Benchmarking the Stability of Variable Selection Methods in the Cox Model

Mathilde Sautreuil
Laboratory of Mathematics and Informatics (MICS)
CentraleSupélec, Université Paris-Saclay
91190 Gif-sur-Yvette, France
mathilde.sautreuil@centralesupelec.fr

Sarah Lemler
Laboratory of Mathematics and Informatics (MICS)
CentraleSupélec, Université Paris-Saclay
91190 Gif-sur-Yvette, France
sarah.lemler@centralesupelec.fr

Paul-Henry Cournède
Laboratory of Mathematics and Informatics (MICS)
CentraleSupélec, Université Paris-Saclay
91190 Gif-sur-Yvette, France
paul-henry.cournede@centralesupelec.fr

March 4, 2021

Abstract

This paper benchmarks the stability and quality of feature selection methods in the Cox model with high dimensional covariates. For this purpose, we consider different classical regularization procedures and screening methods, and we use several indexes to measure the stability and the quality in variable selection of each method. We first study the stability of these methods from simulated data and then the stability of gene selection on a real dataset. The simulation study confirmed the low stability of regularization methods and showed the selection quality was lacking. The screening methods seem to solve the stability problem in high-dimension from our study, but the selection quality is not always correct. We observe a similar behavior of these methods on the real dataset. Finally, the paper highlighted the potential of a screening method using biological knowledge.

1 Introduction

Precision medicine is often seen as the future of medicine. Its objective is to personalize treatments, diagnostic, or prognostic, according to each patient's characteristics. With the advent of high-throughput sequencing, the data enabling to characterize patients can be extremely voluminous, they provide their molecular portraits. However, it can be difficult to extract from this rich mass of data the most relevant information, the key features of interest. These key features are often referred to as "markers". For prognostic purposes, survival analysis models are extremely useful tools to predict patients relapse or death. However, the large number of covariates in view of the usual cohort size hinders model identification as well as model interpretation. In this high-dimensional setting, methods have been developed to reduce the dimension by selecting the most relevant covariates for the model. The aim of this paper is to study the stability and quality of classical feature selection methods for the Cox model.

Survival analysis is the study of the time elapsed until the occurrence of an event of interest. We will call it *death*, but it may as well be relapse or remission. Consider a simple survival regression model of the form:

$$Y_i \sim \mathbb{P}(y|\beta^T X_{i.}), \tag{1}$$

with Y_i the survival time of individual i, $X_{i.} = (X_{i1}, \dots, X_{ip})^T$ the set of variables of individual i that we suppose standardized and $\beta = (\beta_1, \dots, \beta_p)^T$ the regression parameters to be estimated. In this paper, we focus on the Cox model [1] to link the survival time to covariates. The model is written in terms of the conditional hazard and the regression parameter β of this model reflects the effect of variables (genes) on survival duration. It will be detailed in

Section 2.1. The classical procedure for estimating the parameter β consists of minimizing the opposite of the partial log-likelihood $\mathcal{L}$:

$$\widehat{\beta} = \arg\min_{\beta} \left\{ -\mathcal{L}(\beta) \right\}. \tag{2}$$

Variable selection in survival analysis consists of automatically setting to zero the coefficients of variables having the least impact on survival duration, while conserving the variables with the greatest coefficients among the estimated parameters. The most classical solution is the well-known Lasso, which consists in minimizing the opposite of the partial log-likelihood $\mathcal{L}$ to which an L^1 penalty term is added:

$$\widehat{\beta} = \arg\min \left\{ -\mathcal{L}(\beta) + \lambda ||\beta||_1 \right\}. \tag{3}$$

where $\lambda \in \mathbb{R}^+$ is a regularization hyperparameter controlling the compromise between the model fit to the data and its complexity. Many authors have been interested in regularization methods. These methods were first implemented in a linear framework [2, 3, 4, 5] and later adapted to the survival framework [6, 7, 8, 9, 10, 11]. The idea of regularization methods, such as the Lasso method and its derivatives, is to penalize the likelihood so that irrelevant variables are set to zero.

However, when the number of covariates is much larger than the number of patients, regularization procedures haw proves unstable [12, 13]. In the regularization procedures, the regularization hyperparameter λ is obtained by cross-validation, which implies a random choice of sub-samples. Thus, if we run the procedure several times and particularly in high dimension, we can observe that we do not obtain at each time the same set of selected variables. This point is critical since we are unsure if the selected variables are relevant (false positives) or if we have missed some important ones (false negatives). To address this problem, [12] have proposed screening methods. The idea is to coarsely reduce the number of variables using a score before applying finer regularisation methods on a high but smaller number of pre-selected variables. These methods of screening differ in their pre-selection procedure, we present in Section 2.3 the methods SIS [9] and ISIS [9], CoxCS [10] and PSIS [11].

This paper aims to study the regularization and screening methods to examine the quality of their selection and their stability in selection for the Cox model. To evaluate the stability of such methods, we consider similarity indices such as generalizations of the Sørensen and the Jaccard indexes [14, 15, 16]. These indices are used in ecology to measure the variability of species composition in different sites, and we propose a new index. We offer to adapt them to our framework to measure the regularization and screening methods' stability. We also introduce a new index inspired by the F-Score in classification and which is a function of the potential (and unknown) number of significant variables.

The paper is organized as follows. In Section 2, we recall some basics and notation on the Cox model and present some regularization and screening methods developed for this model. To compare the stability of the different methods, we propose to use similarity indexes presented in Section 3. We first assess the methods on simulated datasets. We present the results of these simulations in terms of selection quality in Section 4. Finally, we apply and study the stability of the different methods on a real dataset of Clear Cell Renal Carcinoma, in Section 5, and discuss some genes of interest highlighted by the selection methods.

2 Different selection methods and their application to the Cox model

2.1 Cox model in high dimension

The Cox model [1] is a classical model in the field of survival analysis predicting the survival time from covariates. It allows us to study the time elapsed until an event of interest.

The Cox model is defined for an individual i from the instantaneous risk λ which is a function of time conditional on the explanatory variables given for the individual i $X_{i.} = (X_{i1}, \ldots, X_{ip})^T \in \mathbb{R}^p$:

$$\lambda(t|X_{i.}) = \alpha_0(t) \exp(\beta^T X_{i.}), \tag{4}$$

with $\alpha_0(t)$ baseline risk and $\beta = (\beta_1, ..., \beta_p)^T \in \mathbb{R}^p$ the vector of regression coefficients. The baseline risk is the instantaneous risk of death when all variables are zero. This function $\lambda(t|X_{i.})$ corresponds to the instantaneous risk of death at time t knowing that the individual i is alive before time t. We can separate the instantaneous risk into two parts because the term $\alpha_0(t)$ depends only on time and will be the same for all individuals at a given time. The second term in (4) depends only on the variables specific to each individual. The instantaneous risk between two individuals depends only on the factors to which they are subjected. This characteristic is useful when we are interested in prognosis, i.e., when we only want to know the factors influencing survival.

The Cox model [1] is semi-parametric because the estimation involves the estimation of a vector of parameters of $\mathbb{R}^p$ and a function $\alpha_0(t)$. But it is possible to estimate β without needing to know the baseline risk function $\alpha_0(t)$ thanks to

Cox's partial likelihood [1]. The partial likelihood of Cox [17] is based on the probability that an individual i dies at an observed time knowing that a death occurs. Let $t_1, \ldots, t_n$ be the set of observed times ordered for n individuals and $R(t_i)$ is the set of individuals at risk at time t_i, the probability that the individual i dies knowing that an event occurs at time t_i is:

$$\frac{\exp(\beta^T X_{i.})}{\sum_{l \in R(t_i)} \exp(\beta^T X_{l.})}. \tag{5}$$

The partial likelihood of Cox [1] is part of the total likelihood that does not dependent on the baseline risk function $\alpha_0(t)$. It is written:

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(\beta^T X_{i.})}{\sum_{l \in R(t_i)} \exp(\beta^T X_{l.})} \right], \tag{6}$$

with $R(t_i)$ the individuals at risk at time t_i. Maximizing $\mathcal{L}(\beta)$ allows in reasonable size to estimate correctly the parameter β of Cox's model. In high dimension (*i.e.* when the number of variables is greater than the sample size), the classical estimation procedure consisting of maximizing the Cox's partial log-likelihood no longer works. The solution, therefore, consists of minimizing the opposite of the Cox partial log-likelihood [17] to which we add a penalty term as in (3) with the partial log-likelihood L instead of the log-likelihood $\mathcal{L}$. The addition of this penalty term solves the optimization problem by encouraging a smaller and more easily interpretable model for the large dimension. Many penalties exist with different interpretability properties. We refer to [18] for more details on the penalty concept. These penalized functions are also called regularization methods, we will use this term throughout the paper, and we present in Section 2.2 two of them: the Lasso [2, 6] and the Adaptive-Lasso [3, 7]. However, these regularization methods may be unstable in variable selection [13]. Other methods have therefore appeared, called screening methods. Their general idea is to make a pre-selection before applying a regularization procedure such as a Lasso [6] for example. These methods of screening differ by the pre-selection used, we present in Section 2.3 the methods SIS [9] and ISIS [9], coxCS [10] and PSIS [11].

2.2 Regularization methods

2.2.1 The Lasso

The Lasso procedure was first introduced in the framework of a linear regression model by [2] and then in the survival analysis field by [6]. This procedure is classical in large dimensions and is the most known and most used. Its penalization is in the form of:

$$\begin{aligned} pen(\beta) &= \Gamma |||\beta||_1 \\ &= \Gamma \left(\sum_{j=1}^{p} |\beta_j| \right). \end{aligned} \tag{7}$$

The Lasso estimator of the parameter β is obtained by considering the following problem:

$$\widehat{\beta}^{l_1} = \arg\min_{\beta} \left\{ -L(\beta) + \Gamma \sum_{j=1}^{p} |\beta_j| \right\}, \tag{8}$$

where $L(\beta)$ is the Cox's partial log-likelihood defined by (6). This optimization problem is convex in β and thus allows the use of convex optimization algorithms for the estimation of β. The optimization problem is thus equivalent to minimize the log partial likelihood of Cox [1] by adding a constraint of the type:

$$\sum_{j=1}^{p} |\beta_j| \leq s,$$

with $s \in \mathbb{R}^+$. This amounts to constraining β to be in a ball of standard l_1 radius s in $\mathbb{R}^p$. The obtained $\widehat{\beta}^{l_1}$ estimator is then sparse, that is to say, that a certain number of coefficients of $\widehat{\beta}^{l_1}$ are zero. This gives it interpretability in the variable selection which is our objective in this paper.

