ESAIM: PS 18 (2014) 770–798
DOI: 10.1051/ps/2014006

ESAIM: Probability and Statistics
www.esaim-ps.org

VARIABLE SELECTION THROUGH CART *

Marie Sauve[1] and Christine Tuleau-Malot[2]

Abstract. This paper deals with variable selection in regression and binary classification frameworks. It proposes an automatic and exhaustive procedure which relies on the use of the CART algorithm and on model selection via penalization. This work, of theoretical nature, aims at determining adequate penalties, *i.e.* penalties which allow achievement of oracle type inequalities justifying the performance of the proposed procedure. Since the exhaustive procedure cannot be realized when the number of variables is too large, a more practical procedure is also proposed and still theoretically validated. A simulation study completes the theoretical results.

Résumé. Cet article aborde le thème de la sélection de variables dans le cadre de la régression et de la classification. Il propose une procédure automatique et exhaustive qui repose essentiellement sur l'utilisation de l'algorithme CART et sur la sélection de modèles par pénalisation. Ce travail, de nature théorique, tend à déterminer les bonnes pénalités, à savoir celles qui permettent l'obtention d'inégalité de type oracle. La procédure théorique n'étant pas implémentable lorsque le nombre de variables devient trop grand, une procédure pratique est également proposée. Cette seconde procédure demeure justifiée théoriquement. Par ailleurs, une étude par simulation complète le travail théorique.

Mathematics Subject Classification. 62G05, 62G07, 62G20.

Received December 4, 2012. Revised December 26, 2013.

1. Introduction

This paper discusses variable selection in non-linear regression and classification frameworks using CART estimation and a model selection approach. Our aim is to propose a theoretical variable selection procedure for non-linear models and to consider some practical approaches.

Variable selection is a very important subject since one must often consider situations in which the number of variables is very large while the number of variables that are really explanatory can be much smaller. That is why one must focus on their importance. The variable importance is a notion which allows the quantification of the ability of a variable to explain the phenomenon under study. The formula for the computation depends on the model considered. In the field's literature, there are many variable selection procedures that combine

Keywords and phrases. Binary classification, CART, model selection, penalization, regression, variable selection.

* *The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grant ATLAS (JCJC06_137446) "From Applications to Theory in Learning and Adaptive Statistics".*

[1] Lycée Jules Haag, 25000 Besançon, France. m.sauve@sfr.fr

[2] Laboratoire Jean-Alexandre Dieudonné, CNRS UMR 6621, Université de Nice Sophia-Antipolis Parc Valrose, 06108 Nice cedex 2, France. christine.malot@unice.fr

the concept of variable importance and model estimation. If ones refers to the work of Kohavi and John [19] or Guyon and Elisseff [15], these methods are "filter", "wrapper" or "embedded" methods. To summarize, (i) the filter method is a preprocessing step that does not depend on the learning algorithm, (ii) in the wrapper method the learning model is used to induce the final model but also to search for the optimal feature subset, and (iii) for embedded methods the feature selection and the learning part cannot be separated.

We shall mention some of these methods in the regression and/or the classification framework.

1.1. General framework and state of the art

Let us consider a linear regression model $Y = \sum_{j=1}^{p} \beta_j X^j + \varepsilon = X\beta + \varepsilon$, where ε is an unobservable noise, Y the response and $X = (X^1, \ldots, X^p)$ a vector of p explanatory variables. Let $\{(X_i, Y_i)_{1 \leq i \leq n}\}$ be a sample, *i.e.* n independent copies of the pair of random variables (X, Y).

The well-known Ordinary Least Squares (OLS) estimator provides a useful way to estimate the vector β but it suffers from a main drawback: it is not suitable for variable selection since when p is large, many components of β are non-zero: hence, there is a lack of sparsity. However, if OLS is not a convenient method for variable selection, the least squares criterion often appears in model selection. For example, Ridge Regression and Lasso (wrapper methods) are penalized versions of OLS. Ridge Regression (see Hastié *et al.* [17] for a survey) involves a L_2 penalization that produces the shrinkage of β but does require any coefficient of β to be zero. So if Ridge Regression is better than OLS, since there is nevertheless a thresholding of the coefficients of β, it is not a variable selection method, unlike Lasso. Lasso (see Tibshirani [31]) uses the least squares criterion penalized by a L_1 term. In this way, Lasso shrinks some coefficients of β and puts the others to zero. Thus, the latter method performs variable selection, but its implementation requires quadratic programming techniques.

Penalization is not the only way to perform variable or model selection. For example, we can cite the Subset Selection (see Hastié *et al.* [17]) which provides for each $k \in \{1, \ldots, p\}$ the best subset of size k, *i.e.* the subset of size k associated with the smallest residual sum of squares. Then, by cross-validation, the final subset is selected. This wrapper method is exhaustive: it is therefore difficult to carry out in practice when p is large. Often, Forward or Backward Stepwise Selection (see Hastié *et al.* [17]) is preferred since they are computationally efficient methods. However, since these methods are not exhaustive, they may eliminate useful predictors and thus not reach the globally optimal model. In the regression framework and when the number of explanatory variables is low, there is an efficient algorithm developed by Furnival and Wilson [9], which achieves the optimal model for a small number of explanatory variables without exploring all possible models.

