Automatic variables selection: Statistical learning vs Machine learning

ARBOUCH Mounira, JEANNOT Jynaldo, YAKOUB Mélanie
under the supervision of Mr. Philippe DE PERETTI

January 2023

Abstract

Nowadays, data is becoming more and more preponderant in decision making. Due to the constant evolution of big data,the selection of variables becomes very daunting for economists, statisticians and data scientists. Automatic selection methods have been developed to overcome this problem. In this paper,we will review the different methods of variable selection and evaluate their performance on different categories of data. These data will be generated by Monte Carlo simulation under the null hypothesis of independence and the alternative hypothesis of linear dependence. We will contrast statistical learning methods with machine learning methods.

Contents

1 Introduction

In Econometrics, Machine Learning and more broadly in data science, the question of the automatic selection of a set of relevant variables presents an important issue for both the prediction and the explanation of a linear regression model. On a daily basis, data scientists have to deal with a large amount of data, so they have to select the most relevant variables in order to avoid having unnecessarily complex models. The explanatory issue consists in finding the true model, the one that has exactly the useful non-zero variables to explain the variable to be explained. But there is also the phenomenon of overfitting that the study of variable selection allows to limit. Indeed, if the linear regression model is too precise, partly because the number of explanatory variables is too large, the linear regression runs the risk of adapting too much to the data on which it has been estimated and therefore runs the risk of badly predicting new data that it does not know.

Thus, studying, comparing and testing different variable selection methods in different situations, would allow us to obtain the best possible model. In this paper, we will focus on statistical learning and machine learning procedures for variable selection.

Our study aims to determine the best methods of variable selection between statistical learning and machine learning procedures depending on the selection criteria and data types.

To do this, we will make a theoretical presentation of the different selection methods and selection criteria. Then, we will test and compare the performances of these different methods on several types of generated databases : data based on a multivariate normal distribution and data with outliers and correlation between variables.

2 Litterature overview : statistical learning and machine learning methods

There are artificially two categories of variable selection methods. We can differentiate the variable selection procedures of Statistical learning with inference, that is to say, which use a statistical criterion (T-stat, F-test, AIC,...) from the procedures of machine learning without inference.
Stepwise Statistical Learning procedures are based on parameter significance tests, while Machine Learning methods such as LASSO use procedures that shrink coefficients with little economic influence. This explains the main differences between these two categories of procedures. A significant coefficient due to a low standard deviation, for example, does not necessarily have significant economic weight.

2.1 Statistical learning

Statistical learning theory [3]is a machine learning framework drawn from the fields of statistics and functional analysis. It is not only a theoretical analysis tool, but also a tool for creating algorithms that use statistical inference to estimate multidimensional functions.
We will focus on Statistical learning procedures for variable selection by stepwise and more particularly on the forward or backward methods that compose it.
The stepwise selection method is a simplified alternative to the exhaustive search method when the sample size n and the number of variables p are very large.

2.1.1 Backward stepwise selection

The step by step backward selection is a variable selection procedure which consists in starting from a complete model containing all the potential explanatory variables p, and to go down the number of variables step by step. In the first step, we consider k models (for $k = p, p-1, ...1$) with k-1 predictors. This method removes at each step the least significant variable based on a pre-specified statistical criterion. At each iteration a new model is then estimated with the remaining variables. The variables are eliminated one after the other until the predefined stopping rule is satisfied and thus selects the final model containing the variables with the most significant parameters.
The least significant variable that the procedure removes at each step is the one whose absence causes the most statistically insignificant deterioration in the model's fit. Depending on the statistical indicator chosen, it is therefore the variable whose removal from the model causes the smallest increase in the SSE (Error Sum of Squares) or the smallest decrease in the R^2 [1]compared to the absence of the other predictors. If we use the indicators of significance tests with a T-stat[2] or F-stat[3], the variable removed at each step is the one associated with the coefficient that has the highest p-value[4] of the T-stat or the F-stat. The algorithm will

[1] percentage of variance of Y explained by the model

[2] Student test statistic

[3] Fisher test statistic

[4] Probability for a given statistical model under the null hypothesis H0 to obtain a value at least as extreme as the one observed. It depends on the test statistics distribution under given H0.

stop when the p-value reaches a threshold that is usually defined at 5 %. Other statistical indicators can be used such as the AICc(section 2.3.1) or the BIC(section 2.3.2), the variable eliminated at each step would be the one whose absence in the model minimizes the most the AICc or the BIC. A variable would be eliminated at each iteration until the information criterion can no longer be minimized.

One of the advantages of the backward method compared to the forward method is that all the explanatory variables are mobilized since the procedure starts with the complete model, which is important in case of collinearity. But one of the disadvantages compared to the exhaustive method is that it is no longer possible to reintroduce a variable once it has been eliminated from the model.

2.1.2 Forward stepwise selection

The forward selection method proceeds as its name suggests in the opposite direction of the backward method. This method starts with a regression on a constant containing no variables (a null model) and then introduces step by step the most significant variables until a stopping rule is satisfied. This selection consists of fitting a null model at the beginning by performing the p possible regressions at the first step (k=0) with a single explanatory variable. For each of them, a student test and the calculation of a p-value is performed. The model for which the new explanatory variable is the most significant is retained. The process is repeated for p-k models at iteration k (for k = 0,...,p-1) until we can no longer introduce significant variables into the model and thus the stopping rule is reached. As for the backward procedure, other stopping criteria can be used than the threshold of the p-value of a T-stat or an F-test such as the minimization of the criteria AICc or BIC for example. If we choose to use an F-test, the constrained model is the one selected by the previous step, contrary to the backward method where the constrained model is the one of the current step.

An advantage of forward selection over backward is that the selection can be applied even in the high dimensional setting where n < p.

However, as with the backward selection method, one of the major drawbacks of this method compared to the exhaustive method is that once a variable has been introduced into the model, it can no longer be eliminated. The final model selected by the two backward and forward methods may therefore contain variables that were significant in the previous steps and that are no longer significant, or conversely, it may remove variables that were not significant but that become significant in the following steps because of a multicollinearity relationship with the variables introduced afterwards. It is therefore not guaranteed to find the best possible model. The stepwise method, which combines the two methods, solves this difficulty. Indeed, after the introduction of each new variable, the stepwise method re-examines the significance of each variable previously admitted in the model and removes the least significant one. The algorithm stops when the threshold is reached and no additional variable can be added or eliminated from the model.

2.2 Machine learning

Machine Learning is a branch of Artificial Intelligence which aims to creating computer models that can learn from data without being explicitly programmed. There are multiple types of Machine Learning techniques, for instance we can mention supervised learning, unsupervised learning, and reinforced learning. In Machine Learning, variable selection is an crucial step that helps to select the most important variables to be used in building the model. It helps to improve the accuracy of the model by eliminating irrelevant and redundant variables.There are several methods for variable selection such as Lasso, Ridge, Elastic Net, Recursive Feature Elimination (RFE)...etc. We will see in detail the following methods: Stagewise, Lars and Lasso.

2.2.1 Forward Stagewise And LARS regressions

The Forward Stagewise algorithm is a machine learning algorithm. It is a more constrained version of the forward stepwise method that consists of initializing each variable with arbitrary weights for by using a learning step. Then the algorithm makes iterations to update the weights of the variables. At each iteration, it selects the variable that has the biggest effect on the residual and adds it to the set of variables used for the regression. Then it updates the weights for this variable using a residual gradient. This continues until a predefined number of iterations or a minimum error is attained. It should be noted that this algorithm can be sensitive to noisy data and subsequently the selected variables will not be the most relevant for the final model.

Algorithm : Monotone Incremental Forward Stagewise Regression[1]

1. Start with $r = y - \bar{y}, \beta_1, \beta_2, ..., \beta_p = 0$
2. Find the predictor x_j most positively correlated with r
3. Update $\beta_j \leftarrow \beta_j + \delta_j$ where $\delta_j = \epsilon \cdot \text{sign}\left[\text{corr}(r, x_j)\right]$
4. Update $r \leftarrow r - \epsilon x_j$ and repeat steps 2 and 3 until no predictor has any correlation with r

The LARS (Least Angle Regression) is a similar method to the Stagewise algorithm for variable selection for linear regression. It uses a gradient descent algorithm approach to minimize the prediction error. It relies on the correlation of the variables with the target instead of on the potential improvement in model performance. The LARS algorithm applies an iterative method, in each iteration it adds or removes variables to the model according to their correlation with the target. It uses a selection criterion based on the absolute correlation between the variables and the target to the variables which are the most important for the prediction.
The LARS algorithm is considered as an extension of the Stagewise algorithm, but he contents improvements, so for this it is most effective when we have a high number of variables and the variables are correlated, moreover it's more efficient when the number of examples is low.

