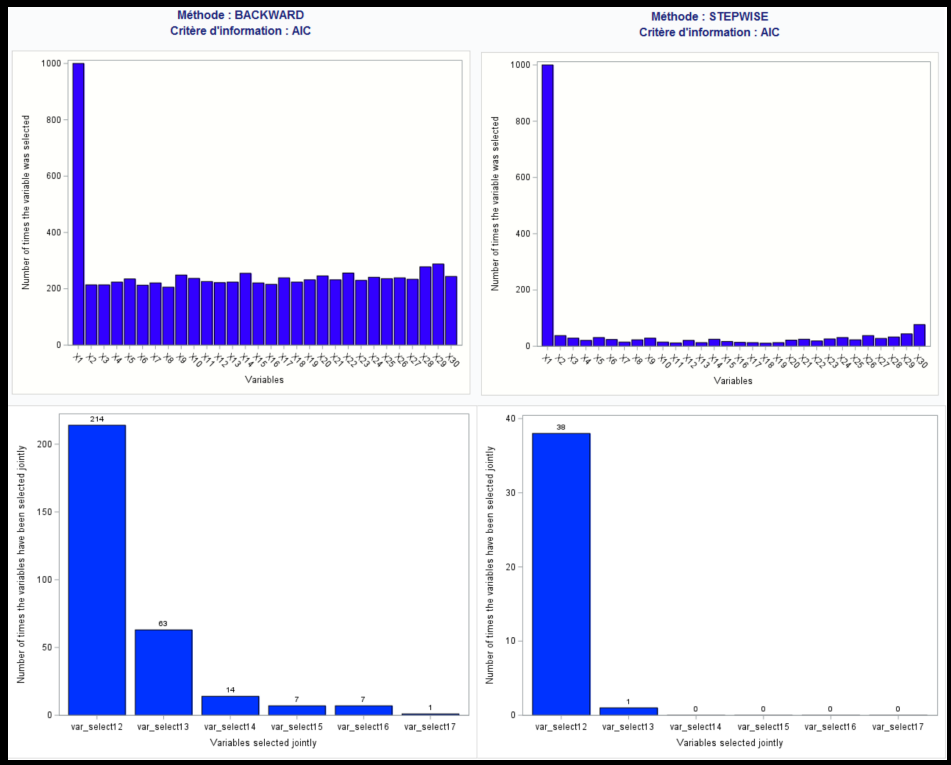

- Very high correlations between $Y$ and variables from $X_1$ to $X_{10}$
- Low correlations between $Y$ and $X_{20}$ to $X_{30}$
- Each variable is highly correlated with the closest ones in the matrix

# DGP1

## Results (1/2)

Statistical Learning

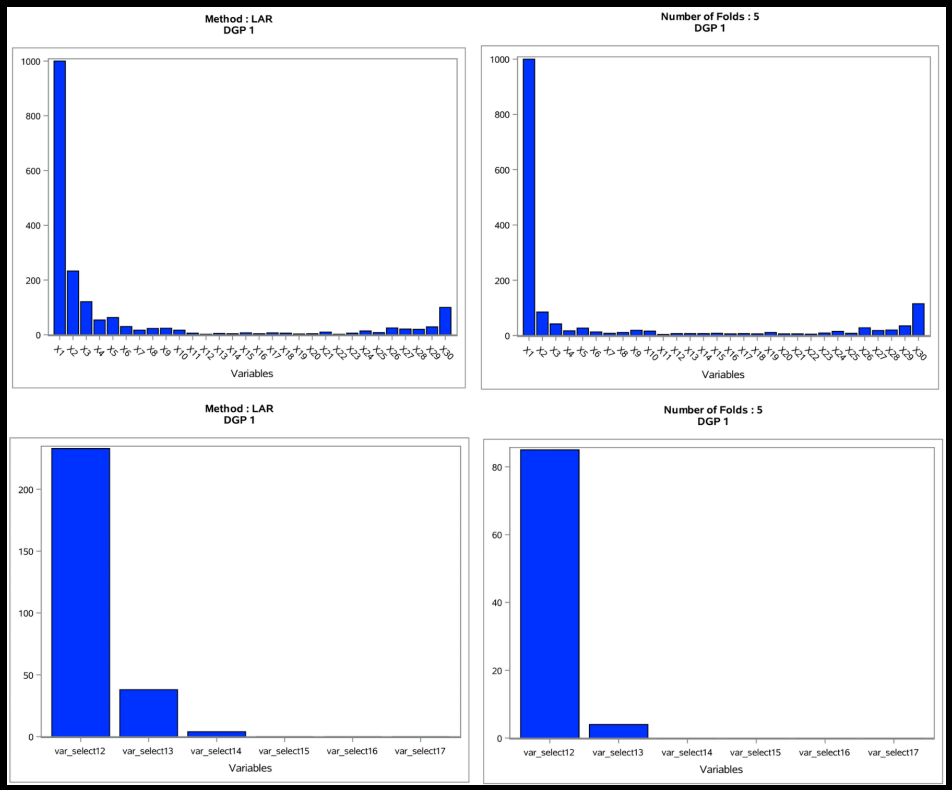
- Probabilities : how many times among a 1 000?
- $X_1$ picked at 100%
- $X_2 \rightarrow$ var_select12
- Stepwise :
  - ⪰ $X_{30}$ freq. $= 10\%$
- Backward :
  - ⪰ $X_2$ to $X_{30}$ freq. are $\frac{20}{100}$
  - ⪰ Until $7_{th}$ $X$s