Least Angle Regression (LARS) from the work of Efron *et al.* [7], is another usefull method. Let $\mu = x\beta$, where $x = (X_1^T, \ldots, X_n^T)$. LARS builds an estimator of μ by successive steps. It proceeds by adding, at each step, one covariate to the model, as Forward Selection. At first, $\mu = \mu_0 = 0$. In the first step, LARS finds the predictor X^{j_1} most correlated with the response Y and increases μ_0 in the direction of X^{j_1} until another predictor X^{j_2} has a larger correlation with the current residuals. Then, μ_0 is replaced by μ_1. This step corresponds to the first step of Forward Selection, but unlike Forward Selection, LARS is based on an equiangular strategy. For example, in the second step, LARS proceeds in an equiangular way between X^{j_1} and X^{j_2} until another explanatory variable enters into the model. This method is computationally efficient and produces good results in practice. However, a complete theoretical elucidation requires further investigation, even if additional survey exist on it (see for example Hesterberg *et al.* [18]).

There exist also recent studies for high and very high dimensional data (see Fan and Lv [8]) however such data are not really considered in our paper.

For linear regression, some work is also based on variable importance assessment; the aim is to produce a relative importance of regressor variables. Grömping [14] proposes a study of some estimators of the relative importance based on the variance decomposition.

In the context of non-linear models, Sobol [30] proposes an extension of the notion of relative importance of variables through Sobol sensitivity indices, *i.e.* those involved in the sensitivity analysis (*cf.* Saltelli *et al.* [27]). The concept of variable importance is not so new since it can be found in the book about Classification And

Regression Trees of Breiman *et al.* [5] which introduces the variable importance as the decrease of node impurity measures, or in studies about Random Forests by Breiman *et al.* [3,4], where the variable importance is rather a permutation importance index. With this notion, the variables can be ordered and we can easily deduce some filter or wrapper methods to select some of them. There are also some embedded purposes based on those notions or some others. Thus, Dìaz-Uriarte and Alvarez de Andrés [6] propose the following recursive strategy. They compute the Random Forests variable importance and they remove 20% of variables with the smallest importance: with the remaining variables, they construct a new forest and repeat the process. At the end, they compare all the models resulting from forests and they keep the one having the smallest Out Of Bag error rate. Poggi and Tuleau [25] develop a method using CART and on a stepwise ascending strategy including an elimination step, while Genuer *et al.* [10] propose a procedure combining Random Forest and elimination, ranking, and variable selection steps. Guyon *et al.* [16] propose a method of selection, called SVM-RFE, using Support Vector Machine methods based on Recursive Feature Elimination. Recently, this approach is a base of a new survey developed by Ben Ishak *et al.* [13], survey using a stepwise strategy.

1.2. Main goals

In this paper, the objective is to propose, for regression and classification frameworks, a variable selection procedure, based on CART, which is adaptive and theoretically validated. This second point is very important as it establishes a real difference with existing works: currently most practical methods for both frameworks are not validated because of the use of Random Forest or arbitrary thresholds of variable importance. Our method is to apply the CART algorithm to all possible subsets of variables and then to consider the model selection by penalization (*cf.* Birgé and Massart [2]), to select the set that minimizes a penalized criterion. In the regression and classification frameworks, we determine through oracle bounds the expressions of this penalized criterion. Of that feature, this work is in continuation of Birgé and Massart [2], Massart and Nédélec [24] and Gey and Nédélec [12]. The contribution of our work lies in the calibration of the penalty term which must take into account the complexity of the models through the number of variables involved.

More precisely, let $\mathcal{L} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be a sample, *i.e.* independent copies of a pair (X, Y), where X takes its values in $\mathcal{X}$, for example $\mathbb{R}^p$, with distribution μ and Y belongs to $\mathcal{Y}$ ($\mathcal{Y} = \mathbb{R}$ in the regression framework and $\mathcal{Y} = \{0; 1\}$ in the classification one). Let s be the regression function or the Bayes classifier according to the considered framework. We write $X = (X^1, \ldots, X^p)$ where the p variables X^j, with $j \in \{1, 2, \ldots, p\}$, are the explanatory variables. We denote by Λ the set of the p explanatory variables, *i.e.* $\Lambda = \{X^1, X^2, \ldots, X^p\}$, and by $\mathcal{P}(\Lambda)$ the set of all subsets of Λ. The explained variable Y is called the response. When one refers to variable selection, there are two distinct purposes (*cf.* Genuer *et al.* [10]): the first is to determine all the important variables highly correlated to the response Y, while the second is to find the smallest subset of variables to provide a good prediction of Y. Our goal here is to find a subset M of Λ, as small as possible, so that the variables of M are sufficient to predict the response Y.