2.2.2 The Adaptive-Lasso

During the presentation of the Lasso [6] regularization method, we recalled that it is unstable in selection. Indeed, the Lasso will select them randomly among two variables strongly correlated if they both have an effect on the variable to be explained. Therefore, [3] proposed an adaptive version of the Lasso, called Adaptive-Lasso. This procedure was later extended to the Cox model by [7]. This regularization penalizes large coefficients less than smaller ones. This is achieved thanks to the penalization used, which is a weighted Lasso penalty (standard l_1):

$$pen(\beta) = \Gamma \sum_{j=1}^{p} w_j |\beta_j|, \tag{9}$$

with $w_j = \frac{1}{|\widehat{\beta}_j|^\gamma}$ where $\gamma > 0$ and the $\widehat{\beta}_j$ are the coordinates obtained by an estimator in a preliminary step. In the work presented in this paper, we have used as a preliminary estimator $\widehat{\beta}$ the Lasso estimator $\widehat{\beta}^{l_1}$. The penalty used is therefore of the form:

$$pen(\beta) = \Gamma \sum_{j=1}^{p} \frac{|\beta_j|}{|\widehat{\beta}_j^{l_1}| + \epsilon},$$

where ϵ is the minimum of the non-zero $\widehat{\beta}_j^{l_1}$. This constant ϵ is added to the denominator to avoid dividing by zero in practice. The Adaptive-Lasso [7] procedure thus corresponds to the minimization of the Cox partial log-likelihood with the addition of a weighted Lasso penalty:

$$\widehat{\beta}^{l_{ada}} = \arg\min_{\beta} \left\{ -\sum_{i=1}^{n} \left(\beta^T X_{i.}\right) - \sum_{i=1}^{n} \delta_i \log \left(\sum_{l \in R_{i.}} \exp\left(\beta^T X_{l.}\right) \right) + \Gamma \sum_{j=1}^{p} \frac{|\beta_j|}{|\widehat{\beta}_j^{l_1}| + \epsilon} \right\},$$

with $R_{i.}$ the individuals at risk at time $t_{i.}$, δ_i the censoring indicator and the regularization parameter $\Gamma \in \mathbb{R}^+$ chosen by cross-validation. The Adaptive-Lasso estimator $\widehat{\beta}^{l_{ada}}$ is better in variable selection but is more biased than Lasso. The solution consists of re-running the unpenalized classical estimation procedure with only the selected variables by adaptive-lasso to reduce the bias.

2.3 Screening methods

The regularization methods are considered unstable in selection [13]. Moreover, this instability phenomenon reinforces when the number of variables increases significantly, which is the case with molecular data. There are several screening methods whose general principle summarizes in two steps. The first one consists of reducing the number of variables by keeping only those with a score obtained from the Cox model (and specific to each method) superior to a certain threshold. The second step consists of executing a Lasso procedure to select the most significant variables among those chosen in the first step.

2.3.1 (I)SIS methods

[8] introduce the SIS and ISIS methods, and a *package* R has been realized in which we can find the different variants of SIS and ISIS.

SIS

For each variable, a score is calculated, called *marginal utility*, which corresponds to the estimation of the regression coefficients by maximum of the log partial likelihood of the marginal model containing only the m^e variable:

$$u_m = \arg\max_{\beta_m} \left(\sum_{i=1}^{n} (\delta_i \beta_m X_{im}) - \sum_{i=1}^{n} \delta_i \log \left(\sum_{j \in R_{i.}} \exp(\beta_m X_{jm}) \right) \right).$$

The variables are ranked according to the value of the score in descending order. Then, the first d variables with the highest score are selected and their indices will correspond to the set $\mathcal{I}$. The set of these variables constitutes the first selected model. However, we cannot know the order of magnitude of this set and by choosing a value too large for d (the choice of the d value is discussed at the end of this section), the model could contain non-important variables. Reducing the model to the size d allows us to apply the Lasso penalty on the Cox model's partial log-likelihood optimization problem:

$$\arg\max_{\beta_{\mathcal{I}}} \left(\sum_{i=1}^{n} \delta_i \beta_{\mathcal{I}}^T X_{\mathcal{I}} + \sum_{i=1}^{n} \delta_i \log \left(\sum \exp(\beta_{\mathcal{I}}^T X_{\mathcal{I}}) \right) - \Gamma \sum_{m \in \mathcal{I}} |\beta_m| \right),$$

where $\beta_{\mathcal{I}}$ is the vector of regression parameters that correspond to the variables whose indices belong to $\mathcal{I}$. The Lasso procedure is used to set the non-informative variables to zero. The final model will be composed of the variables whose parameters are non-zero, the set of selected variables of the model is noted $\widehat{\mathcal{M}}$ and the estimated coefficients are noted $\beta_{\widehat{\mathcal{M}}}$.

ISIS

The SIS method may not perform well when some significant variables are uncorrelated with the variable to explain. Two uncorrelation situations exist. The first one is when variables are correlated with each other but do not individually impact strongly on the variable to explain. The second is when variables are not necessarily linked with each other but will separately have a more relevant impact on the variable to explain than some significant variables. An iterative version of SIS solves this problem by comparing the selected at each step with the variables of the model chosen by SIS. ISIS, therefore, tries to make more use of the joint information of the variables. The procedure is as follows:

1. The SIS method is initially applied to all the variables of the initial model. The k_1 variables with the highest score are selected. The obtained model will be $\widehat{\mathcal{M}}_1$ of dimension $|\widehat{\mathcal{M}}_1|$ and the set of indices of the variables belonging to the model $\widehat{\mathcal{M}}_1$ is $\mathcal{I}_1$.

2. The set of indices of the unselected variables is noted $\mathcal{I}_1^C$. For each $m \in \mathcal{I}_1^C$, a new score is computed, called *conditional utility*:

$$u_{m|\widehat{\mathcal{M}}_1} = \underset{\beta_m, \beta_{\widehat{\mathcal{M}}_1}}{\arg\max} \left[\sum_{i=1}^{n} \delta_i (\beta_m X_{im} + \beta^T_{\widehat{\mathcal{M}}_1} X_{i\widehat{\mathcal{M}}_1}) - \sum_{i=1}^{n} \delta_i \left\{ \log \sum_{j \in R_i.} \exp(\beta_m X_{jm} + \beta^T_{\widehat{\mathcal{M}}_1} X_{i\widehat{\mathcal{M}}_1}) \right\} \right] .$$

   We compute the score for unselected variables of the first step the score by taking into account the $\widehat{\mathcal{M}}_1$ model. We rank each variable according to this score. The k_2 variables with the highest score are selected. The set $\mathcal{I}_2$ contains the indices of the selected k_2 variables and is called relative set.

3. Next, we maximize the partial log-likelihood of the Cox model with a Lasso penalty term for the two sets of variables selected in steps 1 and 2 ($\mathcal{I}_1 \cap \mathcal{I}_2$). Variables with non-zero coefficients will be selected and will constitute the final model noted $\widehat{\mathcal{M}}_2$.

4. Finally, step 2 (with k_i) and 3 are repeated until the cardinality of the final model reaches the value d defined upstream or until $\widehat{\mathcal{M}}_i = \widehat{\mathcal{M}}_{i+1}$.

For the SIS and ISIS methods, we must define a threshold value d corresponding to the maximum number of variables selected in the model. But this d value is difficult to choose. [19] suggested in the paper accompanying their *package* R SIS to set $d = \lfloor \frac{n}{4\log(n)} \rfloor$ as part of the censored survival data. This parameter base on experiments and its choice is not justified. Other screening procedures have emerged as PSIS avoiding choosing the value of d.

2.3.2 PSIS

The PSIS method is a screening method developed by [11]. This method has similarities with the SIS method, but the score calculation is different between the two methods. The originality consists of computing a threshold to select the number of variables in the intermediate model. The choice of the threshold is justified by [11] to control false positives. The steps of the PSIS method are:

1. The regression coefficients are estimated individually for each variable by maximizing the Cox partial log-likelihood. The estimate of the regression coefficients β_j and the estimate of the variance of the estimated coefficients $\hat{\beta}$ (calculated from the inverse of the Fisher information matrix I) $I_j(\widehat{\beta}_j)^{-1}$ are retrieved.

2. The rate of false positives is fixed, we introduce: $q_n = f/p_n$, where f is the number of tolerated false positives and p_n is the dimension of the variables and thus a threshold γ is calculated as follows: $\gamma = \phi^{-1}(1 - q_n/2)$, where ϕ is the distribution function of the normal distribution.

3. The variables are then classified according to their score value. This score is calculated from the estimated regression coefficients and the variance of these coefficients: it is equal to $I_j(\widehat{\beta}_j)^{1/2}|\widehat{\beta}_j|$. Variables whose score is higher than the γ threshold are selected. These variables, therefore, belong to the intermediate model.

4. Finally, a maximum estimate of the partial log-likelihood associated with a Lasso penalty term is performed on the variables belonging to the intermediate model. Variables with non-zero coefficients are selected and correspond to the final model.

The advantage of this procedure is the property of false-positive control, i.e., it ensures that the false positive rate will be lower than the allowed rate (f/p_n). Although this method base on false-positive control, it does not take into account the false-negative rate. Controlling the latter would avoid omitting some essential variables and thus avoid obtaining an uninformative model. Finally, the non-iterative nature of PSIS leads to the same problems as those generated by SIS in the case of variables that are not correlated to the variable to explain but still influence the variable to explain through joint correlation with other variables.

2.3.3 CoxCS

The CoxCS method [10] allows adding in the screening procedure some covariates that are known to influence the variable of interest. These pre-selected covariates are then always selected by the screening procedure and can help find other relevant covariates. In practice, the CoxCS method uses biological knowledge to make the pre-selection. After the addition of biological knowledge, the procedure is the same as for PSIS.

3 Evaluation of the selection methods

Regularization methods are known as unstable in high-dimension, and screening methods try to answer this problem. In this paper, we try to quantify the stability of these methods from two similarity indexes, such as the Sørensen and Jaccard index [14, 15, 16]. Initially used in ecology to measure the variability of species composition in different sites, we propose to use them in the framework of variable selection. In the context of our study, we want to measure the variation in selection by the different methods run several times. Indeed, the choice of genes by the other methods differs for different *seeds*. A *seed* is an integer used to initialize a random number generator. However, regularization methods perform cross-validation to choose the Γ hyperparameter of the penalty criterion. Cross-validation is made by dividing the sample into k sub-samples, $k-1$ sub-samples will constitute the training set, and the last sub-sample will be the testing set). The filling of these k sub-parts is done randomly, according to the chosen *seed*.

The approach we have followed is to run the methods on 100 different *seeds*, and we have created a matrix within a row representing a *seed*, and a column representing a gene. If the gene j is selected for the i *seed*, then the coefficient (i, j) of the matrix will be worth 1 otherwise it is worth 0.

3.1 Sørensen Index

We use the Sørensen index to measure the similarity of the selected genes between the different *seeds*. The Sørensen index allows us to compare the *seeds* by considering the presence or absence of genes. It corresponds to the ratio between the overlap of selected genes by the different *seeds* and the average number of selected genes.

Let N be the number of genes selected by at least one *seed* and S the number of *seeds*. Let E_i be the set of genes selected by the *seed* i, let $|E_i| = n_i$ be its cardinal. Conversely, we note s_j the number of *seeds* for which the gene j is selected.

If there are only two *seeds*, the Sørensen index interprets in a set term as the ratio between the size of the intersection of the selected gene sets and the average size of the sets. The Sørensen index is thus given by:

$$S_2 = \frac{|E_1 \cap E_2|}{\frac{1}{2}\left(|E_1| + |E_2|\right)}$$

A gene being in $E_1 \cap E_2$ is said to belong to a collection. A gene not belonging to $E_1 \cap E_2$ is not in any collection. The belonging of a gene is generalized to a larger number of *seeds*. Therefore, the number of overlaps of selection sets to which a j gene belongs is simply $s_j - 1$. In the optimal case, the size of this overlay would be $S - 1$, and the overlay rate for the j gene is $(s_j - 1)/(S - 1)$.