# DGP1

## Results (2/2)

### Machine Learning




- ElasticNet : an improvment for LASSO's algorithm.
- 10-fold : too restrictive
- DGP1 takes Machine and Statistical Learning in default :
  - ⩾ $X_1$ correlation's to $Y$ is 90% .
  - ⩾ $X_{30}$ 'the farest from $X_1$' source of explanatory power.

# The next section is...

# DGP2

## Method (1/2)

- New data with the same process (ie. Toeplitz correlation Matrix and Multivariate Normal Distribution)
- Less importance to the first ten variables
- Compared to the DGP1, the correlation between $X_1$ and $Y$ is only 0.7.

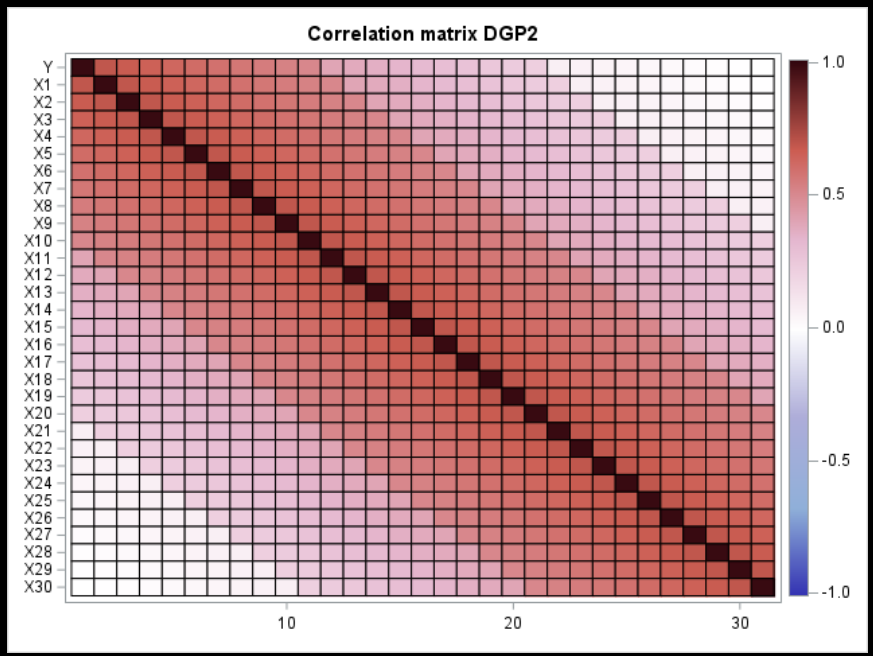
# DGP2

## Method(2/2)

Correlation coefficient between variables and $Y$ :

- [0.7, 0.5[ for the first ten
- [0.5, 0.2[ for the 10 following
- [0.2, 0.001] for the last ten