To achieve this objective, we split the sample $\mathcal{L}$ in three sub-samples $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$ of size n_1, n_2 and n_3 respectively. In the following, we consider two cases: the first one is "$\mathcal{L}_1$ independent of $\mathcal{L}_2$" and the second corresponds to "$\mathcal{L}_1 = \mathcal{L}_2$". Then we apply the CART algorithm to all the subsets of Λ (an overview of CART is given later and for more details, the reader can refer to Breiman *et al.* [5]). More precisely, for any $M \in \mathcal{P}(\Lambda)$, we build the maximal tree by the CART growing procedure using the sub-sample $\mathcal{L}_1$. This tree, denoted $T_{\max}^{(M)}$, is constructed thanks to the class of admissible splits $\mathcal{S}p_M$ which involve only the variables of M. For any $M \in \mathcal{P}(\Lambda)$ and any subtree T of $T_{\max}^{(M)}$, denoted in the sequel $T \preceq T_{\max}^{(M)}$, we consider the space $S_{M,T}$ of $\mathbb{L}^2_{\mathcal{Y}}(\mathbb{R}^p, \mu)$ composed by all the piecewise constant functions with values in $\mathcal{Y}$ and defined on the partition $\tilde{T}$ associated with the leaves of T. At this stage, we have the collection of models

$$\left\{ S_{M,T}, \quad M \in \mathcal{P}(\Lambda) \text{ and } T \preceq T_{\max}^{(M)} \right\},$$

which depend only on $\mathcal{L}_1$. Then, for any (M,T) such that $M \in \mathcal{P}(\Lambda)$ and $T \preceq T_{\max}^{(M)}$, we denote $\hat{s}_{M,T}$ the $\mathcal{L}_2$ the empirical risk minimizer on $S_{M,T}$.

$$\hat{s}_{M,T} = \underset{u \in S_{M,T}}{\operatorname{argmin}}\, \gamma_{n_2}(u) \text{ with } \gamma_{n_2}(u) = \frac{1}{n_2} \sum_{(X_i,Y_i)\in\mathcal{L}_2} (Y_i - u(X_i))^2 .$$

Finally, we select $\widehat{(M,T)}$ by minimizing the penalized contrast:

$$\widehat{(M,T)} = \underset{(M,T),\ M\in\mathcal{P}(\Lambda) \text{ and } T\preceq T_{\max}^{(M)}}{\operatorname{argmin}} \{\gamma_{n_2}(\hat{s}_{M,T}) + pen(M,T)\}$$

and we denote the corresponding estimator $\tilde{s} = \hat{s}_{\widehat{M,T}}$.

Our purpose is to determine the penalty function *pen* so that the model $\widehat{(M,T)}$ is close to the optimal one. This means that the model selection procedure should satisfy an oracle inequality *i.e.*:

$$\mathbb{E}\left[l(s,\tilde{s})\ |\mathcal{L}_1\right] \le C \inf_{(M,T),\ M\in\mathcal{P}(\Lambda) \text{ and } T\preceq T_{\max}^{(M)}} \left\{\mathbb{E}\left[l(s,\hat{s}_{M,T})\ |\mathcal{L}_1\right]\right\}, \quad C \text{ close to } 1$$

where l denotes the loss function and s the optimal predictor. The main results of this paper give adequate penalties defined up to two multiplicative constants α and β.

Thus, we have a family of estimators $\tilde{s}(\alpha,\beta)$ from which the final estimator is selected by means of the test sample $\mathcal{L}_3$. This third sub-sample is introduced for theoretical reasons, since it provides theoretical results more easily.

We can therefore summarize this by the following definition:

Definition 1.1. Let define

$$\tilde{s} = \tilde{s}(\alpha,\beta) := \hat{s}_{\widehat{M,T}},$$

where

$$\widehat{(M,T)} = \underset{(M,T),\ M\in\mathcal{P}(\Lambda) \text{ and } T\preceq T_{\max}^{(M)}}{\operatorname{argmin}} \{\gamma_{n_2}(\hat{s}_{M,T}) + pen(M,T,\alpha,\beta)\}$$

the expression of the *pen* function is dependent on the study framework, namely $pen\,(M,T,\alpha,\beta) := pen_c\,(M,T,\alpha,\beta,h)$ for the classification framework and $pen\,(M,T,\alpha,\beta) = pen_r\,(M,T,\alpha,\beta,\rho,R)$ for the regression framework.

Let also define for some real α_0 and β_0

$$\tilde{\tilde{s}} = \underset{\tilde{s}(\alpha,\beta),\ \alpha>\alpha_0,\ \beta>\beta_0}{\operatorname{argmin}} \gamma_{n_3}(\tilde{s}(\alpha,\beta)).$$

The procedure described is of course a theoretical one, since when p is too large, it is impractical to consider all the 2^p sets of variables. One solution is, initially, to determine, using data, some subsets of variables that are appropriate and in a second time to apply our procedure. Since the restricted family of subsets, denoted $\mathcal{P}^*$, is included in the 2^p sets, the theoretical penalty remains valid (see the proofs and Sect. 5.1) even though it may over-penalize a little because the penalty is too large. However, obtaining an optimal calibration of an appropriate penalty would be difficult due to the randomness of the choice of $\mathcal{P}^*$. From this perspective, the theoretical penalty appears to be the best idea we have. Indeed, it seems clear that the construction of a suitable penalty will depend on the method of construction of $\mathcal{P}^*$. Therefore, the proposed variable selection procedure loses generality since the construction of $\mathcal{P}^*$ depends on the context of study but above data.

