3 Method

This paper aims to test the performance of the variable selection algorithms presented in the previous sections. To do so, we need to generate data sets with a known and precise correlation structure (manually parameterized). Indeed, in order to determine whether or not an algorithm performs effectively, it is necessary to ensure that the data we work on follows a precise correlation structure. Thus, we created 1,000 databases from 4 DGPs (Data Generating Processes), for a total of 4,000 databases. Each database contains a set of 51 variables (1 explained variable Y and 50 explanatory variables X1 - X50) and 100 observations. Finally, we test each algorithm on the 4,000 datasets and conclude on their results.

3.3 Data Generating Processes (DGP)

(Each DGP is made up of 1,000 databases, themselves composed of 100 rows (sample size) and 51 columns (number of variables), in which the first column is column Y and is followed by columns X1 to X50.)

In each DGP, Y based on the following equation : Y = 1.5 · X1 + 0.9 · X2 + 1 · X3 + 0.1 · X4 − 0.5 · X5, and all datasets (which is the X matrix) of 50 variables using the multivariate normal distribution.

For DGP1, we generated a classical dataset representing the perfect model, without the presence of correlation between variables, and without extreme values.

For the second DGP, we generated a dataset in which only the variables X1 to X5 are positively correlated with each other, whereas the variables X6 to X50 are not.

For the third DGP, we generated a dataset in which the variables X1 to X50 are positively correlated with each other.

For the DGP4, we generated a data set in which the variables X1 to X50 are not correlated but have extreme values.

3.3.1 First Data Generating Process (DGP1)

For DGP1, we generated a classical dataset representing the perfect model, without the presence of correlation between variables, and without extreme values. For the creation of a database, we simulated a dataset (the X matrix) of 50 variables using the identity matrix as the variance-covariance matrix:

A picture containing text, different

Description automatically generated

Finally, we joined the column Y with the columns from X1 to X50, to create a database with 51 columns. The value of Y is dependent on the randomly obtained values of the variables X1 to X5. We could interpret the relationship between Y and the variables X1 to X5 as positive and negative correlations, and the relationship between Y and the variables X6 to X50 as zero correlations.

3.3.2 Second Data Generating Process (DGP2)

For the second DGP, we generated a dataset in which only the variables X1 to X5 are positively correlated with each other, whereas the variables X6 to X50 are not. To create this database, we simulated a first data set (the Xa matrix) of 5 variables (X1 to X5) following a multivariate normal distribution (mean = 0), and with positive and symmetric correlation between the first 5 variables. The correlation matrix for the variables X1 to X5 was done using the Toeplitz matrix. We obtain a variable X1 which is weakly correlated to X2, moderately correlated to X3 and X4, and strongly correlated to X5 (COV(X1,X5) = 0.714).

Table

Description automatically generated

Then, we simulated a second data set (the Xb matrix) of 45 variables (X6 to X50) following a multivariate normal distribution (mean = 0) and zero correlation between these variables. For the correlation matrix, we used the identity matrix as in DGP1. After simulating the two datasets, we created a matrix X that joins the two datasets, in order to have only one data table. Then, we generated Y with the same equation in DGP1.

Finally, we joined the Y column to the X matrix, in order to create a database with 51 columns in which only Y is correlated to the first 5 variables, which are themselves correlated with each other. The variable Y is not correlated to the variables X6 to X50, and they are themselves not correlated with each other.

3.3.3 Third Data Generating Process (DGP3)

For the third DGP, we generated a dataset in which the variables X1 to X50 are positively correlated with each other. To create the database, we simulated a dataset (X) of 50 variables (X1 to X50) following a multivariate normal distribution with mean 0, and positive and symmetric correlation between the 50 variables. The correlation matrix for the variables X1 to X50 was done using the Toeplitz matrix. Each variable is highly correlated with the closest variables in the matrix. The further away a variable is from another, the less correlated it is. For example, we obtain a variable X1 that is highly correlated with the variables X2 to X16, moderately correlated with the variables X17 to X36, and weakly correlated with the variables X37 to X50.

Table

Description automatically generated

Then we generate Y according to the same equation in DGP1. Finally, we joined the column Y to the matrix X, in order to create a database with 51 columns in which Y is the result of the random generation of the first 5 variables. The variables X1 to X50 are correlated with each other in a positive and symmetric way.

3.3.4 Fourth Data Generating Process (DGP4)

For the DGP4, we generated a data set in which the variables X1 to X50 are not correlated but have extreme values. For the creation of the database, we simulated a dataset (the X matrix) of 50 variables using the multivariate normal distribution with extreme values and used the identity matrix as the variance covariance matrix. Then, we generate Y via the same equation in DGP1. Finally, we joined the column Y to the matrix X, in order to create a database with 51 columns in which Y is the result of the random generation of the first 5 variables. The variables X1 to X50 have extreme values without correlation.