It is important to note that LARS is sensitive to noisy data and can select irrelevant variables,so it is often used in conjunction with regularization techniques like Lasso regression.

Algorithm : LARS[2]

1. Standardize the predictors to have a mean zero and unit norm. Start with $r = y - \bar{y}, \beta_1, \beta_2, ..., \beta_p = 0$.
2. Find the predictor x_j most correlated with r.
3. Move β_j from 0 towards its least-squares coefficient $\langle x_j,r \rangle$, until some other competitor x_k has as much correlation with the current residual as does x_j.
4. Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on (x_j , x_k), until some other competitor x_l has as much correlation with the current residual.
5. Continue in this way until all p predictors have been entered. After min(N - 1, p) steps, we arrive at the full least-squares solution.

2.2.2 Least Absolute Shrinkage and Selection Operator

LASSO is a constrained optimization program introduced by Robert Tibshirani[5]. It aims to find the estimator $\hat{\beta}$ that minimizes the squared error in a linear regression, under an additional constraint regularizing the parameters. It allows to build models with as few variables as possible, which makes these models interpretable and robust to overfitting. This penalization is recommended when facing a high variable dimension datas, i.e. when p> n. The optimization problem is the following : [1]

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad \text{where} \quad \sum_{j=1}^{p} |\beta_j| \leq t \tag{1}$$

where $\lambda \geq 0$ is the parameter that controls the strength of the penalty, the larger the value of, the greater the amount of shrinkage.

The penalty parameter λ has a crucial importance in penalized regression. To resume, for λ =0 (t very large) we essentially just get the OLS estimates of the full model, and for very high λ (t equal to zero) all lasso estimates are exactly zero. Lasso Regression uses L_1 penalty which instead of squaring the coefficient like Ridge Regression's L_2 penalty, takes its absolute value. The L_1 type penalisation shrinks some coefficients, while the others are canceled out exactly, thus leading to parsimonious models. Contrary to classical linear regression, this regularized regression's objective is to eliminate unnecessary variables and only those variables since the variables x_i are not all relevant.

Shrinking the coefficient estimates significantly decreases their variance. When we apply

[5] https://statistics.stanford.edu/people/robert-tibshirani

shrinking, we most importantly bring the coefficient estimates closer to 0. The need for this method appears as a consequence of underfitting or overfitting the data issues. When we want to minimize the mean error in case of Linear Regression, we need to optimize the bias-variance trade-off. The bias-variance trade-off means the level of underfitting or overfitting of the data according to the Linear Regression model tested for it. The model is underfitted when a high bias-low variance occurs,the model is overfitted when we have and a low bias-high variance. We need to find a balance between bias and variance to achieve the perfect combination for the minimum Mean Squared Error.

In the predictive perspective, the focus is on selecting the best model that can accurately predict new observations. This is typically done using techniques such as cross-validation or resampling methods such as the bootstrap, which estimate the generalization error of the model. The most viable strategy is K-fold cross-validation, by choosing the number of folds K and then dividing the data into training and testing sets. Once we define a grid of values for λ, for each λ we calculate the validation Mean Squared Error within each fold, for each λ we calculate the overall cross-validation MSE, finally we find under which λ cross-validation MSE is minimized. This value of λ is known as the $minimum_{CV}$ λ. We must note that cross-validation error estimates prediction error at any fixed value of λ, and so by applying it, we are implicitly assuming that our purpose is finding minimal prediction error.

When we try to chose the best value of the tuning parameter for selecting the right model, we focus here in selecting the best model that can explain the underlying relationships in the data. This is typically possible using cross-validation or information criteria such as AIC or BIC, which balance the fit of the model to the data with the complexity of the model. Cross-validation is typically used to estimate the performance of the model for different values of the tuning parameter. So we must split the data into training and validation sets, and training the model on the training set while evaluating its performance on the validation set.The final value of the tuning parameter λ is selected when the best performance on the validation set is achieved. This process is repeated several times with different splits of the data to obtain a robust estimate of the model's performance.

LASSO selects only some significant variables, this is a real limitation of the algorithm, because when dealing with multicollinearity, LASSO chooses one, the one that is most related to the target, ignoring the impact of the others. This disadvantage is inherent to techniques integrating a variable selection mechanism (e.g. decision trees, etc.).

2.3 selection criteria

The selection criteria [7] and more particularly the information criteria which were mentioned previously measure the quality of a statistical model. They penalize the log likelihood according to the number of parameters. Their equation reflects a compromise between a model with a high likelihood and therefore a strong explanatory power and a limited number of variables to avoid overfitting in particular. The selected model is the one that minimizes these criteria. We are particularly interested in the Akaike and Bayesian criteria.

2.3.1 AIC and AICc

The akaike information criterion is based on the fact that the addition of a parameter reduces the SSE and thus increases the likelihood of a model. It is an information criterion that satisfies the parsimony criterion based on a trade-off between the goodness of fit measured by the likelihood or the SSE and the complexity of the model. It penalizes models with a large number of parameters. The model chosen is the one that minimizes the AIC criterion. Adding an additional variable to this model makes it more complex than it fits. The AIC is an asymptotically justified criterion. It is known for selecting the model that will make the best predictions in the future. The Akaike information criterion is defined as :

$$AIC = -2\log(L) + 2k \qquad \mathbf{(2)}$$

where
k : the number of parameters to be estimated in the model
L : the likelihood function of the model
SSE : Residuals Sum of Squares

The AICc is a correction of the AIC. It is recommended to use it when the number of parameters is large compared to the number of observations. It penalizes additional variables more than the AIC. AICc is defined by :

$$AICc = AIC + \tfrac{2k(k+1)}{n-k-1} \qquad \mathbf{(3)}$$

where
n: the number of observations
k : the number of parameters to be estimated in the model

2.3.2 BIC

The Bayesian information criterion BIC is an information criterion that is also asymptotically justified. The model selected by this criterion is asymptotically the most probable. By introducing the sample size into its equation, the BIC also penalizes the number of parameters more strongly than the AIC. Indeed, the BIC is defined by :

$$BIC = -2log(L) + klog(n) \qquad \mathbf{(4)}$$

where
k : the number of parameters to be estimated in the model
L : the likelihood function of the model
n : the number of observations

Moreover, it should be noted that the BIC is called SBC by the SAS language. There is also a criterion named BIC by SAS but which corresponds to another criterion which is calculated in the following way :

$$BIC(SAS) = nlog(\tfrac{SSE}{n}) + 2(p+2)q - 2q^2 \qquad \mathbf{(5)}$$
$$\text{where } q = \tfrac{n\hat{\sigma}^2}{SSE}$$

where
n = the number of observations
p = the number of parameters including the intercept
$\hat{\sigma}^2$ = Estimate of pure error variance from fitting the full model
SSE = Error sum of squares

3 Selection method performance comparison

3.1 Methodology and materials

In the following sections we will look at the performance of selection methods on different types of data. Our different metrics : the probability of fitting the model well, the probability of overfitting the model, the probability of underfitting the model and the probability of finding any number of true variables conditional on underfitting the model. To calculate these, we use a sample size of 1000 and 50 potential candidate variables among which are the variables composing the true model. We will use the GLMSELECT procedure in SAS to select the model for each selection method and criterion. This process will be repeated a number of times (from 500 to 1000) and we will calculate the completion rate per goodness-of-fit. We will present the means of these metrics, their minima and maxima to capture extreme events, if any. As the data are generated in a sufficiently large sample size and a sufficiently large number of times, we can assume that the data are always comparable for a given structure and number of simulations. This assumption is very important when comparing the performance of selection methods.

Data types

We will use different types of data in our work. The aim is to compare how the selection procedures and stopping criteria behave in the presence of these data types. We have :

- **Gaussian and independent data** : These data were generated under the null hypothesis of independence. In effect, they follow a multivariate normal distribution, where each is a centred reduced normal.
- **Gaussian and correlated data** : These data were generated under the alternative assumption of linear dependence. If the variables follow normal distributions, linear dependence has been applied to them. In effect, we took a subsample of variables (including those in the true model) and ran a particular correlation structure on them which we specified with a Toeplitz[4] form. To ensure that the correlation matrix is semi-definitely positive, we used Higham's[5] algorithm to find the nearest correlation matrix in the Frobenius A sense that is semi-definitely positive. The latter is used to pass the correlation to the variables with the Iman-Conover[4] algorithm. The rest of the variables are independent and identically distributed Gaussian.
- **Independent data with extreme values** : Again, considering a subsample of variables (including those of the true model), we use mixtures of multivariate normal distri-

butions to generate extreme values to perturb the moments of our variables. However, we used Iman-Conover to force independence between all candidate variables.