Finally, the recovery measure is the sum of all genes of this recovery rate. In the case where $S = 2$, this recovery is directly the size of the intersection

$|E_1 \cap E_2| = \sum_{j=1}^{N}(s_j - 1)/(S - 1)$

Then, S_2 rewrites:

$$S_2 = \frac{\sum_{j=1}^{N}(s_j - 1)}{\frac{1}{2}\left(n_1 + n_2\right)}$$

To generalize to a larger number of *seeds*, divide the overlap measure by the average size overall selection sets:

$$S_S = \frac{\frac{1}{S-1}\sum_{j=1}^{N}(s_j - 1)}{\frac{1}{S}\sum_{i=1}^{S} n_i}.$$

In the case where all the *seeds* select the same N genes, the denominator (i.e., the overlap) is worth N, which is also the average size of the sets, so the index is worth 1. Conversely, if each gene is selected only once, $s_j = 1$ for all j, and the index is worth 0.

3.2 Jaccard index

The Jaccard index also enables to compute the similarity of gene selection between the different *seeds*. The Sørensen index tends to indicate nested selection scenarios as more stable, even if the number of variables varies, whereas the Jaccard index penalizes this type of scenario. The Jaccard index divides the number of genes shared by all two seeds samples with the total number of genes present in all samples seeds:

$$J_2 = \frac{a}{a+b+c}. \tag{10}$$

Let N be the number of genes selected by at least one *seed* and S be the number of *seeds*. We now present these two indexes in multiple cases. We recall that S is the number of seeds and N is the number of genes observed in at least one site. We call E_i the set of genes observed in the seed site i, we note $|E_i| = n_i$ its cardinal. Conversely, one notes s_j the number of seeds where the gene j is present.

By considering two seeds, the Jaccard index interprets as the ratio between the intersection size of the observed gene sets and the union size on of the sets of observed genes:

$$J_2 = \frac{|E_1 \cap E_2|}{(|E_1| + |E_2| - |E_1 \cap E_2|)}.$$

If a gene belongs to $E_1 \cap E_2$, it is said to belong to recovery. If it does not belong to $E_1 \cap E_2$, it does not belong to any recovery. The gene belonging generalizes to a larger number of seeds. Therefore, the number of overlaps of observation sets to which a genes j belongs is simply $s_j - 1$. In the optimal case, the size of this recovery would be $S - 1$, and it is called the recovery rate for the gene j: $(s_j - 1)/(S - 1)$. Finally, the recovery measure is the sum of all cash of this recovery rate. In the case where $S = 2$, this recovery is directly the size of the intersection

$|E_1 \cap E_2| = \sum_{j=1}^{N}(s_j - 1)/(S - 1)$.

and J_2 is rewritten:

$$J_2 = \frac{\sum_{j=1}^{N}(s_j - 1)}{N}.$$

By generalizing to a larger number of seeds, the overlap measure divides by the union size of the sets for the Jaccard index:

$$J_S = \frac{\frac{1}{S-1}\sum_{j=1}^{N}(s_j - 1)}{N}.$$

If we observe the same number N at all sites, the denominator (i.e., the recovery) is worth N, and this is also the average size of the sets, so the index is worth 1. Conversely, if we observe each species only once, $s_j = 1$ for all j, and the index is worth 0.

3.3 F_{score} metrics

The Sørensen and Jaccard indexes enable checking the stability of the regularization and screening methods. But the values of these indexes depend on the number of selected covariates. We are interested in two other metrics: F_{score} and $F_{score}(n^\star)$. The F_{score}, detailed in Section 3.3.1, enables to check the quality of the selection by combining the recall and the precision. If the value of F_{score} is close to 1, better is the quality of the variable selection. The $F_{score}(n^\star)$, introduced in Section 3.3.2, enables the quantification of the stability of selection by taking into account the number of true pertinent covariates selected. If this value is close to 1, better is the stability of the selection. For our study of stability, it is so important to look at the F_{score} and the $F_{score}(n^\star)$ together.

3.3.1 The classical F_{score}

The classical F_{score} [20] is a metric combining the precision and the recall. The precision is the ratio between the number of true positives and the number of considered positives (true positives and false positives). The recall is the ratio between the number of true positives and the number of positives (false negatives and true positives).

3.3.2 New index: F_{score} based on the number of hypothetical true covariates

We propose a new index, called $F_{score}(n^*)$, based on the number of the "true" variables that influence the survival time to compare on simulations the similarity indices on the different methods. We suppose that the selection sets are nested (assuming a proper index rearrangement):

$n_1 \leq n_2 \leq \dots \leq n_S = N$

Let's denote by n^* the real number of significant features, and we also suppose that the total N genes selected contain these features.

s_i is considered a proportion (the number of times gene i is selected divided by the number of experiments)

$num(s)$ is the number of genes that is selected at least s times, $num(0^+) = N$.

We suppose that the methods are "consistent": the bigger s_i, the more probable the i^{th} gene is a truly important one, such that the genes can be ranked according to s. With the proper index rearrangement, we can suppose that:

$$s_1 \geq s_2 \geq \dots \geq s_N, \text{ and } s_i = 0, \forall i > N$$

Let n^* be the number of true positive genes.

Let us denote $s^* = \begin{cases} \arg\max \{s \,|\, num(s) \geq n^*\} = s_{n^*} \text{ if } N \geq n^* \\ 0 \text{ otherwise} \end{cases}$

As a consequence, we also extrapolate by saying that, if $N < n^*$, then all genes such that $s_i > 0$ are true positive genes, while if $N \geq n^*$, then all genes such that $s_i \geq s^*$ are true positive genes.

Based on the assumptions that these genes are positive, we would expect that for all of them, $s_i = 1$, that is to say, that they are selected all the time. Therefore, $(1 - s_i)$ is the proportion of times that the gene is a false negative, and $\sum_{1 \leq i \leq n^*}(1 - s_i)$ is the average number of a false negative.

Note that the formula remains valid when $n^* > N$, since the $n^* - N$ genes for which $s_i = 0$ are missed.

We then deduce the True Positive as the positive minus the false negative, and the average predicted positive as $\sum_{i=1}^{N} s_i$.

We can finally write Precision and Recall:

$$Precision(n^*) = \frac{n^* - \sum_{i=1}^{n^*}(1 - s_i)}{\sum_{i=1}^{N} s_i} = \frac{\sum_{i=1}^{n^*} s_i}{\sum_{i=1}^{N} s_i}$$

$$Recall(n^*) = \frac{n^* - \sum_{1 \leq i \leq n^*}(1 - s_i)}{n^*} = \frac{\sum_{i=1}^{n^*} s_i}{n^*}$$

A traditional measure to measure the compromise between precision and recall is their harmonic mean (F-Score), which would read:

$$F_{score}(n^*) = 2\frac{Precision(n^*)Recall(n^*)}{Precision(n^*) + Recall(n^*)}$$

This new index depends on the "true" variables that influence the survival time. How can this index be used in practice? First, the idea in this paper is to compare the selection methods on simulations. In this case, we know which variables are relevant. We can compare the strategies from this new index with n^* known. This new index gives us an indication of the stability of the different methods in different configurations. However, when we work with real datasets, we do not know the number of "true" variables that influence the survival time. In this case, we can vary our index according to the number n^* of "true" relevant variables.

Moreover, we are interested in other metrics as AIC and BIC, and we also look at the number of selected genes presented in Section 3.4.

3.4 Supplementary criteria

We also calculated the Akaike Information Criterion (called AIC for *Akaike Information Criterion*) of the model obtained for each *seed*. The AIC criterion is a metric developed by [21] and enables the evaluation of the quality of a model. It allows us to manage both the quality of the fit and the complexity of the model by penalizing models with a large number of parameters. The best model will be the one with the lowest value of the AIC criterion. Therefore, this criterion is based on a compromise between the fit quality and the model complexity by penalizing models with a large number of parameters, limiting the effects of over-fitting (increasing the number of parameters necessarily improves the quality of the fit). We calculate the mean and standard deviation of the AIC criterion for each of the methods over the 100 *seeds*. This calculation also allows us to judge the quality of the selection.

For the simulated datasets, we also compute the classical F-score [20], introduced in Section 3.3.1, as we know the true positives.

From these different criteria, we can measure the various selection methods' performances in terms of stability and quality of the selection.

4 Tests on simulated data

First, we want to compare the regularization and screening method on simulated data when we know exactly which variables influence survival time.

4.1 Simulations

We simulated two datasets from the R package. We generate the survival times of these datasets from a Cox model where the baseline hazard function is modeled by a Weibull distribution $\mathcal{W}(a, b)$. The number of parameters is different between the two datasets: 1000 covariates for the first one and 25 000 covariates for the second one. The number of relevant covariates is equal in both cases to 20.

The simulation of data from the Cox model in the R package base on [22]. We have chosen to carry out this simulation to generate survival data that respects the proportional risk hypothesis. The generation of survival data from a Cox model base on:

$$T = H_0^{-1}\left[\frac{-\log(1-U)}{\exp(\beta^T X_{i.})}\right], \tag{11}$$

where $U \sim \mathcal{U}[0,1]$ and $X_{i.} \sim \mathcal{U}[-1,1]$. For this simulation, we consider that survival times follow a Weibull distribution $\mathcal{W}(a, b)$. In this case, we have the cumulative risk function expressed by:

$$H_0(t) = bt^a \tag{12}$$

and survival times can therefore be simulated from:

$$T = \frac{1}{b^{1/a}}\left(\frac{-\log(1-U)}{\exp(\beta^T X_{i.})}\right)^{1/a}. \tag{13}$$

We take for these simulations $a = 1.969765$ and $b = 7.586963e - 07$ to have a mean equal to 1134 and a median equal to 1134. The idea is to simulate data that are not so far from real data.

4.2 Stability analysis

We study the stability of regularization and screening methods, but also the validity of the selection. The Sorensen, Jaccard, and F-score(n*) indexes enable the evaluation of the stability, while the AIC criterion and the classical F-score allow us to judge the quality of the selection. Table 1 gives the results of these different indexes on the two simulated datasets described above.

For the CoxCS method, we have considered two cases. The first one, denoted CoxCS1, considers 5 among the 20 relevant covariates as pre-selected covariates to add knowledge in the screening procedure. The other one, called CoxCS2, considers 5 among the 20 relevant covariates and 5 covariates known to be non-relevant as pre-selected covariates. The idea for this second case is to see if the procedure gives good results even if we are wrong about the pre-selection.