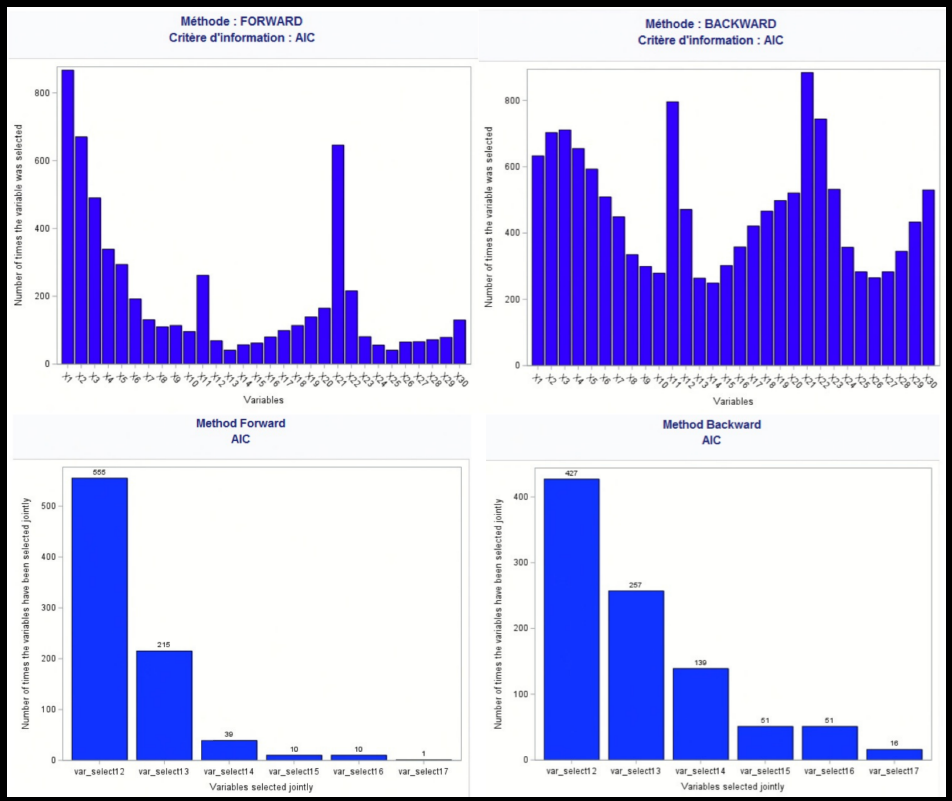
### Correlation Matrix for DGP2.



# DGP2

## Results (1/2)






- Individual probabilities of selection :
  - ⪢ Same results for *Forward* and *Stepwise*
  - ⪢ Different results for the *Backward* and those are slightly better when choosing AIC
- Joint probabilities of selection :
  - ⪢ With Forward the prob of $X_1$- $X_4$ being chosen is 0.39% and only 20% with stepwise
  - ⪢ *Backward* algorithm selects more often than *Forward* and *Stepwise*

# DGP2

## Results (2/2)

### Machine Learning

- Individual probabilities of selection :

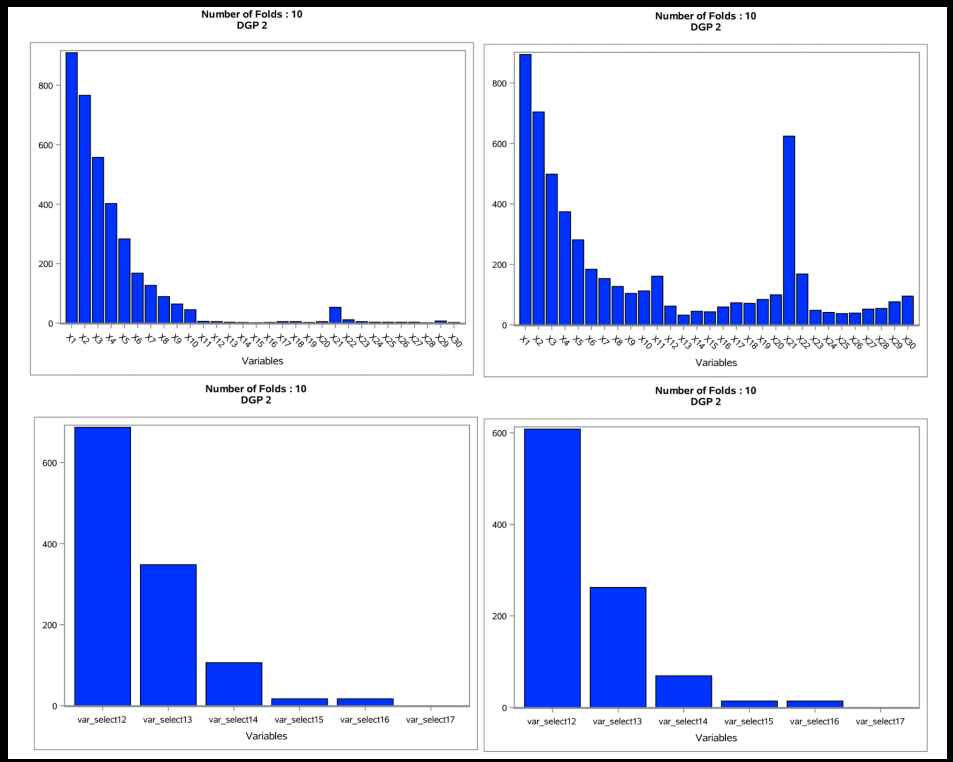
  ➢ Lasso : decrease of the variable selection frequency between the first and the $10^{th}$ and the next more often selected variable is $X_{21}$