- **Correlated data with extreme values** : Similarly, extreme values are generated with mixtures of multivariate normal distributions at the level of a subsample of variables (including those of the true model). The Higham and Iman-Conover algorithms are used to give the correlation structure between the variables.

We will now look at the interpretations of the performances obtained for each method, according to the criteria and the types of data. On the x-axis of our vertical bar charts, we have the cases of overfitting, perfect-fitting and underfitting. On the ordinate we have the probability of each method to select models in each of these cases. We have also represented the conditional probability of underfitting.

3.2 Gaussian and independent data

3.2.1 AICC and BIC : Statistical learning vs Machine learning

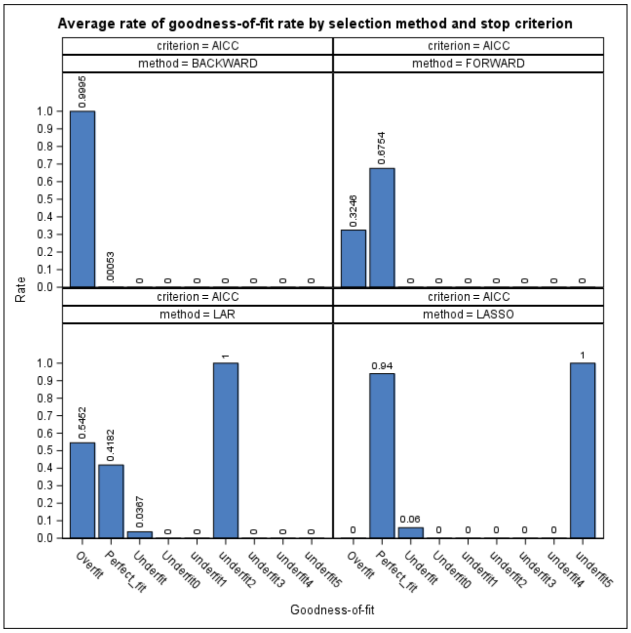


Figure 1: Distribution of the mean rate of goodness-of-fit on normal weakly correlated data for AICC criterion.

Through our vertical bars (Figure1) which concern the perfect data, we can first observe that with the AICc criterion the backward selection method has a very high probability of having the overfitting problem (99%), so with this method we will certainly have more variables selected than those in our true model.

The forward and LARS methods find the true model 67.5% and 41.8% of the time respectively. We find almost no cases of underfitting with the forward and LARS methods.

However, when LARS underfits the model (with a probability of 3.67%), it selects only 2 variables out of 6 of the true model 100% of the time.
Forward is more likely to select the right variables than to overfit unlike LARS which overfits in 54% of the cases. The forward method seems to be better than the backward and LARS using the AICc penalty criterion.

Concerning LASSO, we find the true model 94% of the time and the probability that LASSO selects one variable less than the true model knowing that there has been an underfitting is 99%. Thus, with the AICc criterion, the model selected by LASSO, even when there is underfitting (where it has a 6% chance of happening) remains close to the true model.
Concerning the BIC criterion, the performances of the methods are the same as with the AICc criterion. See Figure 16a.

3.2.2 SBC : Statistical learning vs Machine learning

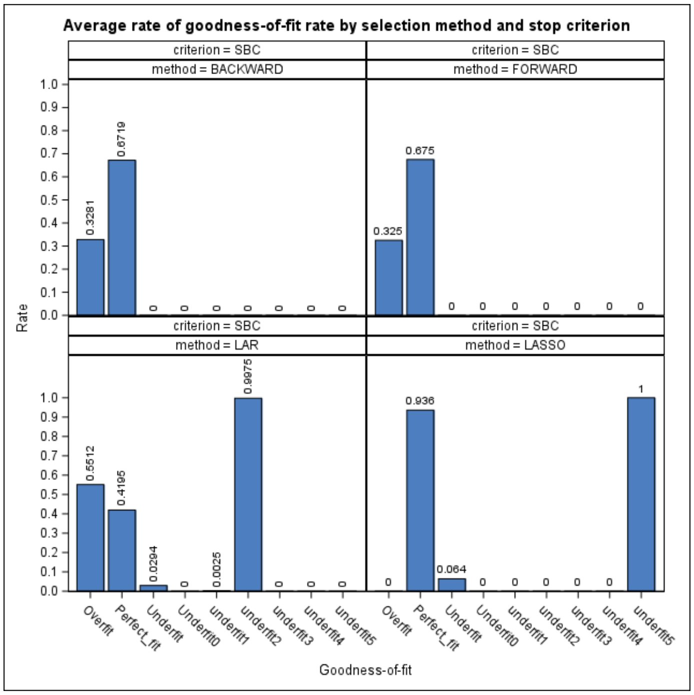


Figure 2: Distribution of the mean rate of goodness-of-fit on normal weakly correlated data for SBC criterion.

When applying the SBC criterion (figure 2), the backward and forward methods have similar performances. The probability of having an overfitting model with the backward method is lower than with the AICc criterion (32% versus 99%). This is due to the fact that the SBC is a more penalizing criterion than the AICc and that the backward method tends to do a lot of overfitting since its algorithm starts with a complete model.
This allowed a considerable improvement of the performance of the backward method. There is indeed a 67% chance of obtaining the true model with the SBC criterion against 0.053% with the AICc. We will notice that for each type of data the SBC criterion will considerably

improve the backward selection method.

3.2.3 Cross-validation : Statistical learning vs Machine learning

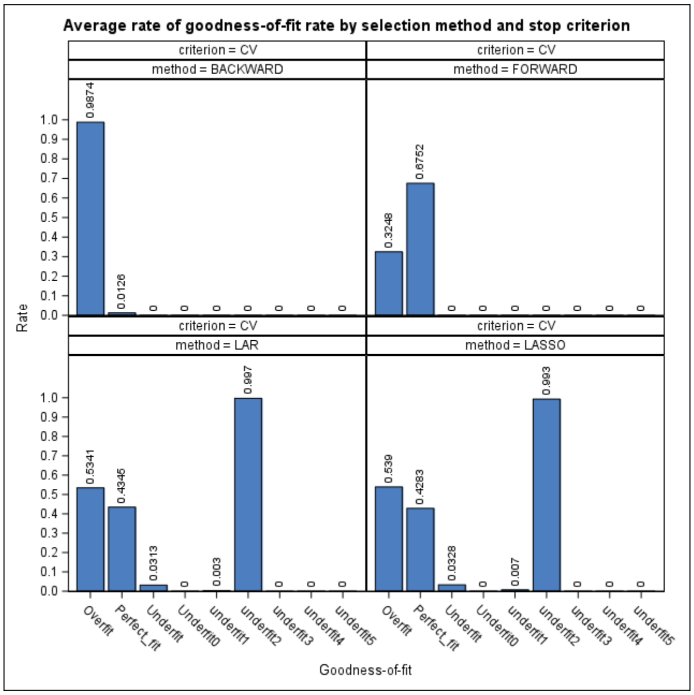


Figure 3: Distribution of the mean rate of goodness-of-fit on normal weakly correlated data for CV criterion.