First, we can see that regularization methods have similar results about stability. The values of Sørensen and Jaccard indexes are close between the Lasso and Adaptive-Lasso methods. Moreover, the value of $F_{score}(n^\star)$ is also similar for these methods. Indeed, the value of $F_{score}(n^{\star})$ is equal to 0.8506 for the Adaptive-Lasso and to 0.8304 for the Lasso for 1000 covariates. We can observe that the value of the $F_{score}(n^{\star})$ of methods is very good on datasets with 1000 covariates. But when the number of covariates increases, the regularization methods have bad results in terms of stability. Indeed, in the high dimensional case, the Adaptive-Lasso is the method with the highest value of the $F_{score}(n^{\star})$ equal to 0.3897, and this value is meager, concluding to a bad level of stability for this method. It is necessary to look at the $F_{score}(n^{\star})$ by comparing it with the F_{score} to quantify both the stability of selection and the quality of selection. We can see that the quality of selection is rather bad for the regularization methods. Indeed, the value of the F_{score} is around 0.38 for all regularization methods. We can also observe that the quality of selection decreases with the dimensionality of the datasets. The F_{score} for all regularization methods is around 0.38 for the dataset with 1000 covariates, and this score is null on the datasets containing 25000 covariates. The results confirm the problem of stability of regularization methods in high-dimension and show us that the quality of the selection of these methods is bad in high-dimension.

For the screening methods, the Sørensen and Jaccard indexes are not the most adapted. Indeed, we can see that SIS and ISIS's screening methods have perfect values of these indexes due to the number of selected covariates correct and do not change according to seeds. These indexes must be completed with other criteria to confirm the stability and the quality of selection. First, we can see by looking at the $F_{score}(n^{\star})$ that SIS and ISIS (with d = 20) have a perfect score, explained by the fact that we impose to these methods to choose not more than 2O covariates. But their values of F_{score} are low (0.35 for 1000 covariates and 0.05 for 25000 covariates), which reflects the fact that the 20 selected covariates are not the right ones. The values of F_{score} for SIS and ISIS are close to those of the regularization methods. Moreover, these values of the F_{score} decrease with the dimensionality of the datasets, despite the values of $F_{score}(n^{\star})$ do not decrease. The methods SIS and ISIS, initially developed to solve the stability problem of regularization methods, seem to maintain selection instability. The PSIS method does not give satisfactory results in terms of variable selection either. This study has shown us that the method coxCS seems to be a good method considering the stability and selection quality. As introduced above, we considered two cases for the coxCS method. The first one, denoted CoxCS1, considers 5 among the 20 relevant covariates as pre-selected covariates to add knowledge in the screening procedure. The other one, denoted CoxCS2 considered 5 among the 20 relevant covariates and 5 covariates known to be non-relevant as pre-selected covariates. We can see that the values of Sørensen and Jaccard indexes are very good for both cases of the coxCS method. Their values are lower than these of SIS and ISIS methods. But the values of the F_{score} and the $F_{score}(n^{\star})$ of the coxCS methods are high, contrary to SIS and ISIS methods. The value of $F_{score}(n^{\star})$ for coxCS1 (5 pertinent covariates in the pre-selection) is equal to 0.919 for 1000 covariates and the one for coxCS2 (5 pertinent covariates and 5 non-pertinent covariates in the pre-selection) is equal to 0.8307. This value slows a few when we add the non-relevant covariates to the pre-selection. This observation shows that knowledge can improve the stability of selection. Besides, this knowledge seems to improve also the quality of selection. The F_{score} of all methods except the coxCS method is lower than 0.5. The Fsocre of coxCS1 is 0.8456 for 1000 covariates and 0.7571 for 25000 covariates, which means that most of the selected covariates are true covariates. We observe the same behavior with coxCS2, where the F_{score} for 1000 covariates is 0.7526, and the one for 25000 covariates is 0.778. We can see that the values of F_{score} and $F_{score}(n^{\star})$ do not decrease or decrease slowly. CoxCS also seems to handle the high-dimensional data. The stability study shows that coxCS is a good method to make the variable selection in high-dimension when some knowledge is available. Finally, the study confirms that it is crucial to detect potential markers in survival analysis to discuss the domain's experts. Indeed, coxCS seems to be the best method in our stability study. However, the users must use the included knowledge in this method correctly. Moreover, for the choice of d in SIS and ISIS, it would seem relevant to consider a higher value of d than one is looking for because one loses little stability but one gains in quality of selection. We, therefore, recommend discussing with professionals in the domain to find out what strategy to consider in practice.

5 Tests on Real dataset

We apply the methods presented above to a real data set. This dataset concerns clear cell renal carcinoma cancer (ccRCC) from the TCGA database (*The Cancer Genome Atlas*). The data are available at `https://www.cancer.gov/tcga`.

From this database, we consider three sets of covariates for our study. First, we consider as covariates only the *Immune-Checkpoints* (IC), *i.e.* genes involved in the immune response process. There are 48 covariates in this first set. Then, we choose to work with the genes considered as differentially expressed in the study published in [23]. The number of genes found differentially expressed by the DESeq2 method [24] is 11 289 genes. Finally, we consider all the coding genes present in humans. After filtering to remove the null genes, we obtain 17 789 genes.

We applied both regularization methods and screening methods on these three sets of covariates. For the pre-selection of the CoxCS method for the second and third datasets, we had the idea to use the latest publications about markers capable of predicting patient survival. For this, we carried out a PubMed search using the following keywords and MeSH terms (for *Medical Subject Headings*): `ccRCC [tiab] prognosis [tiab] survival [tiab] genes NOT RNA`. MeSH is

Methods	#Covariates	Sorensen	Jaccard	#Genes selected	$F_{score}(n^\star)$ $(n^\star = 20)$	F_{score}
Lasso	1000	0.9939	0.6184	28 (6.9515)	0.8304	0.38
	25000	0.9869	0.4298	4.79 (2.5556)	0.3864	0
Adaptive-Lasso	1000	0.9937	0.6122	25.26 (7.0534)	0.8506	0.384
	25000	0.9892	0.4788	4.84 (2.2862)	0.3897	0
SIS (d=10)	1000	1	1	10 (0)	0.6667	0.4
	25000	1	1	10 (0)	0.6667	0.0667
SIS (d=20)	1000	1	1	20 (0)	1	0.35
	25000	1	1	20 (0)	1	0.05
SIS (d = 50)	1000	0.9996	0.9592	47.02 (1.8694)	0.5968	0.3581
	25000	1	1	48 (0)	0.5882	0.1471
SIS (d=100)	1000	0.9944	0.6388	53.96 (25.4899)	0.4911	0.2977
	25000	0.9998	0.978	92.93 (0.9018)	0.3542	0.124
ISIS (d =10)	1000	0.999	0.9082	10 (0)	0.6667	0.3333
	25000	1	1	10 (0)	0.6667	0
ISIS (d =20)	1000	0.9995	0.9519	20 (0)	0.999	0.35
	25000	1	1	20 (0)	1	0.05
ISIS (d = 50)	1000	0.999	0.9082	50 (0)	0.5714	0.2857
	25000	1	1	50 (0)	0.5714	0.1143
ISIS (d =100)	1000	0.9898	0.4924	100 (0)	0.3333	0.2005
	25000	0.999	0.9082	100 (0)	0.3333	0.1
PSIS	1000	1	1	6 (0)	0.4615	0.4615
	25000	0.9987	0.8869	4.44 (0.4989)	0.3633	0
coxCS1	1000	0.9988	0.8963	23.33 (1.9749)	0.919	0.8456
	25000	0.998	0.8352	20.92 (6.0963)	0.8876	0.7571
coxCS2	1000	0.9993	0.9377	28.15 (1.158)	0.8307	0.7526
	25000	0.9996	0.9641	27.97 (3.6803)	0.8197	0.778

Table 1: Summary of selection stability and quality by considering differents metics (Sørensen index, Jaccard index, F_{score}, $F_{score}(n^\star)$, the mean and the sd of selected covariate number) for the simulation study.

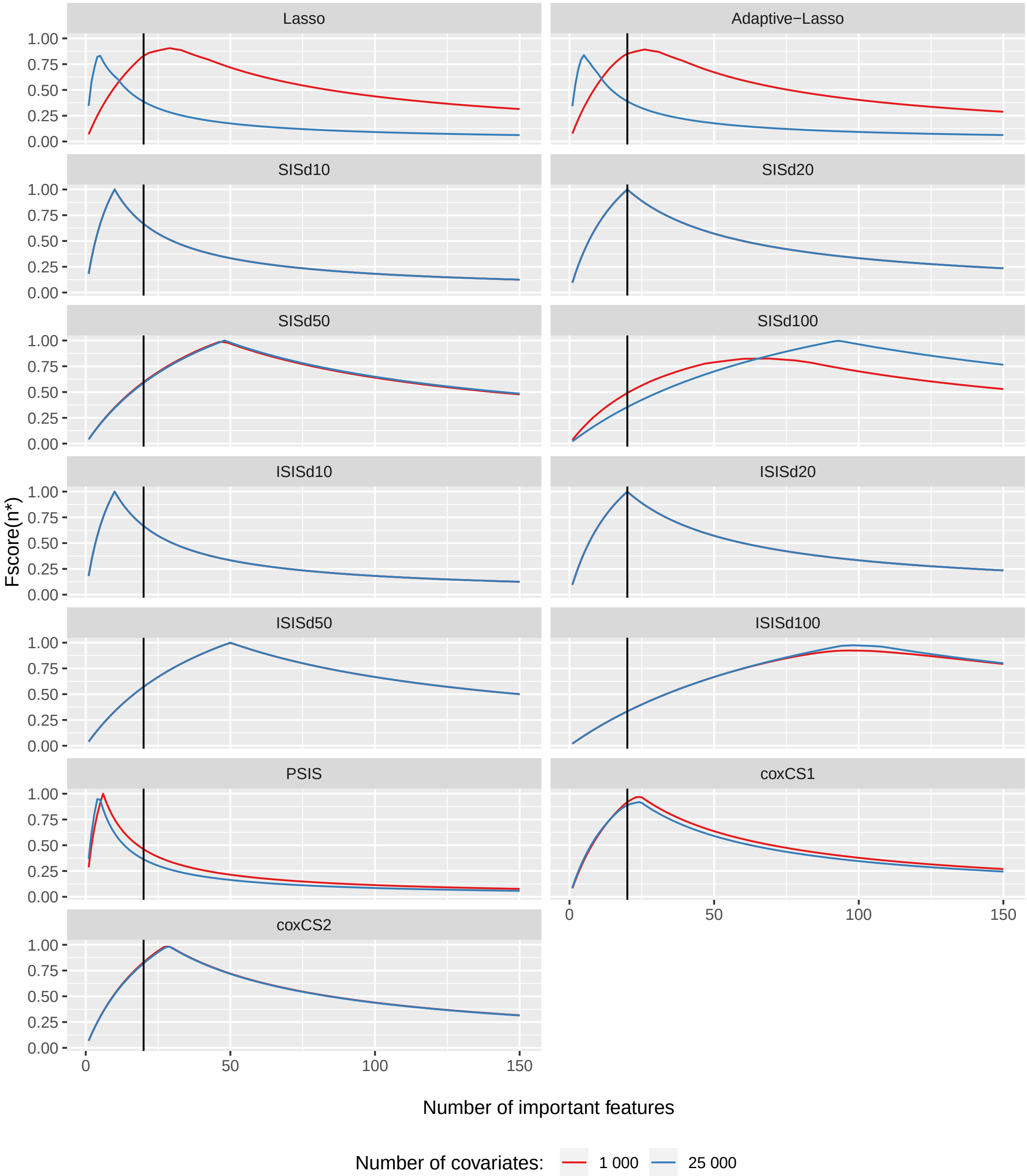

Figure 1: The $F_{score}(n^{\star})$ are plotted as a function of $n*$ for the different methods