A screenshot of a computer

Description automatically generated with low confidence

3.4 Results’ extraction

In each DGP, we evaluate the performance of each method with several selection criteria by calculating the probability that the selection procedure used selects only the 5 variables of interest (which are variables X1 to X5).

We have for each stated procedure () two histograms, one represents the probability that a variable is selected among the 1,000 regressions ; and the other represents the probability that the 5 variables of interest are uniquely selected among the 1,000 regressions.

When the selection procedure is run for a ”generated data table”, the selected variables are assigned a correlation coefficient. By extension, the variables without an assigned coefficient are the ”unselected” variables, they appear as a missing value in the coefficient results, these variables are automatically removed. Then we find a new table that we will call ”selection table” with only the columns of the selected variables. For example, if in a generated data table, the method used selects only the variables X1 to X10, then the selection table will be composed of the columns from X1 to X10.

In order to compare the different selection methods[16], we first compare the individual selection frequencies of each variable according to the different data sets. And in a second step, we compare the selection frequencies of the 5 variables of interest only according to the different data sets. For the individual frequency, we have created a table that we will call the ”selection table in 1”, allowing to gather all the selection tables. Then, we replaced each coefficient of the ”selection table in 1” by 1 and each missing value by a 0. The resulting matrix will be used to build a bar chart that gives the frequency of each variable that has been chosen individually. The frequency of selection of each variable obtained is divided by 1,000 to obtain the frequency of selection in percentages, so we can use them to express the likelihood of variable selection on a scale ranging from 0 to 100.

For the frequency of selection of the 5 variables of interest only, we build a new database consisting of 0 and 1, which assigns 1 when these 5 variables are selected from the same dataset ; 0 if at least one variable is not selected with the others. For example, we define stepwise12345, as the joint selection of X1, X2, X3, X4, and X5 with the Stepwise method : if this condition is fulfilled, the value 1 is written in the column, otherwise 0. Once this is done, we just need to count the number of 1’s among the 1,000 databases and divide the frequency by 1,000 to get the results in percentages. If the joint frequency is 0/100, this means that X1, X2, X3, X4, and X5 are never selected together, whereas if the joint frequency is 100/100, then these variables of interest are joined each time.

We will have for each stated procedure (Stepwise, Backward, Forward, LARS, LASSO, Elastic Net) two histograms, one represents the probability that a variable is selected among the 1,000 regressions, and the other represents the probability that the 5 variables of interest are uniquely selected among the 1,000 regressions.

4 Results

Now that we have the graphs for each model’s probability of selecting a variable, as well as the joint probabilities of selecting multiple variables at the same time, we will summarize them. Indeed, we studied several selection criteria for each approach, including SBC, AIC, and AICc. However, we will only provide the most interesting criteria and method for each DGP in this part. All the criteria can be found in our code SAS. Finally, we will compare the most relevant method between Statistical Learning and Machine Learning.

Conclusion of results :

After examining all of the selection methods in various conditions, it can be concluded that in DGP1, DGP2, and DGP4 type situations, Machine Learning techniques are more successful than Statistical Learning methods. Statistical Learning techniques are more successful than Machine Learning methods only in the scenario of DGP3. If we compare all of the outcomes, we can conclude that LASSO and ElasticNet are the best approaches.

We can see from these results that the LASSO and ElasticNet approaches are limited in some instances. First and foremost, we know that due to the nature of the coefficient restriction, LASSO will prefer to choose a subset of variables, but it will also tend to under-fit. While LASSO is more likely to avoid presenting variables that are of no interest, it is also more likely to overlook some factors that are of interest. Furthermore, if two variables are highly associated and critical for prediction, LASSO will favor one over the other, as in the case of DGP2. When the variables of interest are associated with other variables, correlations might become an issue. The consistency of the LASSO selection is no longer guaranteed in this scenario.

ElasticNet uses a weighted combination of L1 and L2 regularizations. As you can probably see, the same function is used for both LASSO and Ridge regression, only the L1wt argument changes. This argument determines how much weight goes to the L1 norm of the partial slopes. If the regularization is pure L2 (Ridge) and if L1wt = 1.0 the regularization is pure L1 (LASSO). ElasticNet allows to add to LASSO a penalty Ridge, in order to reduce the selection bias of LASSO, ElasticNet thus finds better results compared to LASSO in DGP2 and DGP3 type situations.

Attention, les méthodes de ML ont du mal à détecter les variables qui ont un faible coeff multi, cas où X4 n’est pas sélectionné.