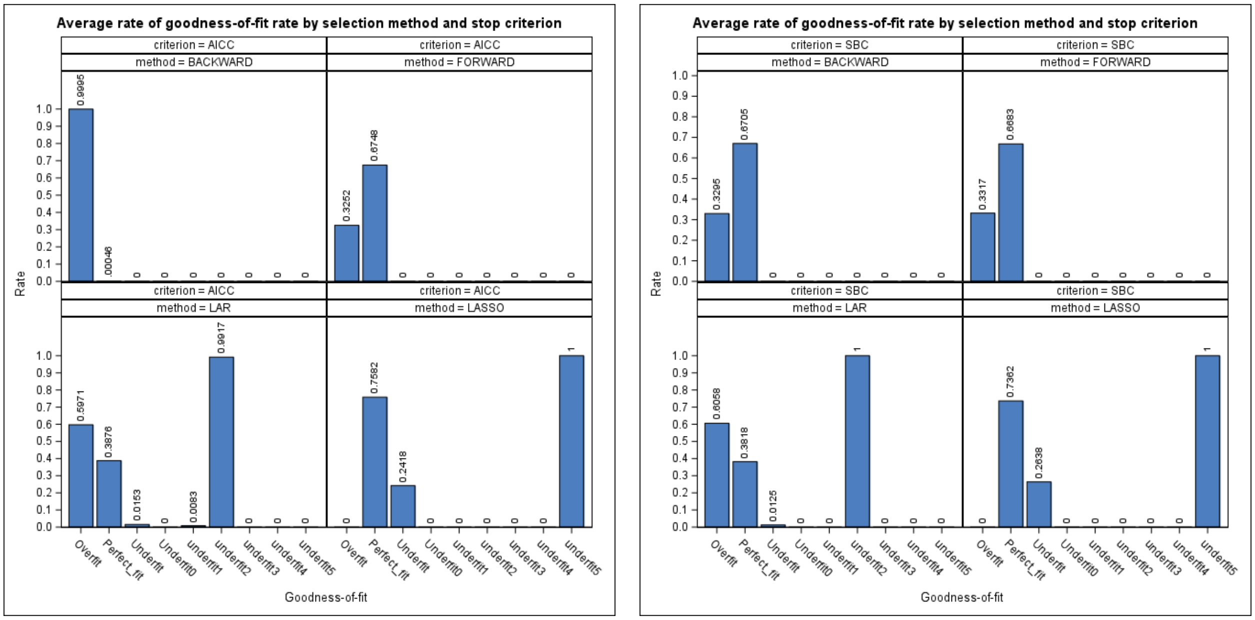
Using an empirical method like cross-validation (figure 3) as a criterion (CV) instead of a sample penalty criterion like AICc, BIC and SBC, we notice that the probability of over-fitting decreased very legerly for the backward method (97% instead of 99%). The use of the CV increased the chances of finding the right model (1.26% instead of 0%), but it remains a very low probability.

In addition, the probability of the LASSO finding the true pattern went from 94% with the other criteria to 42% with the CV. In addition, the probability of over-fitting with the LASSO went from zero probability to 53% with this criterion.

To conclude on the performances of our methods on Gaussian data, we can say that using penalty criteria like the BIC, the LASSO is the best method for the selection of variables. Then we find in order forward, backward with SBC and finally LARS. LARS is generally better than the backward method with the other criteria on this type of data.

3.3 Independent data with outliers

3.3.1 AICC, BIC and SBC: Statistical learning vs Machine learning



(a) Distribution of the mean rate of goodness-of-fit on weakly correlated data with outliers for AICC criterion.

(b) Distribution of the mean rate of goodness-of-fit on weakly correlated data with outliers for SBC criterion.

Figure 4: AICC and SBC

In figure4, we see that the performances of the backward and forward methods are exactly the same as on Gaussian data without extreme values, when we use AICc, BIC (see figure 16b)and SBC criterions.
Concerning LARS the difference is very small compared to what we have seen previously on normal values. There is a very slight decrease in the probability of LARS to find the true model of about 2% in the presence of extreme values whatever the criterion used.
However, LASSO performs better on data without outliers, we notice that the probability of having the true model is about 75% with the AICc, BIC (see figure 16b). and SBC criteria.

3.3.2 Cross-validation: Statistical learning vs Machine learning

The performance of this method also appears to be almost identical to that of the normal data without extreme values(figure5). The performance of LARS was reduced by 13 percentage points compared to the normal data without extremes.The performance of LASSO with CV also seems unchanged from the data without extreme values. But CV is obviously not a good selection criterion for LASSO.
We can conclude that the forward, backward and LARS methods are not sensitive to extreme values.

Despite the loss of the probability of LASSO to find the true model of 20 percentage points

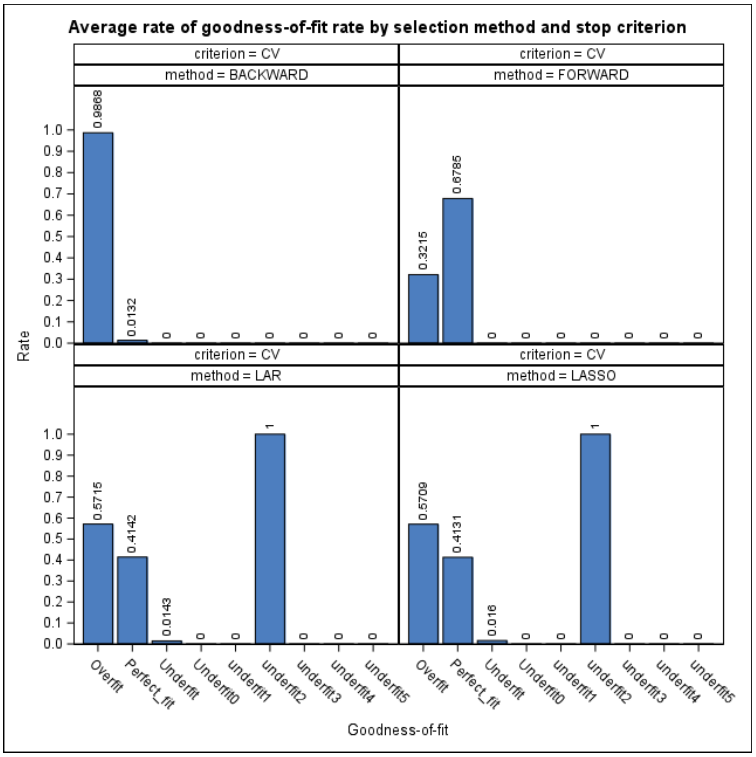


Figure 5: Distribution of the mean rate of goodness-of-fit on weakly correlated data with outliers for CV criterion.

compared to normal data, it remains the most performing method on data with outliers. The order of performance of the tested methods remains the same as on normal data without outliers.

3.4 Gaussian correlated data

3.4.1 AICC and BIC : Statistical learning vs Machine learning

With the AICc criterion(figure6), LASSO is the best performing method with a probability of 3.87% of finding the true model, the probability of the other methods to select all the variables of the true model are almost null.
The methods tend to do more overfitting than underfitting. All the methods select at least 50% of the time more variables than the true model, i.e. they are more often overfitting.
In particular, LASSO does 75% of the overfitting and 20% of the underfitting with this criterion. When the LASSO does underfitting, it selects in 97% of the cases, 5 variables out of 6 of the true model.

Concerning the methods : backward, forward and LASSO the performances are almost the same using the AICc criterion while using the BIC criterion (see figure 17a). The probability of LASSO to find the true model has slightly increased, it remains the best selection method.

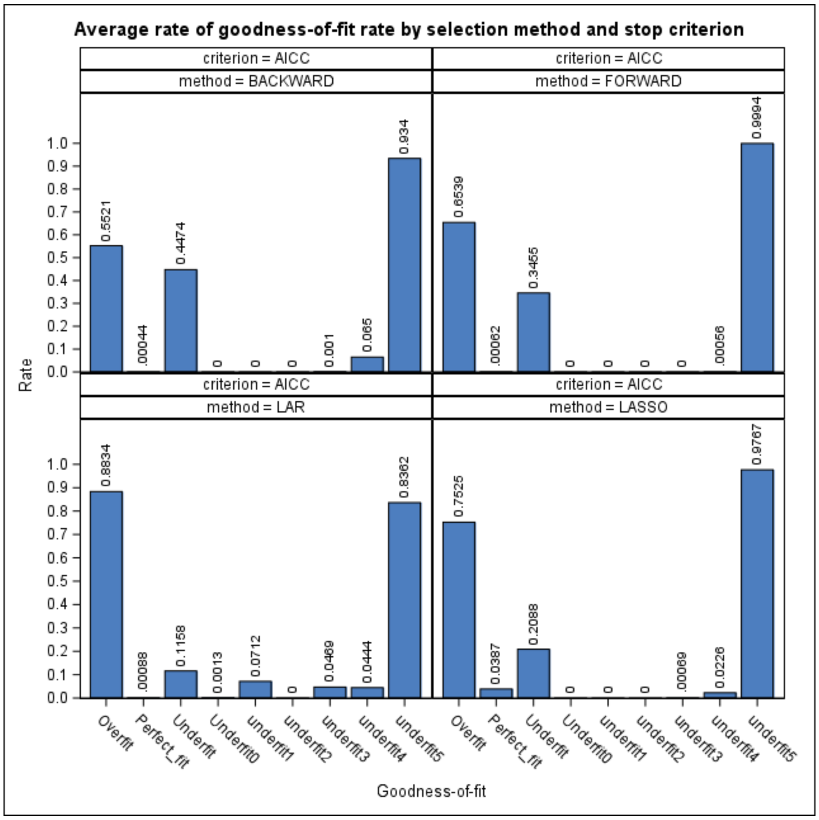


Figure 6: Distribution of the mean rate of goodness-of-fit on normal highly correlated data for AICC criterion.