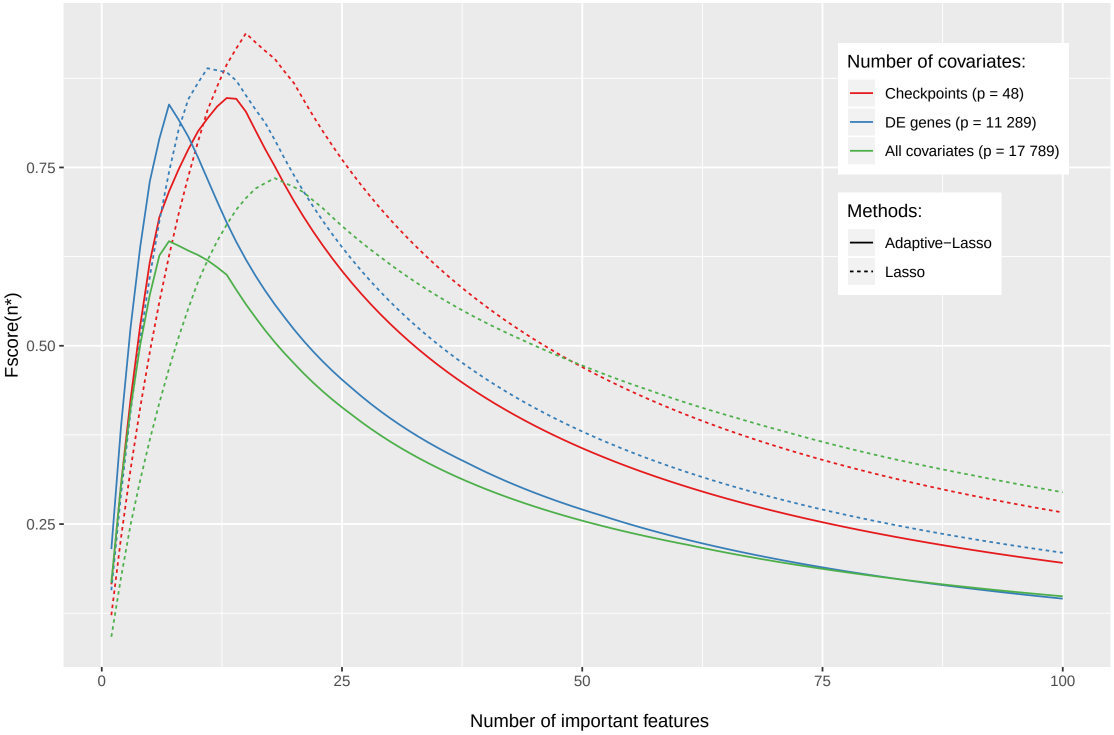

Figure 2: The $F_{score}(n^{\star})$ are plotted as a function of $n*$ for the different methods for the real datasets

a hierarchically organized and controlled vocabulary produced by the *National Library of Medicine*. We thus obtained 32 results corresponding to our search, and we decided to keep the first 10 *abstracts*. These 10 *abstracts* were saved and later loaded into the web application `beca annotate` nunesbecas2013. This web application aims to enable the identification and annotation of medical concepts in a text. The identification and annotation of genes and proteins are performed using *machine learning* methods present in Gimli camposgimli2013. Gimli camposgimli2013 is a *open-source* tool for the automatic recognition of biomedical terms. The list of genes obtained is: $KDM2B$, $HMGCS2$, $HSD11B1$, $IL10$, $KDM5B$, $KDM5A$, $KDM5C$, $KDM5D$, $KDM1B$, $OGDHL$, $SSBP2$, $VSIG4$ and $XCR1$. We decided to use this list as biological knowledge for pre-selection in coxCS for the real dataset.

We applied the methods on the 100 different *seeds* (seed is an integer used to initialize a random number generator) to study their stability and better interpret the selection. In addition to the calculation of similarity indices, we have also calculated a selection percentage for each gene (see Table 4 and Table 5 in the Supplementary material). This selection percentage allows us to avoid the conclusion that a gene that would have been selected only a few times, by chance, is important. We carried out the biological analysis of the selected genes from the site `GeneCards` [25] accessible at `https://www.genecards.org/` and from the site `COSMIC` [26] (for *Catalogue Of Somatic Mutations In Cancer*) available at `https://cancer.sanger.ac.uk/cosmic`. `GeneCards` is a database that provides complete information on all predicted and annotated human genes. `COSMIC` is also a database, but this one exclusively dedicates to cancer. For `COSMIC` we have particularly relied on the *Cancer Gene Census* catalog, which lists genes with mutations involved in cancer.

5.1 Stability analysis and marker discovery

5.1.1 Regularization methods

TABLE 2 presents the results of the Sørensen and Jaccard indexes, the mean number of selected genes (and the standard deviation), the $F_{score}(n^{\star})$ and the AIC for the Lasso and the Adaptive-Lasso for the three sets of covariates: the Immune-Checkpoints, the differential expressed covariates and all genes. As expected, the Sørensen index is high for the regularization methods when the number of covariates is small. When the number of covariates increases, the Adaptive-Lasso seems to be less stable than the Lasso. Indeed the more covariates we consider, the worse the Sørensen and Jaccard indexes are for the Adaptive-Lasso. The Lasso also has more difficulties when the number of covariates

		Lasso	Adaptive Lasso
Immune-Checkpoints	Sørensen index	0.9960	0.9933
	Jaccard index	0.73	0.60
	$F_{score}(n^\star)$ (n*=20)	0.8682	0.703
	Number of selected genes	15.36 (2.83)	10.84 (3.53)
	AIC	1915.50 (4.33)	1917.71 (11.06)
Differential expressed genes	Sørensen index	0.9946	0.9436
	Jaccard index	0.65	0.14
	$F_{score}(n^\star)$ (n*=20)	0.739	0.523
	Number of selected genes	11.72 (2.34)	7.84 (3.01)
	AIC	1867.35 (1.95)	1862.47 (23.04)
All genes	Sørensen index	0.9332	0.8284
	Jaccard index	0.12	0.05
	$F_{score}(n^\star)$ (n*=20)	0.7225	0.4754
	Number of selected genes	17.70 (3.57)	8.65 (3.64)
	AIC	1873.43 (24.95)	1870.42 (40.97)

Table 2: Results of similarity indexes (Sørensen and Jaccard index), of the average AIC and of the average of selected genes on the whole seeds for the studied regularization methods according to the set of genes given as input. The standard deviations are precised in the brackets.

becomes large. However, it seems to be still stable when we work with the differential expressed covariates, whereas the Adaptive-Lasso has rather bad similarity indexes in this case. However, these two indexes are not the most adapted because they depend strongly on the number of selected covariates. The index $F_{score}(n^\star)$ we have proposed enables us to vary the number of pertinent covariates and to see the effects of this number on the stability of selection. By considering $F_{score}(n^\star)$, the Lasso method seems to be the most stable. Figure 2 shows that the dotted curves are higher than the solid curves, confirming this observation. That means that the values of the $F_{score}(n^\star)$ for the Adaptive-Lasso are the highest and the most precise when the number of relevant covariates is larger than 10. However, the selection quality seems better for the Lasso method; these values are lower than those of the Adaptive-Lasso method. The values of AIC stay close between these two methods. We can also observe that the values of the indexes (Sørensen, Jaccard, $F_{score}(n^\star)$) decrease with the dimensionality of data, which confirms that the regularization methods are not stable in high-dimension.

However, we can see from TABLE 4 in the Supplementary material that many genes are selected in only one *seed* for the Lasso and Adaptive-Lasso methods. We can see that the Sørensen and Jaccard indexes do not highlight this phenomenon, which is not a good point for the stability of these methods.

5.1.2 Screening methods

We can see on TABLE 3 that the different screening methods give different results as far as stability is concerned. Indeed, if the SIS method seems very stable even when the number of variables increases, the other methods have more difficulties in high-dimension. The behavior of stability appears to be confirmed with the $F_{score}(n^\star)$ for the SIS and ISIS methods. Contrary to the values of Sørensen and Jaccard indexes, the values of $F_{score}(n^\star)$ for the coxCS and PSIS methods increase with the dimensionality of data. This observation confirms the drawback of Sørensen and Jaccard indexes due to their too strict or soft behavior. In Figure 3, we can see that SIS and ISIS always have best $F_{score}(n^\star)$ for a small $n^\star$ even when the number of covariates in the study increases, these methods are very sparse, but we know that they do not select the relevant covariates. The screening methods select very few variables, notably SIS and ISIS. For example, for the Immune-Checkpoints, the average of the variables chosen for SIS is 2.57. The Lasso, on the contrary, selects an average of 15.36 genes. However, the number of variables in this dataset is small, and we chose these variables early because they had a potential impact on clear cell Renal Cell Carcinoma (ccRCC). Many variables could therefore correlate with each other, and this could explain this lower selection stability result. PSIS and CoxCS select more variables when the number of covariates increases. We can nevertheless note that for the sets constitute with the 48 Immune-Checkpoints, the Lasso remains the most stable method. The ISIS method seems to be the worst stable method, and the CoxCS method based on the biological knowledge gives good stability results in terms of $F_{score}(n^{\star})$. However, the knowledge used for CoxCS is maybe not the best biological knowledge that we can add to the method. The performances of this method could be even better with better biological knowledge. From the AIC, we can also conclude that the variables selected by PSIS and coxCS seem to explain less well the survival duration than the other methods. This observation could also explain by a wrong chosen knowledge added to the method.

		SIS	ISIS	PSIS	coxCS
Immune-Checkpoints	Sørensen index	0.9708	0.6089	0.9983	0.9974
	Jaccard index	0.2495	0.798	0.8566	0.796
	$F_{score}(n^{\star})$ (n*=20)	0.2277	0.3532	0.6004	0.5704
	Number of selected genes	2.57 (2.23)	4.29 (0.69)	8.58 (0.57)	7.98 (2.01)
	AIC	1953.37 (21.36)	1935.35 (3.66)	1961.22 (4.33)	1946.50 (9.92)
Differential expressed genes	Sørensen index	0.9905	0.382	0.9662	0.8885
	Jaccard index	0.5101	0.5841	0.2207	0.8127
	$F_{score}(n^{\star})$ (n*=20)	0.3416	0.3519	0.8824	0.6188
	Number of selected genes	4.12 (1.57)	4.27 (1.02)	18.51 (5.11)	8.96 (1.47)
	AIC	1903.63 (7.78)	1895.25 (5.92)	1944.20 (5.27)	1960.39 (3.50)
All genes	Sørensen index	0.9962	0.8956	0.9610	0.9341
	Jaccard index	0.7222	0.6212	0.2492	0.1231
	$F_{score}(n^{\star})$ (n*=20)	0.4496	0.424	0.7494	0.7814
	Number of selected genes	5.80 (1.04)	5.39 (1.50)	27.21 (9.18)	25.85 (14.44)
	AIC	1873.80 (0.71)	1880.01 (24.69)	1931.38 (12.55)	1937.71 (13.69)

Table 3: Results of similarity indexes (Sørensen and Jaccard index), of the average AIC and of the average of selected genes on the whole seeds for the studied screening methods according to the set of genes given as input. The standard deviations are precised in the brackets.