The paper is organized as follows: After this introduction, Section 2 outlines the different steps of the CART algorithm and introduces some notations. Respectively Sections 3 and 4 present the results obtained in the

classification and regression frameworks, in both sections, the results are in the same spirit: however, as the framework differs, the assumptions and the penalty functions are different. That is why, for the sake of clarity, we present our results in each context separately. In Section 5, we apply our procedure on simulated datasets, and compare our results with what was expected, firstly when we implement the theoretical procedure and secondly the simplified procedure involving a small number of subsets, subsets determined from the importance of variables defined by Breiman *et al.* [5]. In this section, we also look at the influence of correlated variables. Sections 6 and 7 collect lemmas and proofs.

2. Preliminaries

2.1. Overview of CART and variable selection

In the regression and classification frameworks and thanks to a training set, CART recursively divides the observations space $\mathcal{X}$ and defines a piecewise constant function on the partition induced, function called predictor or classifier as appropriate. CART proceeds in three stages: the construction of a maximal tree, the construction of nested models by pruning and a final step of model selection. In the following, we give a brief summary; for details, we invite the reader to refer to the seminal book of Breiman *et al.* [5] or to Gey's vulgarization articles [11,12].

The first step involves the construction of a nested sequence of partitions of $\mathcal{X}$ using binary splits. A useful representation of this construction is a tree composed of non-terminal and terminal nodes. At each non-terminal node is associated a binary split that is in the form of a question such as $(\mathbf{X^j} \leq \mathbf{c_j})$ for numerical variables and type $(\mathbf{X^j} \in \mathbf{S_j})$ for qualitative ones. Such a split involves only one explanatory variable and is determined by the maximization of a quality criterion induced by an impurity function. For example, in the regression framework the quality criterion associated with a node $\mathbf{t}$ is the decrease of $\mathbf{R(t)}$ where $\mathbf{R(t)} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{X_i} \in \mathbf{t}} (\mathbf{Y_i} - \bar{\mathbf{Y}}(\mathbf{t}))^{\mathbf{2}}$ with $\bar{\mathbf{Y}}(\mathbf{t})$ the arithmetical mean of $\mathbf{Y}$ over $\mathbf{t}$. This is simply the error estimate. In the classification framework, the quality criterion is the decrease of the impurity function which is often given by the Gini index $\mathbf{i(t)} = \sum_{\mathbf{i} \neq \mathbf{j}} \mathbf{p(i|t)p(j|t)}$ with $\mathbf{p(i|t)}$ the posterior probability of the class $\mathbf{i}$ in $\mathbf{t}$. In this case, the criterion is less intuitive but the estimate of the misclassification rate has too many drawbacks to be used like what has been done in regression. The tree associated with the finest partition, *i.e.* one that contains only one observation in each element of the partition or at least an observation of the same response, is called the maximal tree. This tree is too complex and too faithful to the training sample to be used as is. This is the reason for the next step called pruning.

The principle of pruning is, from a maximal tree, to extract a sequence of nested subtrees that minimize a penalized criterion proposed by Breiman *et al.* [5]. This penalized criterion realizes a trade-off between the goodness of fit and the complexity of the tree (or model) measured by the number of leaves (terminal nodes).

Finally, using a test sample or cross-validation, a subtree is selected in the previous collection.

CART is an algorithm which builds a binary decision tree. An initial idea for performing variable selection from a tree is to retain only the variables that appear in the binary splits defining the tree. This idea has many drawbacks, since on the one hand the number of selected variables may be too large, and on the other hand some very influential variables may not appear, as they were hidden by the selected ones.

A second idea is based on the Variable Importance (VI), a concept introduced by Breiman *et al.* [5]. This concept, calculated with respect to a given tree (typically coming from the procedure CART), quantifies the contribution of each variable by assigning a score between 0 and 100 (see [5] for more details). The variable selection consists of keeping only the variables whose rating is greater than an arbitrary threshold. But to-date, there is no procedure to determine automatically this threshold and also such a selection does not remove the variables that are highly dependent on relevant variables.

In this paper, we propose a new approach based on the application of CART to all subsets of variables and on the choice of the set that minimizes an adapted penalized criterion.

2.2. The context

The paper deals with two frameworks: regression and binary classification. In both cases, we denote

$$s = \underset{u:\mathbb{R}^p \to \mathcal{Y}}{\operatorname{argmin}} \mathbb{E}\left[\gamma(u,(X,Y))\right] \text{ with } \gamma(u,(x,y)) = (y-u(x))^2. \tag{2.1}$$

The quantity s represents the best predictor according to the quadratic contrast γ. Since the distribution P is unknown, s is unknown too. Thus, in the regression and classification frameworks, we use $(X_1,Y_1),\ldots,(X_n,Y_n)$, independent copies of (X,Y), to construct an estimator of s. The quality of this one is measured by the loss function l defined by:

$$l(s,u) = \mathbb{E}[\gamma(u,.)] - \mathbb{E}[\gamma(s,.)]. \tag{2.2}$$

In the regression case, the expression of s defined in (2.1) is

$$\forall x \in \mathbb{R}^p, \quad s(x) = \mathbb{E}[Y|X=x],$$

and the loss function l given by (2.2) is the $\mathbb{L}^2(\mathbb{R}^p,\mu)$-norm, denoted $\|.\|_\mu$.