3.4.2 SBC : Statistical learning vs Machine learning

With this criterion (figure7), the backward and forward methods have the highest probability of finding the true model even more than the LASSO. We find once again the relevance of using the SBC criterion for the backward method. Indeed, the probability that the backward selection finds the true model was almost zero for all the other criteria and now goes to a small probability of 1.75% with this criterion.
LARS has a very high probability of overfitting compared to other methods.

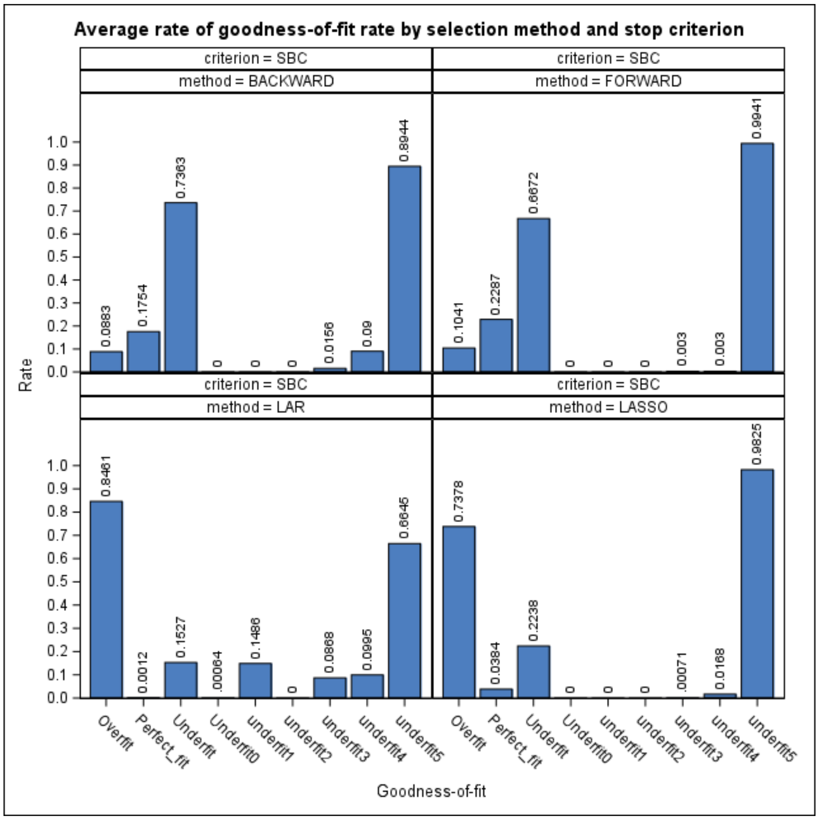


Figure 7: Distribution of the mean rate of goodness-of-fit on normal highly correlated data for SBC criterion.

3.4.3 Cross-validation : Statistical learning vs Machine learning

With this criterion (figure8) we notice that the best performing methods are forward and LASSO but this remains a low performance because the probability of falling on the true model is 3.4% and 3.92% respectively. They are also the methods that have the least chance of overfitting contrary to backward and LARS.

To summarize, on correlated normal data, we notice that the performance of LASSO is not the best when using the SBC criterion. The performance of the methods and in particular LASSO has extremely decreased with the correlation of the variables. This confirms what we have seen in the theory about LASSO and its sensitivity to multicollinearity. However, LASSO remains the best selection method on this type of data when the criterion used is not the SBC. On correlated data without outliers, the selection of variables by forward using the SBC criterion is the best.

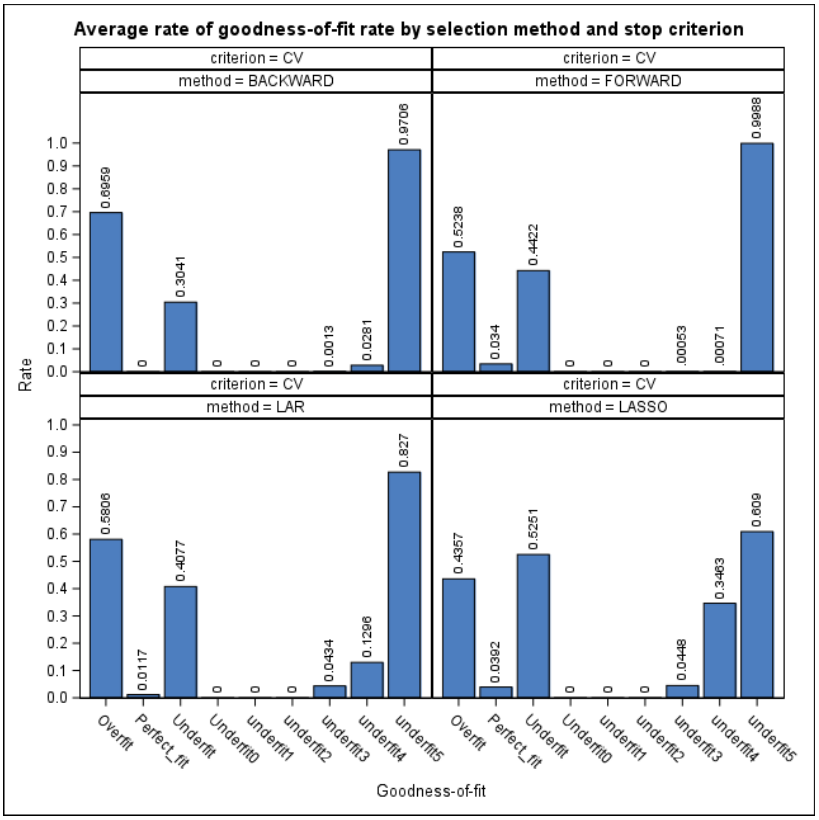


Figure 8: Distribution of the mean rate of goodness-of-fit on normal highly correlated data for CV criterion.

3.5 Correlated data with extreme values

3.5.1 AICC and BIC : Statistical learning vs Machine learning

This criterion (figure9) performed well with the forward and LASSO methods, which selected the true model in 68.62% and 47.66% of the cases, respectively. Their performance was below 4% with correlated data without extreme values.

Furthermore, backward selection remains a very inefficient method when applying this criterion with a very high percentage of overfitting. Concerning LARS, this method selects the true model only 6.24% of the time, and tends to have a very high overfitting percentage. The backward and LARS methods therefore continue to have very poor performances on correlated data even with extreme values.

We find again here on our correlated data with extreme values that the performances of all the selection methods with the BIC criterion are very similar to those of the AICc criterion. See figure17b.

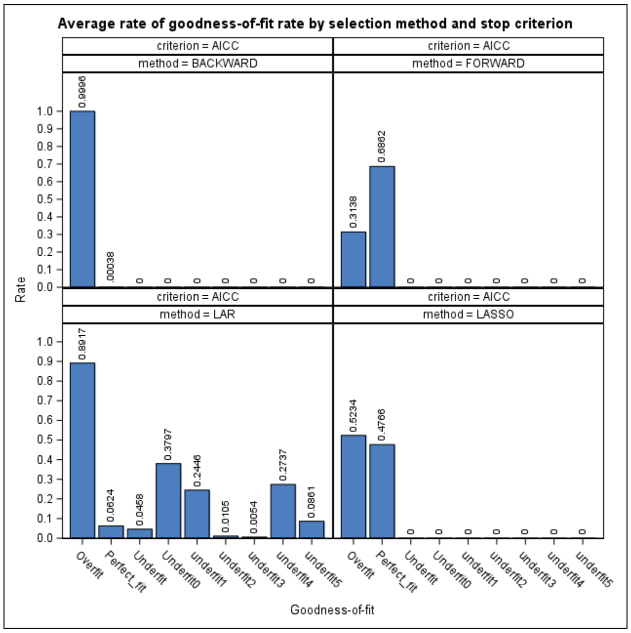


Figure 9: Distribution of the mean rate of goodness-of-fit on highly correlated data with outliers for AICC criterion.