  ➢ *ElasticNet* : no divergence between the number of folds (3, 5 et 10) but it's selection frequency is higher than Lasso's one

- Joint probabilities of selection :

  ➢ Lasso : Big focus in favor of the first two variables

  ➢ *ElasticNet* provides sensitively the same probabilities



# The next section is...

# DGP3

## Method

Purpose/point :

- Only 3 variables significantly correlated to $Y$ ;
- No significant correlation between variables.

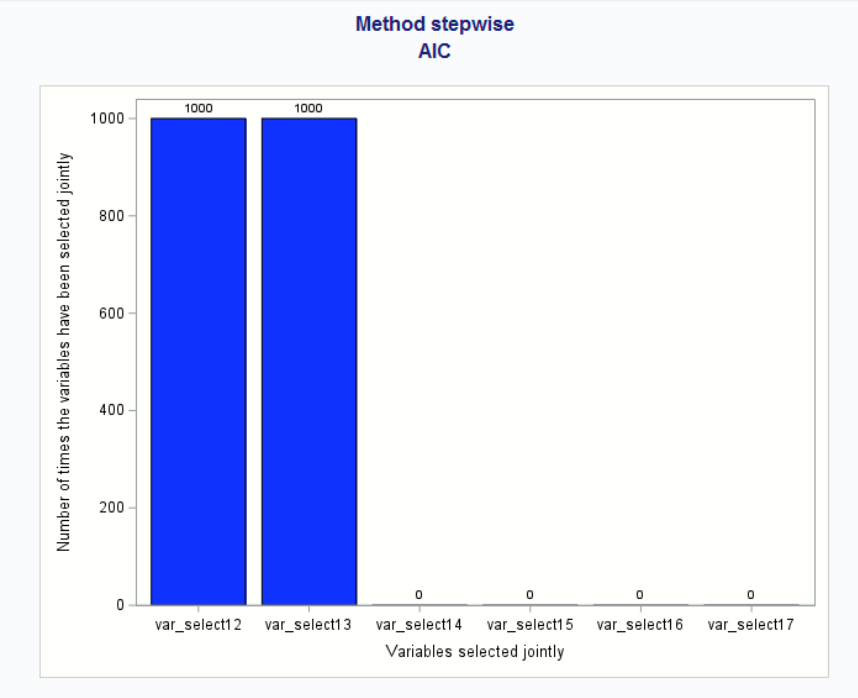
To obtain our third DGP we have:

- Simulated 1000 data sets of 30 variables using $X \sim \mathcal{MVN}(\overrightarrow{0}, I_{30})$ ;
- Generated $Y$ according to :

$$Y = 0.6777 \times X_1 + 0.414 \times X_2 - 0.5814 \times X_3 \tag{1}$$

# DGP3

## Results




Results were the same for Statistical Learning and Machine Learning :

- Only $X_1$, $X_2$ and $X_3$ are systematically selected by each algorithms ;
- Without multi-correlation SL is able to find which variable is correlated with the target.

# The next section is...

# Discussion

3 main results :

- SL does not perform well when correlation between one variable and $Y$ is too high (DGP1) ;
- ML is more robust than SL when considering multi-correlations between variables in the data set (DGP2) ;
- ML and SL provide the same results when significant correlations are attributed to few variables (DGP3).