In this context, each (X_i,Y_i) satisfies

$$Y_i = s(X_i) + \varepsilon_i$$

where $(\varepsilon_1,\ldots,\varepsilon_n)$ is a sample such that $\mathbb{E}\left[\varepsilon_i|X_i\right] = 0$. In the following, we assume that the variables ε_i have exponential moments around 0 conditionally to X_i. As explained in [28], this assumption can be expressed by the existence of two constants $\sigma \in \mathbb{R}_+^*$ and $\rho \in \mathbb{R}_+$ such that

$$\text{for any } \lambda \in (-1/\rho, 1/\rho)\,, \ \log \mathbb{E}\left[\mathrm{e}^{\lambda\varepsilon_i}\middle|X_i\right] \le \frac{\sigma^2\lambda^2}{2\left(1-\rho|\lambda|\right)}. \tag{2.3}$$

σ^2 is necessarily greater than $\mathbb{E}(\varepsilon_i^2)$ and can be chosen as close to $\mathbb{E}(\varepsilon_i^2)$ as desired, but at the price of a larger ρ.

Remark 2.1. If $\rho = 0$ in (2.3), the random variables ε_i are said to be sub-Gaussian conditionally to X_i.

In the classification case, the Bayes classifier s, given by (2.1), is defined by:

$$\forall x \in \mathbb{R}^p, \quad s(x) = \mathbb{1}_{\eta(x)\ge 1/2} \text{ with } \eta(x) = \mathbb{E}[Y|X=x].$$

As Y and the predictors u take their values in $\{0;1\}$, we have $\gamma(u,(x,y)) = \mathbb{1}_{u(x)\neq y}$ so we deduce that the loss function l can be expressed as:

$$l(s,u) = \mathbb{P}(Y \neq u(X)) - \mathbb{P}(Y \neq s(X)) = \mathbb{E}\left[|s(X)-u(X)||2\eta(X)-1|\right].$$

For both frameworks, we consider two situations:

- $(M1)$: the training sample $\mathcal{L}$ is divided in three independent parts $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$ of size n_1, n_2 and n_3, respectively. The sub-sample $\mathcal{L}_1$ is used to construct the maximal tree, $\mathcal{L}_2$ to prune it and $\mathcal{L}_3$ to perform the final selection;
- $(M2)$: the training sample $\mathcal{L}$ is divided only in two independent parts $\mathcal{L}_1$ and $\mathcal{L}_3$. The first one is both for the construction of the maximal tree and its pruning whereas the second one is for the final selection.

The $(M1)$ situation is theoretically easier since all the sub-samples are independent, thus each step of the CART algorithm is performed on independent data sets. With real data, it is often difficult to split the sample in three parts because of the small number of data. That is the reason why we also consider the more realistic situation $(M2)$.

3. Classification

This section deals with the binary classification framework. In this context, we know that the best predictor is the Bayes classifier s defined by:

$$\forall x \in \mathbb{R}^p, \quad s(x) = \mathbb{1}_{\eta(x)\geq 1/2}$$

A problem appears when $\eta(x)$ is close to $1/2$, because in this case, the choice between the label 0 and 1 is difficult. If $\mathbb{P}(\eta(x) = 1/2) \neq 0$, then the accuracy of the Bayes classifier is not really good and the comparison with s is not relevant. For this reason, we consider the margin condition introduced by Tsybakov [32]:

$$\exists h > 0, \text{ such that } \forall x \in \mathbb{R}^p, \quad |2\eta(x) - 1| \geq h. \tag{3.1}$$

For details about this margin condition, we refer to Massart [23]. Otherwise in [1] some considerations about margin-adaptive model selection can be found more precisely in the case of nested models and with the use of the margin condition introduced by Mammen and Tsybakov [21].

The following subsection gives results on the variable selection for the methods $(M1)$ and $(M2)$ under margin condition. More precisely, we define convenient penalty functions which lead to oracle bounds. The last subsection deals with the final selection by test sample $\mathcal{L}_3$.

3.1. Variable selection *via* $(M1)$ and $(M2)$

- $(M1)$ case:

Given the collection of models

$$\left\{ S_{M,T},\ M \in \mathcal{P}(\Lambda) \text{ and } T \preceq T_{\max}^{(M)} \right\}$$

built on $\mathcal{L}_1$, we use the second sub-sample $\mathcal{L}_2$ to select a model $\widehat{(M,T)}$ which is close to the optimal one. To do this, we minimize a penalized criterion

$$crit(M,T,\alpha,\beta) = \gamma_{n_2}\left(\hat{s}_{M,T}\right) + pen\left(M,T,\alpha,\beta\right)$$

The following proposition gives a penalty function *pen* for which the risk of the penalized estimator $\tilde{s} = \hat{s}_{\widehat{M,T}}$ can be compared to the oracle accuracy.