3.5.2 SBC : Statistical learning vs Machine learning

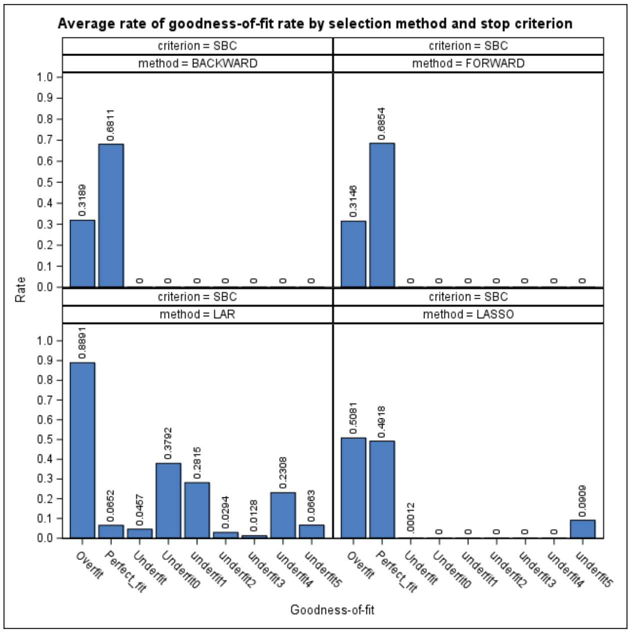


Figure 10: Distribution of the mean rate of goodness-of-fit on highly correlated data with outliers for SBC criterion.

There is a great improvement of the backward method with this criterion (figure10) even on data correlated with outliers. The probability of this method to obtain the right model is

68% with this criterion. It is therefore an efficient criterion on stepwise backward whatever the type of data. But this big difference in performance with this criterion is surprising because the backward with the SBC has a better performance than the LASSO while with all the other criteria the probability of the backward to find the right model is almost zero. We will therefore not generalize this finding.

3.5.3 Cross-validation : Statistical learning vs Machine learning

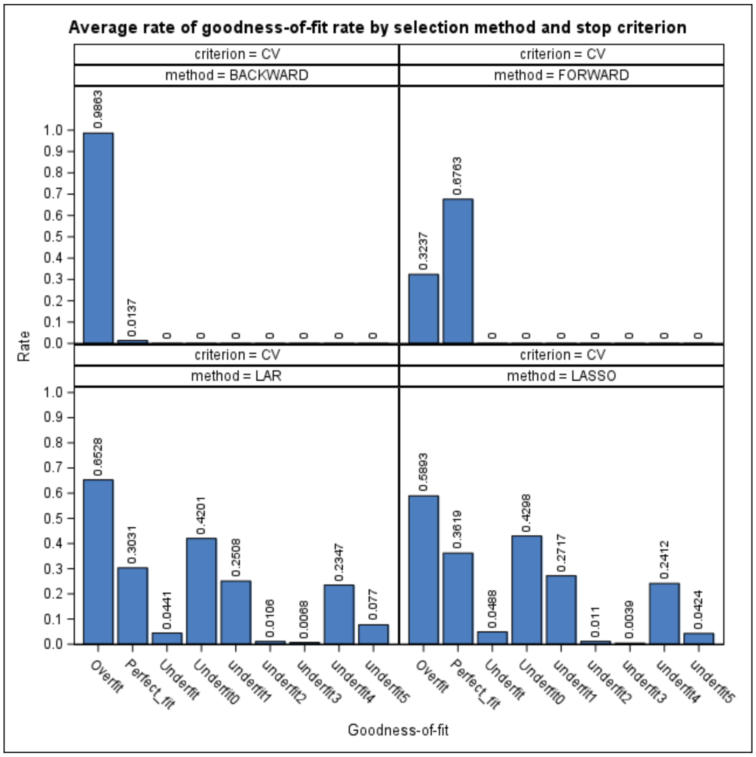


Figure 11: Distribution of the mean rate of goodness-of-fit on highly correlated data with outliers for CV criterion.

By applying this criterion (figure 11)we see that forward selects the true model 67% of the time while LASSO does so only 36.19% of the time. Forward is therefore the most efficient selection method whatever the selection criteria. The efficiency of LASSO to find the right model has decreased (by 12% points) with this information criterion while the performance of LARS has clearly increased (by 24% points) with this criterion. We observe that the CV is the most adapted criterion to LARS whatever our data.

Thus, with data in the presence of correlations between the variables and extreme values, the performances are better than in the presence of correlation only. This result is less surprising when we notice by comparing the data of the hypothesis H0 normal and H0 extrem that the selection methods do not seem sensitive to the extreme values.
It is interesting to note the surprising differences in performance between the backward and forward statistical stepwise learning methods even though both methods have high and close

performance with the SBC.
The forward and LASSO methods are the best methods in general in the presence of correlation and extreme values, even if we should note the high results of backward with the SBC. Forward is even the best performing selection method whatever the selection criterion on this type of data. We will discuss these results in the general conclusion.

4 Further looks : Monte Carlo simulation diagnostics

To measure the performance of the selection methods, we performed Monte Carlo simulations with different numbers of iterations. Indeed, we can obtain different values for the probabilities we have calculated. This is why we have presented the averages. However, it is possible that the latter are influenced by extreme realizations. Therefore, it is wise to look at the sensitivity of the simulation to the number of iterations performed. To do this, we will analyse the range of variation of the probabilities of fit, over-fit and under-fit on our different types of data.

Gaussian and independent data

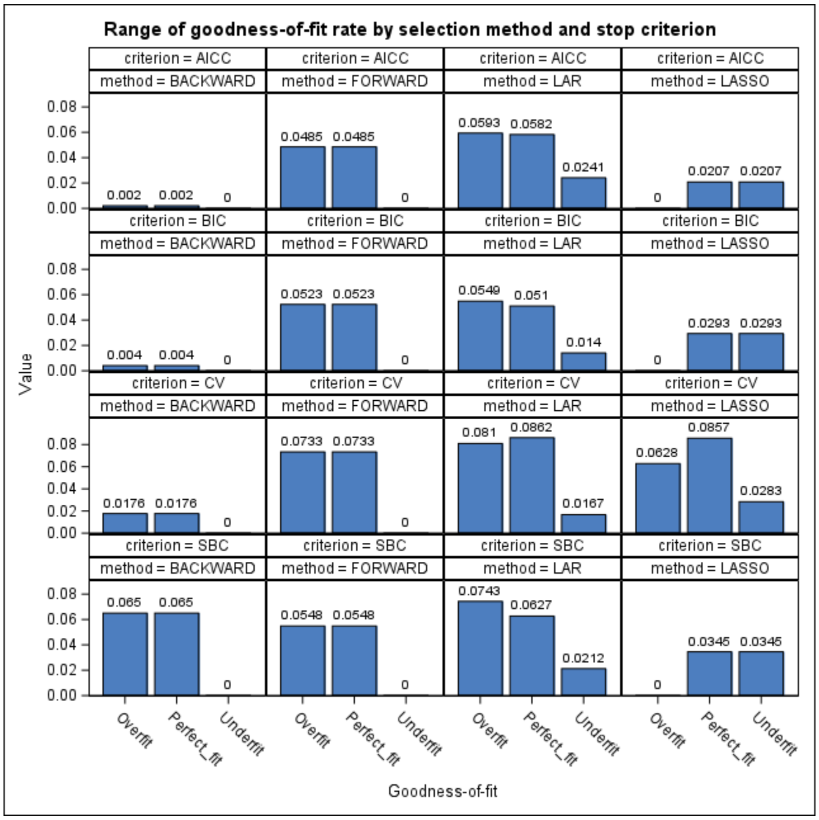


Figure 12: Range of goodness-of-fit rate by selection method and stopping criterion, for weakly correlated gaussian data.

We notice that there is a larger difference between the maxima and minima of the perfect-fit probabilities of the LARS and LASSO methods using the cross-validation criterion, compared to the other methods and selection criteria. After these two, comes the forward method. The CV criterion seems to be the one that occaions more extreme realizations in the performance of the methods. This is probably due to the partitioning of the sample by k-fold.(See figure 12)

Independent data with extreme values

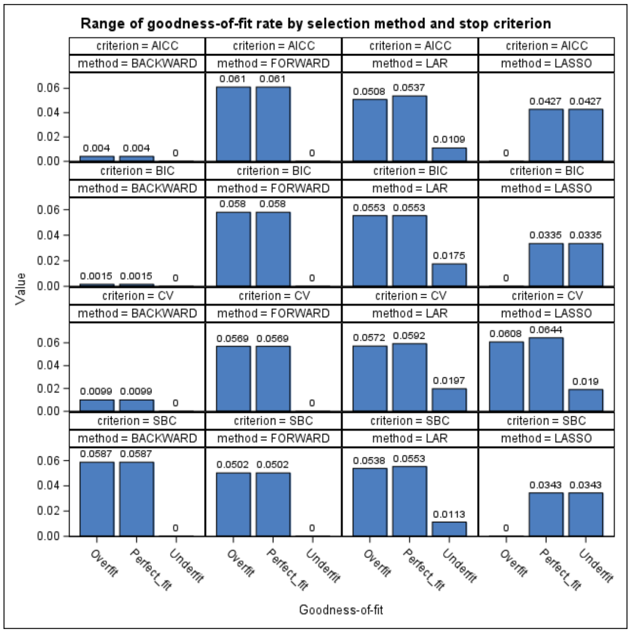


Figure 13: Range of goodness-of-fit rate by selection method and stopping criterion, for weakly correlated data with extrem values.