Nuance :

On a remarqué que les méthodes de statistical learning ont plus de chances de retrouver les variables d'intérêt dans le DGP 3, mais ils ont aussi plus de chance de retrouver des variables supplémentaires et donc plus de chance de proposer des modèles avec des erreurs de sélection. Ainsi il y aura plus de chances que la méthode propose un modèle composé de toutes les variables d'intérêts mais avec des variables d'erreurs.

tandis que les méthodes de machine learning, ont moins de chances de retrouver les variables d'intérêt dans le DGP 3, mais ont aussi moins de chances de retrouver des variables supplémentaires et donc moins de chance de proposer des modèles avec des erreurs de sélection. Ainsi il y aura plus de chances que la méthode propose un modèle seulement composé d’une partie des variables d'intérêts mais sans variables d'erreurs.

( In each DGP, we evaluate the performance of each method with several selection criteria by calculating the probability that the selection procedure used selects only the 5 variables of interest (which are variables X1 to X5).

We have for each stated procedure () two histograms, one represents the probability that a variable is selected among the 1,000 regressions ; and the other represents the probability that the 5 variables of interest are uniquely selected among the 1,000 regressions.)

(nous remarquons que les méthodes de statistical learning ont plus de chances de proposer un modèle composé de toutes les variables d'intérêts mais avec des variables d'erreurs dû à l’under fitting. Tandis que les méthodes de machine learning ont plus de chances de proposer un modèle seulement composé d’une partie des variables d'intérêts mais sans variables d'erreurs dû à l’over fitting.)

(c’est mieux d’avoir un modèle avec peu de varirables d’intérêt mais dans lequel on sait que se sont toutes des variables d’intérêts, plutôt qu’un modèles composé de toutes les variables d’intérêts avec des variables supplémentaires mais dans lequel on ne connait pas réellement les variables d’intérêts.)

Soutenance SAS:

1. Introduction

Good evening everyone, to begin with, our presentation's overall structure is divided into 4 sections : first, I will introduce you the topic and explain why we chose it. Secondly, Romain will discuss the methods used and the results obtained. Then, Claire will give a brief overview of the findings from the diabetic database application. Finally, I will conclude with a discussion of the paper’s results.

[INTRODUCTION]

Dealing with high-dimensional data leads to some issues and to overcome them, we need to find the most efficient algorithm to select only the relevant explanatory variables of a model among a large number of variables. To this purpose, there are several variable selection approaches and they are classified into two main categories : Statistical Learning and Machine Learning. Statistical Learning refers to a set of tools for modeling and understanding complex datasets whereas Machine Learning is defined as “the study of computer algorithms that allow computer programs to automatically improve through experience”.

We quickly and unanimously agreed on this topic because it was the one we found the most interesting and challenging. Machine Learning is becoming a huge concern for companies, and understanding how it works and its limitations gives us a significant competitive advantage in the future.

Intro: How to test the performance ?

First of all, we will see “methods used” and “the results obtained” in this part.

To test the performance of “the variable selection algorithms”, we generate “datasets” with a “precise correlation structure” for each “Data Generating Process” (which is called DGP). Thus, we created “1,000 databases” from 4 DGPs. Each database contains “100 observations” and “51 variables” in which the first variable is “column Y” and is followed by “columns X1 to X50”. At last, we test each algorithm on the “4,000 datasets” and “conclude on their results”.

3.3 Data Generating Processes (DGP)

In each DGP, Y is based on this equation as you can see on the slide, and “all datasets of 50 variables” using the “multivariate normal distribution”.

For DGP1, we generated a classical dataset representing the perfect model, without “the presence of correlation between variables”, and without “extreme values”.

For the second DGP, we generated a dataset in which “only the variables X1 to X5” are “positively correlated with each other”, whereas “the variables X6 to X50” are not.

For the third DGP, we generated a dataset in which “the variables X1 to X50” are “positively correlated with each other”.

For the DGP4, we generated a “dataset” in which “the variables X1 to X50” are “not correlated” but have “extreme values”.

Conclusion of results :

After examining “all the selection methods in various conditions”, we can conclude that in DGP1, DGP2, and DGP4, Machine Learning techniques “are more successful than” Statistical Learning methods. However, Statistical Learning techniques find the “variables of interest” more easily than Machine Learning methods “only in the scenario DGP3”, but Statistical Learning is not necessarily better.

As we can see in DGP3 case, we noticed that statistical Learning methods are more likely to propose “a model composed of” “all the interests variables” but with “error variables”, due to under-fitting. While machine learning methods are more likely to propose “a model composed of” “only a part of interests variables” but without “error variables”, due to over-fitting.

If we compare “all of the outcomes”, we can conclude that LASSO and ElasticNet are “the best approaches”. “For instance”, we know that “our variables of interest” range from X1 to X5. It is better to have “a model involving only the variables X1 to X3”, but we know that “these are variables of interest”. “Rather than” having “a model with X1 to X5” plus “irrelevant variables” like X10, X20, and X40, but we don’t know which are “ the variables of interest”.