Proposition 3.1. *Let consider $\tilde{s}$ the estimator defined in Definition 1.1, let h the margin defined by 3.1 and let consider a penalty function of the form: $\forall\ M \in \mathcal{P}(\Lambda)$ and $\forall\ T \preceq T_{\max}^{(M)}$*

$$pen_c(M,T,\alpha,\beta,h) = \alpha\frac{|T|}{n_2 h} + \beta\frac{|M|}{n_2 h}\left(1 + \log\left(\frac{p}{|M|}\right)\right).$$

There exists two theoretical constants α_0 and β_0 such that if $\alpha > \alpha_0$ and $\beta > \beta_0$, then there exists two positive constants $C_1 > 1$ and C_2, which only depend on α and β, such that:

$$\mathbb{E}\left[l\left(s,\tilde{s}\right)|\mathcal{L}_1\right] \leq C_1 \inf_{(M,T),\ M\in\mathcal{P}(\Lambda) \text{ and } T\preceq T_{\max}^{(M)}} \left\{ l\left(s,S_{M,T}\right) + pen_c\left(M,T,\alpha,\beta,h\right) \right\} + C_2\frac{1}{n_2 h}$$

where $l(s,S_{M,T}) = \inf_{u\in S_{M,T}} l(s,u)$.

The penalty is the sum of two terms. The first one is proportional to $\frac{|T|}{n_2}$ and corresponds to the penalty proposed by Breiman *et al.* [5] in their pruning algorithm. The other one is proportional to $\frac{|M|}{n_2}\left(1 + \log\left(\frac{p}{|M|}\right)\right)$ and is due to the variable selection. It penalizes models that are based on too much explanatory variables. For a given value of $|M|$, this result validates the CART pruning algorithm in the binary classification framework, result proved also by Gey [11] in a more general situation since the author consider a less stronger margin condition.

Thanks to this penalty function, the problem can be divided in practice in two steps:

– First, for every set of variables M, we select a subtree $\hat{T}_M$ of $T_{\max}^{(M)}$ by

$$\hat{T}_M = \underset{T \preceq T_{\max}^{(M)}}{\operatorname{argmin}} \left\{ \gamma_{n_2}(\hat{s}_{M,T}) + \alpha' \frac{|T|}{n_2} \right\}.$$

This means that $\hat{T}_M$ is a tree obtained by the CART pruning procedure using the sub-sample $\mathcal{L}_2$

– Then we choose a set $\hat{M}$ by minimizing a criterion which penalizes the big sets of variables:

$$\hat{M} = \underset{M \in \mathcal{P}(\Lambda)}{\operatorname{argmin}} \left\{ \gamma_{n_2}(\hat{s}_{M,\hat{T}_M}) + pen_c(M, \hat{T}_M, \alpha, \beta, h) \right\}.$$

The $(M1)$ situation permits to work conditionally to the construction of the maximal trees $T_{\max}^{(M)}$ and to select a model among a deterministic collection. Finding a convenient penalty to select a model among a deterministic collection is easier, but we have not always enough observations to split the training sample $\mathcal{L}$ in three sub-samples. This is the reason why we study now the $(M2)$ situation.

• $(M2)$ case:

We extend our result to only one sub-sample $\mathcal{L}_1$. But, while in the $(M1)$ method we work with the expected loss, here we need the expected loss conditionally to $\{X_i,\ (X_i, Y_i) \in \mathcal{L}_1\}$ defined by:

$$l_1(s,u) = \mathbb{P}\left(u(X) \neq Y | \{X_i,\ (X_i, Y_i) \in \mathcal{L}_1\}\right) - \mathbb{P}\left(s(X) \neq Y | \{X_i,\ (X_i, Y_i) \in \mathcal{L}_1\}\right). \tag{3.2}$$

Proposition 3.2. *Let consider $\tilde{s}$ the estimator defined in Definition 1.1, let h the margin defined by 3.1 and let consider a penalty function of the form: $\forall\ M \in \mathcal{P}(\Lambda)$ and $\forall\ T \preceq T_{\max}^{(M)}$*

$$pen_c(M,T,\alpha,\beta,h) = \alpha \left[1 + (|M|+1)\left(1 + \log\left(\frac{n_1}{|M|+1}\right)\right)\right] \frac{|T|}{n_1 h} + \beta \frac{|M|}{n_1 h}\left(1 + \log\left(\frac{p}{|M|}\right)\right).$$

There exists two theoretical constants α_0 and β_0 such that if $\alpha > \alpha_0$ and $\beta > \beta_0$, then there exists three positive constants $C_1 > 2$, C_2, Σ which only depend on α and β, such that, with probability $\geq 1 - \mathrm{e}^{-\xi}\Sigma^2$:

$$l_1(s,\tilde{s}) \leq C_1 \inf_{(M,T),\ M \in \mathcal{P}(\Lambda)\ and\ T \preceq T_{\max}^{(M)}} \left\{ l_1\left(s, S_{M,T}\right) + pen_c(M,T,\alpha,\beta,h) \right\} + \frac{C_2}{n_1 h}(1+\xi)$$

where $l_1(s, S_{M,T}) = \inf_{u \in S_{M,T}} l_1(s,u)$.

When we consider the $(M2)$ situation instead of the $(M1)$ one, we only obtain an inequality with high probability instead of a result in expectation. Indeed, since all the results are obtained conditionally to the construction of the maximal tree, in this second situation, it is impossible to integrate with respect to $\mathcal{L}_1$ whereas in the first situation, we integrated with respect to $\mathcal{L}_2$.