On the data with outliers simulated under the null hypothesis of independence, we can notice on the graph opposite that the Backward and Forward selection methods have the same ranges of variation for overfit and underfit, given the information criterion. Furthermore, we notice that the Backward method gives the least dispersed probabilities for the AICC, BIC and CV criteria. On the other hand, it gives the widest probabilities with the SBC criterion. Moreover, the range of probabilities of the Forward method is relatively stable regardless of the criteria. The same is true for the LAR method. In contrast, the LASSO shows higher differences in the range of variation. Indeed, the probability of overfitting is fixed whatever the number of simulations for the AICC, BIC and SBC criteria, due in particular to the use of a fixed value for the Lagrange multiplier λ_1. With cross-validation, the probability of perfect-fit admits the largest range of variation. (See figure 13)

Correlated and Gaussian data

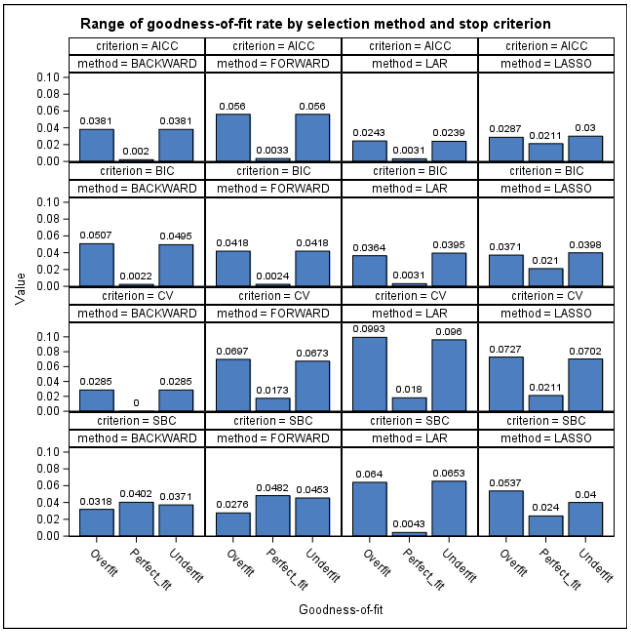


Figure 14: Range of goodness-of-fit rate by selection method and stopping criterion, for correlated gaussian data.

On correlated and identically distributed Gaussian data, the LAR method (CV criterion) presents minima and maxima furthest from the overfitting and underfitting probabilities of the ensemble, i.e. 0.0993 and 0.098 respectively (to one decimal). Indeed, due to the cross-validation and multicollinearity between the predictors, the metrics are much more sensitive to the number of iterations. The Backward metrics are less stable than before for the AICC, BIC and CV criteria. It can also be seen that the minimum and maximum of the underfit probability are further apart in the present framework. With multicollinear predictors and other variables, this could be expected as the algorithms will tend to err on the side of moving variables in and out of the model.(See figure 14)

Correlated data with extreme values

While Forward selection performed quite well on average on multicollinear data with extreme values, there is a significant difference between the minimum and maximum perfect-fit rates, i.e., 0.0923. Thus, there are extreme realizations in the perfect-fit series. Moreover, the average performance of the LAR method when performing cross-validation is also driven by extreme realizations. Note that cross-validation is the criterion that gave the best results for the LAR method.(See figure 15)

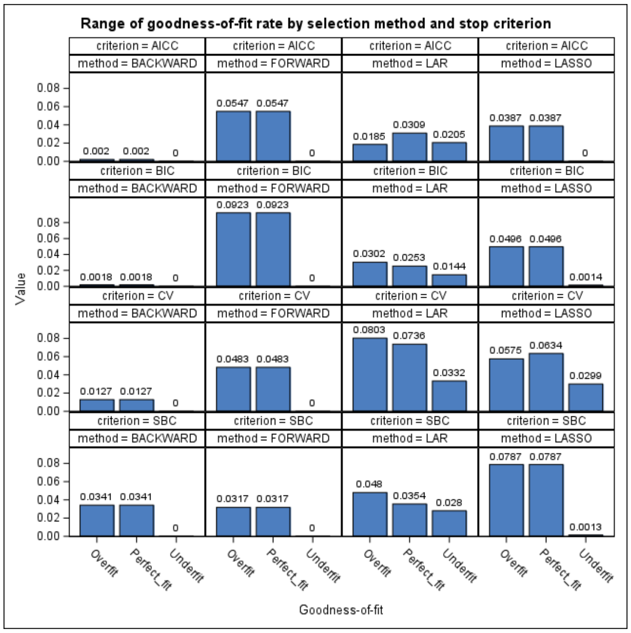


Figure 15: Range of goodness-of-fit rate by selection method and stopping criterion, for correlated data with extrem values.

5 Conclusion and discussions

In this paper, we studied the performance of variable selection methods in different settings. To do so, we presented the theoretical framework of the different methods and stopping criteria. Then, we generated data calibrated in different ways from Monte Carlo simulations. We computed the perfect fit, overfit and underfit probabilities. Also were calculated the conditional probabilities of finding k variables given the fact that the algorithm underfit the model. We used algorithms that, including the data generation, determine the realization rate of the events after a given number of iterations.

On perfect data, we saw in section 3.2 that LASSO performed better than the other methods, except when using cross-validation. But, in this context, we will retain the results given by the SBC because it is a criterion that favors obtaining the true model. In the literature, it is an asymptotically justified criterion that allows us to find the right model. This theoretical finding can also be justified empirically because on all the data types we generated, the SBC generally obtained the best results for each method. Its results were superior to the CV and very close to the AICc with the difference that it is the only selection criterion that works well on the backward method. LASSO is thus followed by Forward, Backward and finally LARS according to the SBC. However we noticed that LARS is better than Backward with the other criteria on all types of data except in the presence of linear dependence (H1 normal) where its results were very close to 0 for all criteria.

On the data generated with extreme values under the null hypothesis of independence, we saw in section 3.3 that LASSO wins again the prize of the method that best selects the true variables of the model, considering all criteria except CV. It is then followed by forward which obtains the second position for all criteria, then backward and LARS if we consider the SBC criterion.

In the case of normally distributed data under the alternative hypothesis of linear dependence, we saw in section 3.4 that: considering the AICC only LASSO does not give a almost zero probability of finding the true model.It is also the least bad when using cross-validation. On the other hand, with the SBC we have this order: Forward, Backward, LASSO and LARs. The Forward method with the SBC is the best performing combination on these data.
In the case of correlated data with extreme values, we obtain the following ranking: The Forward, the Backward, the LASSO and finally the LARS with the SBC criterion. (see Section 3.5). The LASSO obtains the second position with the other criteria after the Forward which achieves the best performance regardless of the criterion used.

Thus the conclusions of the debate between statistical learning and machine learning are mixed because Forward statistical learning and LASSO machine learning methods have complementary performances on different types of data. Nevertheless, Machine learning selection methods such as LASSO has an advantage over the others in that it has a hyper-parameter on which we can play to improve the results. Indeed, for reasons of internal and external comparability, we did not vary the choice of lambda in the framework of the study. But, it's maybe possible to improve the LASSO performance because it remains a large set of possibilities for adjustment.

The results presented above depend to some extent on the way the data we have generated are structured. But, in reality, the work is done on real data that will not necessarily be calibrated in this way. So, it is possible to prefer one method to the other but also one criterion to the other. Moreover, it is important to underline that the problems of explanation or prediction are not the only ones encountered. Indeed, the cost of using the methods and criteria is also to be considered. For example, a criterion like PRESS which is heavy in terms of execution time is not very encouraging. Also, the AIC is not very interesting in high dimension, i.e. p $\simeq n$.

Moreover, we could notice that the generated data with extreme values pose problems to the selection algorithms but rather less than expected. This is due to the fact that the extreme values were simulated from a mixture of multivariate normal distributions. Indeed, due to hardware problems, we could not simulate them in a univariate way. This could be even more interesting. So, if an observation (row) of the explanatory variables (x) has extreme elements, almost all of them are. Since, moreover, the dependent variable (y) is a linear function of a part of these variables, the value corresponding to an extreme observation of x is also an extreme value except in case of compensation by the parameters. Thus, we have leverage points. Therefore, the algorithms can detect the leverage effect. Hence, it could be interesting to see how the results would be if some of the variables of the true model were normally distributed. In this sense, we could decrease the occurrence of leverage points.