5.2 Potential markers explaining the survival duration

Concerning the gene selection from the regularization methods, we refer the reader to Table 4 in the supplementary material. Several genes are often selected, and some have biological interpretability. First, we have the $FBXL5$ gene that is involved in the immune system. We report several phenotypes for this gene, one of which corresponds to chronic kidney disease. The location of this gene is on the same segment as the $BST1$ and $CD38$ *Immune-Checkpoints* on chromosome 4. The $FBXL5$ gene, therefore, appears to be an essential gene to explain survival for clear cell renal cell carcinoma. Then, the $CKAP4$ gene occurs in the selection of the different methods and is involved in the immune system. A disease associated with the $CKAP4$ gene is cystitis, an infection that affects the bladder, a part of the urinary system like the kidneys. These observations may reinforce the idea that the $CKAP4$ gene would be a useful marker for the survival of patients with kidney cancer. Although they are not methodically selected and/or selected with a small percentage, two other genes that we find interesting are the $CHEK2$ and $C10orf90$ genes, which are tumor suppressors. The function of a tumor suppressor gene is to prevent the runaway of cell division. If it is present in a cancer cell, it tends to slow down cell proliferation. These genes are, therefore, good indicators of a patient's survival to the disease in question.

We can also observe that the selected genes are most of the time the same considering only the differentially expressed genes and all the genes (cf. TABLE 4) for the Lasso and Adaptive-Lasso. For both methods, we have the minimum set $\{GDF5$, $CKAP4$,$CUBN$,$OTOF$,$SORBS2\}$. In this set, we have the genes $CKAP4$ that already appeared to be promising biomarkers of patient survival. The $FBXL5$ gene mentioned above as the right biomarker candidate is a selected gene by both methods. Finally, the $CUBN$ gene also appears to be a good biomarker to explain patient survival. It is a receptor located on the epithelial tissue of the intestine and kidney. The $CUBN$ may be a prognostic marker for clear cell renal cell carcinoma from the study of [27].

Concerning the gene selection from the screening methods, we can see that SIS and ISIS select the $CHEK2$ gene from Table 5. In Section 5.1.1, we already mentioned that it could be interesting as markers for the prognosis of patients

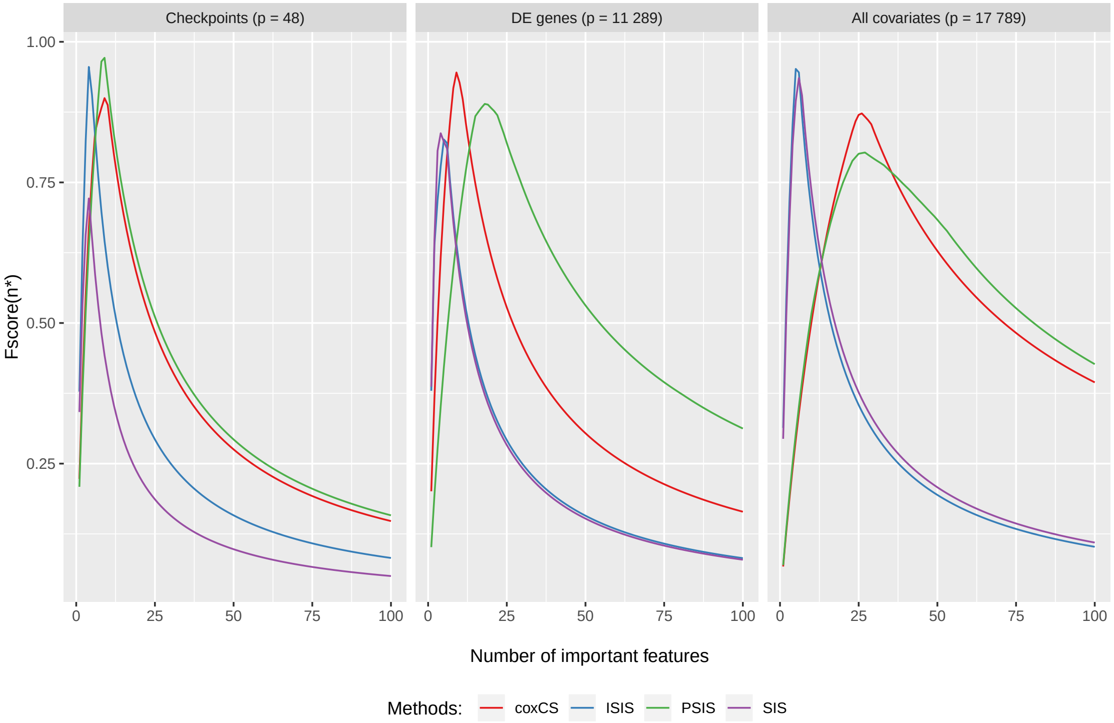

Figure 3: The $F_{score}(n^{\star})$ are plotted as a function of $n*$ for the different screening methods for the real datasets.

with ccRCC. Finally, we can notice that the screening method's biological knowledge has changed the gene selection compared to SIS, ISIS, and PSIS methods that do not use knowledge. Indeed, none of the genes selected in coxCS are found in SIS, ISIS or PSIS (cf. TABLE 5 in the supplementary material). It is more difficult to conclude on these results because none of the selected genes is by all methods simultaneously. One explanation of these bad results is that we introduce knowledge from a preliminary study on abstracts. Maybe this study lacks informed advice from a biologist or doctor.

6 Discussion

In this study, we have compared several feature selection methods in the Cox model, with a focus on high-dimension: Lasso, Adaptive-Lasso, SIS, ISIS, PSIS, CoxCS. These comparisons were achieved in a simulation study and on a real dataset. In the simulation study and for a low number of covariates, we showed that Lasso and adpative-Lasso are stable in selection from the values of the similarity indexes. However, the selection quality remains poor, the situation being better for adaptive-Lasso. But in high-dimension, Lasso and adpative-Lasso are both unstable and the quality of selection is also low. The screening methods seem to solve, partially, the problem of stability noticed for the regularization methods. Indeed, SIS and ISIS methods have rather good stability results, but the selection is incorrect. By choosing a pertinent value of the selection threshold d, we can improve the selection quality. It thus needs to be chosen carefully. CoxCS seems to be a very good method when some biological knowledge is available: in the simulation study, the selection stability of this method is good, and the selection quality is also correct. In this study, screening methods allow a more straightforward interpretation of the results. The number of variables selected by the screening methods is lower than that of the regularization methods.

On the real dataset, screening methods (PSIS and coxCS) appear to be more stable, especially for the highest dimension, when all genes are kept in the regression model. The $F_{score}(n^{\star})$ values for the screening and regularization methods are in agreement with this. The $F_{score}(n^{\star})$ values of the screening methods are higher in high dimensions than those of the regularization methods. Moreover, we can observe that the value of the $F_{score}(n^{\star})$ of regularization methods decreases with increasing dimension, which is not the case with screening methods. As mentioned in the simulation study, CoxCS is an attractive method because it also has good stability results by considering $F_{score}(n^{\star})$ on the real dataset. If there is no prior available biological information, then PSIS seems like the best compromise in terms of both stability and quality. Finally, we noticed that the Sørensen index tended to indicate nested selection scenarios as more stable, even if the number of variables varies. In contrast, the Jaccard index penalizes this type of design. That is why we proposed the $F_{score}(n^{\star})$, inspired by the F_{score}, allowing to consider a variable number of pertinent covariates, since it is generally unknown. A critical point to consider in future studies is to further explore why higher regularity indexes do not necessarily translate into a better behavior of the model in terms of AIC.

The study of the selection of variables to explain patient survival from *Immune-Checkpoints* shows that some of them would be interesting to study further. Indeed, the $B7$-$H3(CD276)$ gene seems to be a good biomarker to explain patient survival. It is also the case with $HLA.G$. We have seen in the study and in [23] that $HLA.G$ shares information with other genes. [23] raises the possibility that this $HLA.G/ILT2$ pair may be an alternative to $PD1/PDL1$ treatments when patients do not respond to the latter. A perspective from a biological point of view is to further study this couple.

By extending the variable selection study to differentially expressed genes and the set of genes as a whole, we were able to observe genes that could be of potential interest. It would be relevant to deepen their study. The $CHEK2$,$CKAP4$ genes appear to be important markers to explain survival. The $CHEK2$ gene is already known to have an impact on breast cancer and the $CKAP4$ gene is involved in the immune system. We recall that the clear cell renal carcinoma is an immunogenic cancer, which means that it can stop the immune system. Another gene that appears in the selection to explain the survival of patients is the $CUBN$, gene. In [27], this gene was said to play a key role in the development of clear cell renal carcinoma. Finally, the $FBXL5$ gene also appears to be a potential marker to explain patient survival as it has a role in the immune system, is involved in chronic kidney disease, and is thought to be related to two *Immune-Checkpoints* listed in the study by [23].

References

[1] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[2] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[3] Hui Zou. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[4] Pierre J. M. Verweij and Hans C. Van Houwelingen. Penalized likelihood in Cox regression. *Statistics in Medicine*, 13(23-24):2427–2436, 1994.

[5] Jianqing Fan and Runze Li. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[6] Robert Tibshirani. The Lasso Method for Variable Selection in the Cox Model. *Statistics in Medicine*, 16(4):385–395, 1997.

[7] H. H. Zhang and W. Lu. Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, 94(3):691–703, 2007.

[8] Jianqing Fan, Yang Feng, and Yichao Wu. *High-dimensional variable selection for Cox's proportional hazards model*. Institute of Mathematical Statistics, 2010.

[9] Jianqing Fan and Rui Song. Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, 38(6):3567–3604, 2010.

[10] Hyokyoung G. Hong, Jian Kang, and Yi Li. Conditional screening for ultra-high dimensional covariates with survival outcomes. *Lifetime Data Analysis*, 24(1):45–71, 2018.

[11] Sihai Dave Zhao and Yi Li. Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis*, 105(1):397–411, 2012.

[12] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[13] Stefan Michiels, Serge Koscielny, and Catherine Hill. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet (London, England)*, 365(9458):488–492, 2005.

[14] Héctor T. Arita. Multisite and multispecies measures of overlap, co-occurrence, and co-diversity. *Ecography*, 40(6):709–718, 2017.

[15] Andrés Baselga. Multiple site dissimilarity quantifies compositional heterogeneity among several sites, while average pairwise dissimilarity may be misleading. *Ecography*, 36(2):124–128, 2013.

[16] Andrés Baselga. Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography*, 19(1):134–143, 2010.

[17] D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

[18] Peter J. Bickel, Bo Li, Alexandre B. Tsybakov, Sara A. van de Geer, Bin Yu, Teófilo Valdés, Carlos Rivero, Jianqing Fan, and Aad van der Vaart. Regularization in statistics. *Test*, 15(2):271–344, 2006.