Since the penalized criterion depends on two parameters α and β, we obtain a family of predictors $\tilde{s} = \hat{s}_{\widehat{M,T}}$ indexed by α and β, and the associated family of sets of variables $\hat{M}$. We now choose the final predictor using test sample and we deduce the corresponding set of selected variables.

3.2. Final selection

We have obtained a collection of predictors

$$\mathcal{G} = \{\tilde{s}(\alpha,\beta);\ \alpha > \alpha_0 \text{ and } \beta > \beta_0\}$$

which depends on $\mathcal{L}_1$ and $\mathcal{L}_2$.

For any M of $\mathcal{P}(\Lambda)$, the set $\left\{T \preceq T_{\max}^{(M)}\right\}$ is finite. As $\mathcal{P}(\Lambda)$ is finite too, the cardinal $\mathcal{K}$ of $\mathcal{G}$ is finite and

$$\mathcal{K} \leq \sum_{M \in \mathcal{P}(\Lambda)} \mathcal{K}_M$$

where $\mathcal{K}_M$ is the number of subtrees of $T_{\max}^{(M)}$ obtained by the pruning algorithm defined by Breiman *et al.* [5]. In practice, $\mathcal{K}_M$ is much smaller than $\left|\left\{T \preceq T_{\max}^{(M)}\right\}\right|$. Given the sub-sample $\mathcal{L}_3$, we choose the final estimator $\tilde{\tilde{s}}$ by minimizing the empirical contrast γ_{n_3} on $\mathcal{G}$.

$$\tilde{\tilde{s}} = \underset{\tilde{s}(\alpha,\beta)\in\mathcal{G}}{\operatorname{argmin}}\ \gamma_{n_3}\left(\tilde{s}(\alpha,\beta)\right)$$

The next result validates the final selection for the $(M1)$ method.

Proposition 3.3. *Let consider $\tilde{\tilde{s}}$ the final estimator defined in Definition 1.1 and let h the margin defined by 3.1. For any $\eta \in (0,1)$, we have:*

$$\mathbb{E}\left[l\left(s,\tilde{\tilde{s}}\right)\ |\mathcal{L}_1,\mathcal{L}_2\right] \leq \frac{1+\eta}{1-\eta}\inf_{(\alpha,\beta),\ \alpha>\alpha_0,\ \beta>\beta_0}\left\{l\left(s,\tilde{s}(\alpha,\beta)\right)\right\} + \frac{\left(\frac{1}{3}+\frac{1}{\eta}\right)\frac{1}{1-\eta}}{n_3 h}\log\mathcal{K} + \frac{\frac{2\eta+\frac{1}{3}+\frac{1}{\eta}}{1-\eta}}{n_3 h}.$$

where α_0 and β_0 are the two theoretical constants of Proposition 3.1 or Proposition 3.2.

For the $(M2)$ method, we get exactly the same result except that the loss l is replaced by the conditional loss l_1 (3.2).

For the $(M1)$ method, since the results in expectation of the Propositions 3.1 and 3.3 involve the same expected loss, we can compare the final estimator $\tilde{\tilde{s}}$ with the entire collection of models, where the constants may depend on α and β:

$$\mathbb{E}\left[l(s,\tilde{\tilde{s}})\ |\mathcal{L}_1,\mathcal{L}_2\right] \leq \tilde{C}_1 \inf_{(M,T),\ M\in\mathcal{P}(\Lambda)\text{ and }T\preceq T_{\max}^{(M)}}\left\{l(s,S_{M,T}) + pen_c(M,T,\alpha,\beta,h)\right\} + \frac{C_2}{n_2 h} + \frac{C_3}{n_3 h}\left(1+\log\mathcal{K}\right).$$

In the classification framework, it may be possible to obtain sharper upper bounds by considering for instance the version of Talagrand concentration inequality developed by Rio [26], or another margin condition as the one proposed by Koltchinskii (see [20]) and used by Gey [11]. However, the idea remains the same and those improvement do not have a real interest since we do not get in our work precise calibration of the constants.

4. Regression

Let us consider the regression framework where the ε_i are supposed to have exponential moments around 0 conditionally to X_i (*cf.* (2.3)).

In this section, we add a stop-splitting rule in the CART growing procedure. During the construction of the maximal trees $T_{\max}^{(M)}$, $M \in \mathcal{P}(\Lambda)$, a node is split only if the two resulting nodes contain, at least, $N_{\min}$ observations.

As in the classification section, the following subsection gives results on the variable selection for the methods $(M1)$ and $(M2)$ and the last subsection deals with the final selection by test sample $\mathcal{L}_3$.

4.1. Variable selection *via* $(M1)$ and $(M2)$

In this subsection, we show that for convenient constants α and β, the same form of penalty function as in classification framework leads to an oracle bound.