Furthermore, if the Gaussian correlated data did a lot of harm to the selection methods, we have seen that the latter performed better on correlated data with extreme values. As can be seen in Figure 18, the extreme values put distortion in the structure of the correlations, apparently non-linear. It follows, therefore, that the algorithms are more likely to err in the presence of multicollinearity than nonlinear dependence.

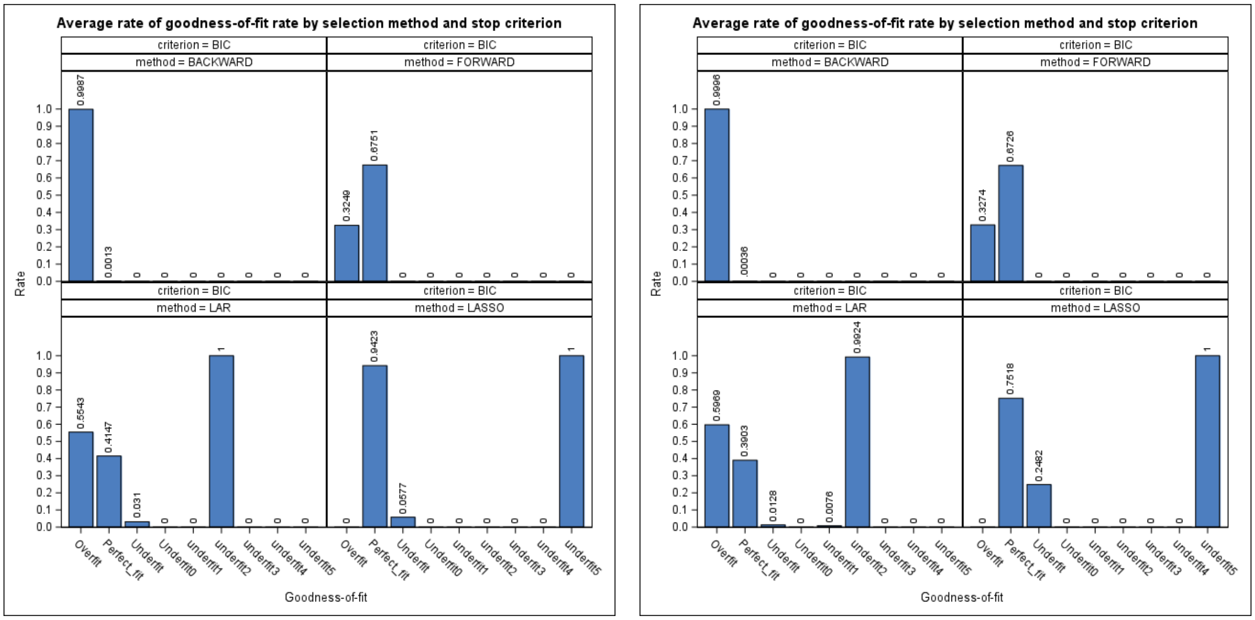
A Frobenius distance

Let A, $X \in \mathcal{M}_n(\mathrm{R})$, two symmetric matrices. The Frobenius distance is defined by:

$$\|A - X\|_F^2 \equiv A^T A + X^T X - 2A^T X.$$

where A^T and X^T are the transpose of A and X matrices respectively.

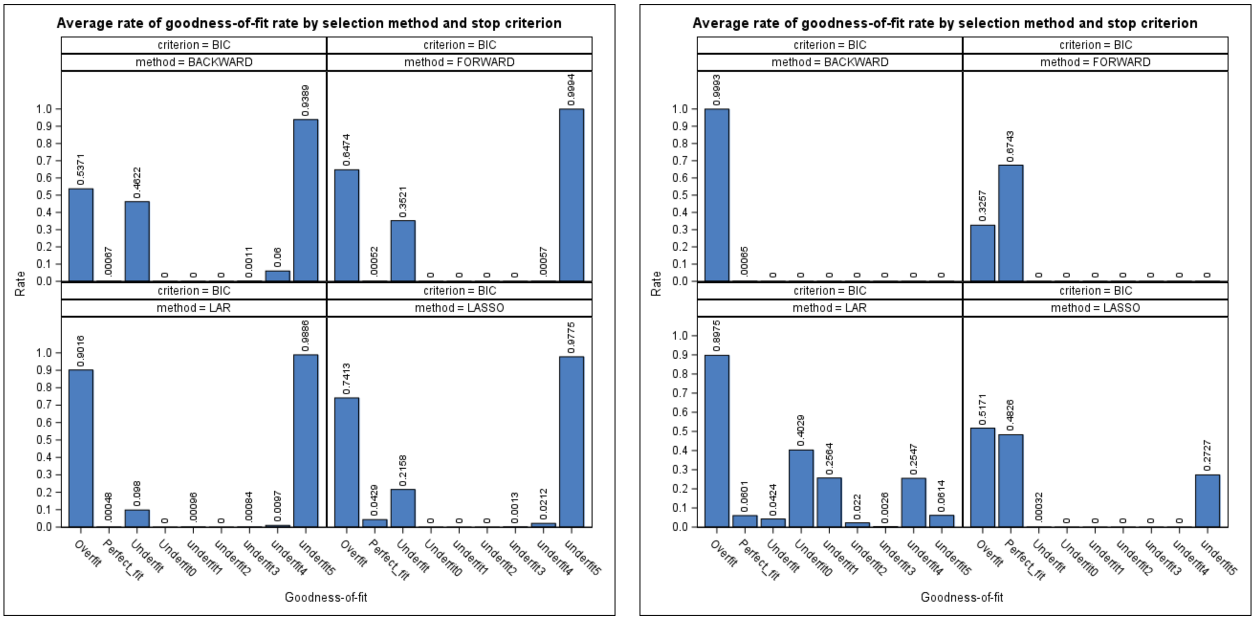
B Additional relevant figures.



(a) Distribution of the mean rate of goodness-of-fit on normal weakly correlated data for BIC criterion.

(b) Distribution of the mean rate of goodness-of-fit on weakly correlated data with extreme values for BIC criterion.

Figure 16: The BIC under the independence null hypothesis.



(a) Distribution of the mean rate of goodness-of-fit on normal highly correlated data for BIC criterion.

(b) Distribution of the mean rate of goodness-of-fit on highly correlated data with extreme values, for BIC criterion.

Figure 17: The BIC under the alternative hypothesis

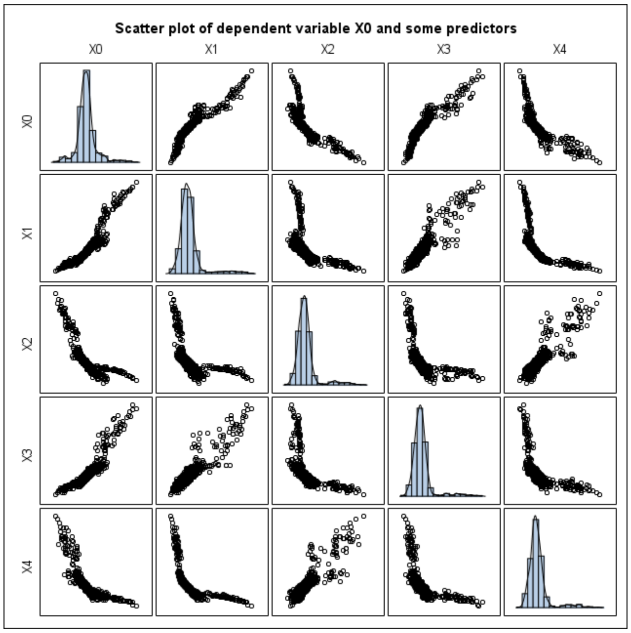


Figure 18: Scatter plot matrix of dependent variable X0 and some predictors, correlated data with extreme values.

References

[1] Hastie, T. et al (2007), ”Forward stagewise regression and the monotone lasso”, *Electronic Journal of Statistics*,1,1-29.

[2] Hastie, T., Tibshirani, R. and Friedman, J. (2009), ”The Elements of Statistical Learning: Data mining Inference and Prediction”, *Springer*, 2nd Edition.

[3] Hastie, T, James, G., Tibshirani, R. and Witten, D. (2021), ”An introduction to Statistical Learning ”, Springer, 2nd Edition.

[4] Wicklin, R. (2013), ”Simulating data with SAS”, *SAS Institute.*

[5] Higham, N. J. (2002), ”Computing the Nearest Correlation Matrix- a problem from finance”, *IMA Journal of Numerical Analysis*, 22, 329-343.

[6] De Peretti, P. Sélection de variables, statistical learning,

[7] GLMSELECT SAS Documentation