[19] Diego Franco Saldana and Yang Feng. SIS : An R Package for Sure Independence Screening in Ultrahigh-Dimensional Statistical Models. *Journal of Statistical Software*, 83(2), 2018.

[20] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge University Press, New York, 2008. OCLC: ocn190786122.

[21] Hirotogu Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer Science+Business Media, New York, 1998.

[22] Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723, 2005.

[23] Diana Tronik-Le Roux, Mathilde Sautreuil, Mahmoud Bentriou, Jérôme Vérine, Maria Belén Palma, Marina Daouya, Fatiha Bouhidel, Sarah Lemler, Joel LeMaoult, François Desgrandchamps, Paul-Henry Cournède, and Edgardo D. Carosella. Comprehensive landscape of immune-checkpoints uncovered in clear cell renal cell carcinoma reveals new and emerging therapeutic targets. *Cancer Immunology, Immunotherapy*, 2020.

[24] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

[25] Gil Stelzer, Naomi Rosen, Inbar Plaschkes, Shahar Zimmerman, Michal Twik, Simon Fishilevich, Tsippi Iny Stein, Ron Nudel, Iris Lieder, Yaron Mazor, Sergey Kaplan, Dvir Dahary, David Warshawsky, Yaron Guan-Golan, Asher Kohn, Noa Rappaport, Marilyn Safran, and Doron Lancet. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics*, 54(1):1.30.1–1.30.33, 2016.

[26] Zbyslaw Sondka, Sally Bamford, Charlotte G. Cole, Sari A. Ward, Ian Dunham, and Simon A. Forbes. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, 18(11):696–705, 2018.

[27] Gabriela Gremel, Dijana Djureinovic, Marjut Niinivirta, Alexander Laird, Oscar Ljungqvist, Henrik Johannesson, Julia Bergman, Per-Henrik Edqvist, Sanjay Navani, Naila Khan, Tushar Patil, Åsa Sivertsson, Mathias Uhlén, David J. Harrison, Gustav J. Ullenhag, Grant D. Stewart, and Fredrik Pontén. A systematic search strategy identifies cubilin as independent prognostic marker for renal cell carcinoma. *BMC cancer*, 17(1):9, 2017.

A Supplementary material

A.1 Méthodes de régularisation

A.1.1 Sur l'ensemble des gènes

	gènes sélectionnés			
Lasso (AIC = 1873.43 ±24.95)	ABCB5(1%)	ACRC(2%)	AMZ2(1%)	ANAPC7(6%)
	APOBEC2(2%)	ARHGEF4(2%)	**B3GNTL1**(87%)	BIN1(1%)
	BIRC6(7%)	C10orf90(25%)	C14orf165(1%)	**C19orf76**(80%)
	C1R(1%)	C20orf112(1%)	CACNA1S(2%)	CAMK2N2(1%)
	CAMSAP1(1%)	CCDC19(25%)	CCDC51(1%)	CDC7(1%)
	CDCA3(62%)	CENPBD1(3%)	**CHEK2**(81%)	**CKAP4**(87%)
	CNN2(1%)	CTAGE9(7%)	**CUBN**(87%)	CYB5D2(1%)
	DAG1(2%)	DDX24(3%)	DHRS12(8%)	EHBP1L1(3%)
	EIF4EBP2(59%)	EMILIN3(1%)	FAM133A(7%)	FAM63A(3%)
	FAM64A(1%)	FAM66A(7%)	FAM95B1(1%)	**FBXL5**(81%)
	FLI1(1%)	FUT3(7%)	GABBR1(1%)	GABPB2(1%)
	GDF5(74%)	GTPBP2(1%)	GZMA(1%)	**HBP1**(68%)
	HEG1(7%)	HPCAL1(5%)	HPCA(7%)	IGF2(7%)
	IKBIP(5%)	IKBKG(5%)	ILDR1(7%)	ITFG2(2%)
	ITPRIPL1(1%)	JAGN1(1%)	KIAA1109(1%)	KIAA1524(1%)
	KIF18B(13%)	KIF21B(7%)	KLHL14(7%)	LEO1(2%)
	LOC284233(8%)	LPIN1(2%)	LRRC23(4%)	LRRC8E(13%)
	LRRN4(7%)	MAST4(22%)	**MBOAT7**(72%)	MDM4(2%)
	MGST1(2%)	MIA(1%)	MKRN3(3%)	MOGAT2(7%)
	MRAP(2%)	NAA30(2%)	NDUFA8(7%)	NEK2(8%)
	NUMBL(1%)	OSTC(5%)	OTOF(81%)	P2RX7(2%)
	PABPC3(8%)	PDCD1(5%)	PDCD2(1%)	PHTF2(2%)
	PIGK(5%)	PIGO(5%)	PKD1L3(2%)	PRDM7(1%)
	PRKAA1(1%)	PROCA1(2%)	PROS1(2%)	PRUNE(8%)
	PTPLA(81%)	**RANGAP1**(69%)	RGS17(50%)	RGS20(50%)
	RP9P(6%)	RPL36AL(8%)	SCAP(1%)	SCO1(4%)
	SEC61A2(80%)	SERPINB8(1%)	SH2D4A(6%)	SHQ1(1%)
	SLC12A8(6%)	SLC45A4(4%)	SNCA(6%)	**SORBS2**(81%)
	SYVN1(4%)	TACC2(1%)	TAF9(1%)	TARP(1%)
	THEM4(1%)	THEMIS(4%)	TMEM17(4%)	TMEM203(1%)
	TMEM207(1%)	TMEM71(4%)	TTLL11(6%)	TTLL1(1%)
	TUBGCP5(2%)	TUSC1(4%)	UBE2D2(1%)	UMODL1(2%)
	VWA5B1(1%)	ZNF148(1%)	ZNF232(1%)	ZNF252(1%)
	ZNF626(4%)	ZNF676(7%)	ZNF766(1%)	ZNF90(2%)
Adaptive-Lasso (AIC = 1870.42 ±40.97)	ABCA8(1%)	ABCF3(1%)	ADCY10(1%)	ANAPC7(2%)
	ARHGEF4(1%)	ATP2A3(1%)	**B3GNTL1**(82%)	BIRC6(2%)
	BMP8B(2%)	BRSK1(1%)	C10orf12(1%)	C10orf90(21%)
	C12orf40(3%)	C19orf76(6%)	C21orf45(1%)	C2CD4C(2%)
	C4orf39(1%)	C5orf32(1%)	CAMSAP1(2%)	CASKIN1(1%)
	CBWD2(1%)	CCDC14(1%)	CCDC19(19%)	CCNG1(1%)
	CENPBD1(5%)	CHCHD10(1%)	**CKAP4**(70%)	CLNK(2%)
	COL6A2(1%)	**CUBN**(82%)	CUEDC1(2%)	DACH1(1%)
	DAG1(1%)	DBF4B(2%)	DDX24(1%)	DHRS12(1%)
	DQX1(1%)	DUSP2(1%)	EHBP1L1(1%)	EIF4EBP2(1%)
	ELFN2(1%)	ENGASE(1%)	ENTPD6(1%)	FAM133A(4%)
	FAM63A(1%)	FAM66A(3%)	FAM7A2(1%)	FAM95B1(1%)
	FBXL5(48%)	FBXO39(2%)	FCGR1C(1%)	GABPB2(1%)
	GAS2L3(1%)	**GDF5**(74%)	GFM2(1%)	GJB1(1%)
	GNAI2(1%)	GSTO1(2%)	GZMA(1%)	HBP1(27%)
	HDAC7(1%)	HDGF(1%)	HEG1(1%)	HIST1H2BK(2%)
	IKBKG(1%)	ILDR1(3%)	JAGN1(1%)	JPH3(1%)

	$KIF21B$(4%)	$KIF22$(2%)	$KLHL14$(1%)	$LAD1$(1%)
	$LAMB3$(2%)	$LCA5L$(1%)	$LILRB2$(1%)	$LOC284233$(2%)
	$LOC646471$(1%)	$LPIN1$(1%)	$LRRC8E$(4%)	$LRRN4$(2%)
	$LST1$(1%)	$MAST4$(1%)	$MDM4$(2%)	$MED20$(1%)
	$MFSD6$(1%)	$MIOX$(2%)	$MKI67IP$(1%)	MME(1%)
	$MOGAT2$(2%)	$MOXD1$(2%)	$MRGPRX3$(2%)	$NBPF3$(1%)
	$NCRNA00081$(1%)	$NDUFA8$(2%)	$NDUFS6$(2%)	$NFAM1$(1%)
	$NKRF$(2%)	$NXT1$(2%)	$OR9Q1$(1%)	$OSTC$(1%)
	OTOF(79%)	$OXCT1$(1%)	$PARP1$(1%)	$PKD1L3$(1%)
	$POLD2$(1%)	$POLI$(2%)	$PROS1$(2%)	$PROSC$(1%)
	$PRR12$(1%)	$PTGER3$(1%)	$PTPLA$(27%)	$RAB40AL$(5%)
	$RANGAP1$(2%)	$RGS20$(24%)	$RNF216L$(1%)	$RPL36AL$(3%)
	$SCO1$(3%)	$SEC13$(1%)	**SEC61A2**(26%)	$SERPINA7$(1%)
	$SERPINB8$(1%)	$SETD5$(2%)	$SFTA1P$(1%)	$SH2D4A$(1%)
	$SHQ1$(1%)	$SKA2$(1%)	$SLCO1A2$(1%)	$SLITRK3$(1%)
	$SMARCA2$(1%)	$SNCA$(1%)	$SNORA39$(1%)	**SORBS2**(78%)
	$SPIC$(2%)	$STX1A$(1%)	$SYNC$(1%)	$TAF1L$(1%)
	$TAF5L$(2%)	$TAF9$(1%)	$TBC1D15$(1%)	$TMCC1$(1%)
	$TMEM31$(1%)	$TMEM71$(1%)	$TNFSF12$(2%)	$TRAF7$(1%)
	$TUSC1$(3%)	$UAP1$(1%)	$UBE2R2$(1%)	$UMODL1$(1%)
	$USP42$(2%)	$WNT1$(4%)	$ZAP70$(1%)	$ZGPAT$(2%)
	$ZNF252$(1%)	$ZNF283$(2%)	$ZNF766$(2%)	$ZNF833$(1%)
	$ZNF90$(1%)			

Table 4: Résultats des méthodes de régularisation sur l'ensemble de gènes

A.2 Méthodes de *Screening*

A.2.1 Sur l'ensemble des gènes

	gènes sélectionnés			
SIS (AIC = 1873.80 ± 0.71)	**C5orf23**(100%)	**CAD**(83%)	**CHEK2**(100%)	**CUBN**(100%)
	$DHRS12$(27%)	**GDF5**(100%)	$SPC24$(1%)	$TOP2A$(69%)
ISIS (AIC = 1880.01 ±24.69)	$APOBEC2$(1%)	**C5orf23** (94%)	**CAD**(88%)	**CHEK2**(94%)
	$COMMD5$(1%)	$CTSO$(1%)	**CUBN**(94%)	$DHRS12$(45%)
	$DUOX1$(1%)	$EHBP1L1$(1%)	$FAM155B$(1%)	**GDF5** (93%)
	$GNA11$(2%)	$GRM7$(2%)	$IGF2$(2%)	$KIF21B$(2%)
	$LPIN1$(1%)	$NDUFA8$(2%)	$NKD1$(2%)	$NUP153$(2%)
	$PHTF2$(2%)	$PROS1$(2%)	$SNCA$(2%)	$SPAG5$(2%)
	$STK40$(2%)	$TMEM17$(2%)	$TUSC1$(2%)	$ZGPAT$(1%)
PSIS (AIC = 1931.38 ±12.55)	**ABTB2**(70%)	$ACER2$(2%)	$ACSBG1$(1%)	**ADAM17**(97%)
	ADCK1(85%)	$ADCY7$(1%)	**AFAP1L2**(96%)	$AGPAT5$(29%)
	$AHNAK$(2%)	$ALG12$(85%)	$ALKBH6$(10%)	$ALPK2$(22%)
	AMH(24%)	$AMIGO2$(29%)	$ANKH$(29%)	$ANKRD12$(13%)
	ANKRD22(77%)	**AP2A1**(99%)	$AP3B2$(15%)	$APH1B$(19%)
	APOL4(78%)	$ARHGEF6$(14%)	$ASB12$(16%)	$ASF1B$(29%)
	ASNA1(84%)	$ATG9A$(29%)	$ATP10B$(43%)	**ATP2C2**(97%)
	$ATP8B3$(8%)	AUH(43%)	**AXIN2**(96%)	$BAGE2$(19%)
	$BCL2L10$(1%)	$BCL9$(2%)	BCR(1%)	$BIN1$(2%)
	$BMP8A$(22%)	**BTN3A1**(78%)	$C10orf11$(22%)	**C11orf41**(82%)
	C12orf43(85%)	$C12orf52$(20%)	$C12orf62$(15%)	$C14orf129$(56%)
	$C14orf37$(2%)	$C14orf79$(19%)	$C16orf93$(8%)	$C17orf44$(56%)
	$C17orf46$(1%)	$C18orf8$(1%)	$C19orf36$(1%)	**C19orf55**(99%)
	C22orf26(84%)	**C2orf62**(97%)	$C3orf47$(76%)	**C3orf50**(95%)
	$C4orf44$(1%)	**C6orf163**(98%)	$C6orf59$(1%)	$C7orf55$(70%)
	$C7orf59$(1%)	$C8orf84$(15%)	$C9orf167$(13%)	**C9orf93**(99%)
	$CAPRIN2$(13%)	$CCDC102A$(1%)	$CCDC160$(8%)	$CCDC3$(70%)
	$CCL13$(29%)	$LILRA6$(1%)	$LMAN2$(1%)	$LOC100130691$(1%)
	$LOC284232$(1%)	$LOC441177$(1%)	$LOC613037$(1%)	$LOC81691$(1%)
	MAK(1%)	$MAPKAPK5$(1%)	$MCM3AP$(1%)	$MRPS10$(1%)
	$MRPS18C$(1%)	$MYT1$(1%)	$NDUFAF3$(1%)	$NEK2$(1%)
	$NFKBIE$(1%)	$NFKBIL1$(1%)	NNT(1%)	$NTN1$(1%)
	$RAB34$(1%)	$RAP1GAP2$(1%)	RFK(1%)	$RGS2$(1%)
	$RHEBL1$(1%)	$RNF121$(1%)	$RPL36$(1%)	$RSPO4$(1%)
	$SALL4$(1%)	$SCGN$(1%)	$SLA2$(1%)	$SLC41A1$(1%)
	$SNF8$(1%)	$SNRPA1$(1%)	$SOAT2$(1%)	$SOBP$(1%)
	$SPHK2$(1%)	$STAG3$(1%)		
coxCS (AIC = 1937.71 ±13.69)	**A1CF**(91%)	**A2BP1**(91%)	**A2ML1**(91%)	**A2M**(91%)
	A4GALT(91%)	**A4GNT**(91%)	**AAAS**(91%)	**AACSL**(89%)
	AACS(91%)	**AADAC**(89%)	**AADAT**(89%)	**AAGAB**(89%)
	AAK1(90%)	$AAMP$(73%)	**AARS2**(91%)	**AARSD1**(89%)
	$AASDH$(28%)	**AASS**(91%)	**AATF**(91%)	**AATK**(86%)
	ABCA10(91%)	**ABCA11P**(89%)	$ABCA12$(27%)	$ABCA13$(23%)
	$ABCA1$(50%)	**ABCA2**(91%)	**ABCA3**(89%)	**ABCA4**(91%)
	ABCA5(91%)	$CCDC109B$(1%)	$CCDC110$(1%)	$CCDC112$(1%)
	$CCDC113$(1%)	$CCDC114$(1%)	$CCDC115$(1%)	$CCDC116$(1%)
	$CCDC117$(1%)	$CCDC11$(1%)	$CCDC120$(1%)	$CCDC121$(1%)
	$CCDC122$(1%)	$CCDC123$(1%)	$CCDC125$(1%)	$CCDC126$(1%)
	$CCDC12$(1%)	$CCDC130$(1%)	$CCDC132$(1%)	$CCDC134$(1%)
	$CCDC135$(1%)	$CCDC137$(1%)	$CCDC138$(1%)	$CCDC13$(1%)
	$CCDC144A$(1%)	$CCDC144B$(1%)	$CCDC144C$(1%)	$CCDC146$(1%)
	$CCDC147$(1%)	$CYP4A22$(2%)	$CYP4B1$(2%)	$CYP4F12$(2%)
	$CYP4F22$(2%)	$CYP4F2$(2%)	$CYP4F3$(2%)	$CYP4V2$(2%)
	$CYP4X1$(2%)	$CYP4Z2P$(2%)	$CYP51A1$(2%)	$CYP7A1$(2%)
	$CYP7B1$(2%)	$CYP8B1$(1%)	$CYR61$(1%)	$CYS1$(2%)

	$CYSLTR1(2\%)$	$CYTH2(2\%)$	$CYTH3(2\%)$	$CYTH4(1\%)$
	$CYTIP(1\%)$	$CYTSA(1\%)$	$CYTSB(1\%)$	$CYYR1(1\%)$
	$D2HGDH(1\%)$	$D4S234E(1\%)$	$DAAM1(1\%)$	$DAAM2(1\%)$
	$DAB1(1\%)$	$HAVCR1(1\%)$	$HAVCR2(1\%)$	$HBA1(1\%)$
	$HBA2(1\%)$	$HBB(1\%)$	$HBD(1\%)$	$HBE1(1\%)$
	$HBEGF(1\%)$	$HBG1(1\%)$	$HBG2(1\%)$	$HBP1(1\%)$
	$HBS1L(1\%)$	$HBXIP(1\%)$	$HCFC1R1(1\%)$	$HCFC1(1\%)$
	$HCG18(1\%)$	$HCG22(1\%)$	$HCG26(1\%)$	$HCG27(1\%)$
	$HCG4P6(1\%)$	$HCG4(1\%)$	$HCG9(1\%)$	$HCN2(1\%)$
	$HCN3(1\%)$	$HCN4(1\%)$	$HCP5(1\%)$	$HCST(1\%)$
	$KNTC1(1\%)$	$KPNA1(1\%)$	$KPNA3(1\%)$	$KPNA4(1\%)$
	$KPNA5(1\%)$	$KPNA6(1\%)$	$KPNB1(1\%)$	$KPTN(1\%)$
	$KRAS(1\%)$	$KRBA1(1\%)$	$KRBA2(1\%)$	$KRCC1(1\%)$
	$KREMEN1(1\%)$	$KRI1(1\%)$	$KRIT1(1\%)$	$KRT13(1\%)$
	$KRT14(1\%)$	$KRT15(1\%)$	$KRT16(1\%)$	$KRT18(2\%)$
	$KRT19(2\%)$	$KRT1(2\%)$	$KRT23(2\%)$	$KRT24(2\%)$
	$KRT25(2\%)$	$KRT2(2\%)$	$KRT34(2\%)$	$PSMD8(2\%)$
	$PSMD9(2\%)$	$PSME2(2\%)$	$PSME3(2\%)$	$PSME4(2\%)$
	$PSMF1(1\%)$	$PSMG1(2\%)$	$PSMG2(2\%)$	$PSMG3(1\%)$
	$PSMG4(2\%)$	$PSORS1C1(2\%)$	$PSORS1C2(2\%)$	$PSORS1C3(2\%)$
	$PSPC1(2\%)$	$PSPH(1\%)$	$PSPN(1\%)$	$PSTPIP1(2\%)$
	$PSTPIP2(2\%)$	$PTAFR(2\%)$	$PTAR1(2\%)$	$PTBP2(2\%)$
	$PTCD1(2\%)$	$PTCD2(2\%)$	$PTCH2(2\%)$	$PTCHD1(2\%)$
	$PTCHD2(2\%)$	$PTCRA(2\%)$	$PTDSS1(2\%)$	$UBR5(2\%)$
	$UBR7(2\%)$	$UBTD2(2\%)$	$UBTFL1(2\%)$	$UBTF(2\%)$
	$UBXN10(2\%)$	$UBXN11(2\%)$	$UBXN1(2\%)$	$UBXN2A(2\%)$
	$UBXN2B(2\%)$	$UBXN4(2\%)$	$UBXN6(2\%)$	$UBXN7(2\%)$
	$UBXN8(2\%)$	$UCA1(2\%)$	$UCHL1(2\%)$	$UCK1(2\%)$
	$UCK2(2\%)$	$UCKL1AS(2\%)$	$UCKL1(2\%)$	$UCN3(2\%)$
	$UCN(2\%)$	$UCP2(2\%)$	$UCP3(2\%)$	$UFC1(2\%)$
	$UFD1L(2\%)$	$UFM1(2\%)$	$UFSP1(2\%)$	$UFSP2(2\%)$

Table 5: Résultats des méthodes de *Screening* sur l'ensemble de gènes

A.3 Getting biological knowledge for coxCS method

The CoxCS method uses biological knowledge to make a pre-selection. For this pre-selection, The proposed idea to use the latest publications referencing predictive biomarkers about patient survival. For this, we performed a PubMed search using the following keywords and MeSH terms (for *Medical Subject Headings*): `ccRCC [tiab] prognosis [tiab] survival [tiab] genes NOT RNA` MeSH is a hierarchically organized and controlled vocabulary produced by the *National Library of Medicine*. We thus obtained 32 results corresponding to our search, and we decided to keep the first 10 *abstracts*. These 10 *abstracts* were saved and later loaded into the web application `beca annotate` nunesbecas2013. The purpose of this web application is the identification and annotation of medical concepts in a text. The identification and annotation of genes and proteins performed by using *machine learning* methods present in Gimli camposgimli2013. Gimli camposgimli2013 is a *open-source* tool for automatic recognition of biomedical terms. The list of genes obtained is: $KDM2B$, $HMGCS2$, $HSD11B1$, $IL10$, $KDM5B$, $KDM5A$, $KDM5C$, $KDM5D$, $KDM1B$, $OGDHL$, $SSBP2$, $VSIG4$ and $XCR1$. We decided to use this list as biological knowledge for pre-selection in coxCS.