- $(M1)$ case:

Proposition 4.1. *Let suppose that $\|s\|_\infty \leq R$, with R a positive constant.*

Let $\tilde{s}$ the estimator defined in Definition 1.1 and let $\rho > 0$ satisfying condition 2.3. Let consider a penalty function of the form: $\forall\ M \in \mathcal{P}(\Lambda)$ and $\forall\ T \preceq T_{\max}^{(M)}$

$$pen_r(M,T,\alpha,\beta,\rho,R) = \alpha\left(\sigma^2+\rho R\right)\frac{|T|}{n_2} + \beta\left(\sigma^2+\rho R\right)\frac{|M|}{n_2}\left(1+\log\left(\frac{p}{|M|}\right)\right).$$

If $p \leq \log n_2$, $N_{\min} \geq 24\frac{\rho^2}{\sigma^2}\log n_2$, $\alpha > \alpha_0$ and $\beta > \beta_0$ where α_0 and β_0 are two theoretical constants, then there exists two positive constants $C_1 > 2$ and C_2, which only depend on α and β, such that:

$$\mathbb{E}\left[\|s-\tilde{s}\|_{n_2}^2 \ \middle|\mathcal{L}_1\right] \leq C_1 \inf_{(M,T),\ M\in\mathcal{P}(\Lambda)\ and\ T\preceq T_{\max}^{(M)}}\left\{\inf_{u\in S_{M,T}}\|s-u\|_\mu^2 + pen_r(M,T,\alpha,\beta,\rho,R)\right\} + C_2\frac{(\sigma^2+\rho R)}{n_2} + C(\rho,\sigma,R)\frac{1\!\!1_{\rho\neq 0}}{n_2\log n_2}$$

where $\|\ .\ \|_{n_2}$ denotes the empirical norm on $\{X_i;\ (X_i,Y_i)\in\mathcal{L}_2\}$ and $C(\rho,\sigma,R)$ is a constant which only depends on ρ, σ and R.

As in classification, the penalty function is the sum of two terms: one is proportional to $\frac{|T|}{n_2}$ and the other to $\frac{|M|}{n_2}\left(1+\log\left(\frac{p}{|M|}\right)\right)$. The first term corresponds also to the penalty proposed by Breiman *et al.* [5] in their pruning algorithm and validated by Gey and Nédélec [12] for the Gaussian regression case. This proposition validates the CART pruning penalty in a more general regression framework than the Gaussian one.

Remark 4.2. In practice, since σ^2, ρ and R are unknown, we consider penalties of the form

$$pen_r(M,T,\alpha,\beta) = \alpha'\frac{|T|}{n_2} + \beta'\frac{|M|}{n_2}\left(1+\log\left(\frac{p}{|M|}\right)\right)$$

If $\rho = 0$, the form of the penalty is

$$pen_r(M,T,\alpha,\beta,\rho,R) = \alpha\sigma^2\frac{|T|}{n_2} + \beta\sigma^2\frac{|M|}{n_2}\left(1+\log\left(\frac{p}{|M|}\right)\right),$$

the oracle bound becomes

$$\mathbb{E}\left[\|s-\tilde{s}\|_{n_2}^2 \ \middle|\mathcal{L}_1\right] \leq C_1 \inf_{(M,T),\ M\in\mathcal{P}(\Lambda)\text{ and }T\preceq T_{\max}^{(M)}}\left\{\inf_{u\in S_{M,T}}\|s-u\|_\mu^2 + pen_r(M,T,\alpha,\beta,\rho,R)\right\} + C_2\frac{\sigma^2}{n_2},$$

and the assumptions on $\|s\|_\infty$, p and $N_{\min}$ are no longer required. Moreover, the constants α_0 and β_0 can be taken as follows:

$$\alpha_0 = 2(1+3\log 2) \quad \text{and} \quad \beta_0 = 3.$$

In this case σ^2 is the single unknown parameter which appears in the penalty. Instead of using α' and β' as proposed above, we can in practice replace σ^2 by an estimator.

- $(M2)$ case:

In this situation, the same sub-sample $\mathcal{L}_1$ is used to build the collection of models

$$\left\{S_{M,T},\ M\in\mathcal{P}(\Lambda)\text{ and }T\preceq T_{\max}^{(M)}\right\}$$

and to select one of them.

For technical reasons, we introduce the collection of models

$$\{S_{M,T},\ M \in \mathcal{P}(\Lambda) \text{ and } T \in \mathcal{M}_{n_1,M}\}$$

where $\mathcal{M}_{n_1,M}$ is the set of trees built on the grid $\{X_i;\ (X_i,Y_i) \in \mathcal{L}_1\}$ with splits on the variables in M. This collection contains the preceding one and only depends on $\{X_i;\ (X_i,Y_i) \in \mathcal{L}_1\}$. We find nearly the same result as in the $(M1)$ situation.

Proposition 4.3. *Let suppose that $\|s\|_\infty \leq R$, with R a positive constant. Let $\tilde{s}$ the estimator defined in Definition 1.1 and $\rho > 0$ satisfying condition 2.3.*

Let consider a penalty function of the form: $\forall\ M \in \mathcal{P}(\Lambda)$ and $\forall\ T \preceq T_{\max}^{(M)}$

$$pen_r(M,T,\alpha,\beta,\rho,R) = \alpha\left(\sigma^2\left(1+\frac{\rho^4}{\sigma^4}\log^2\left(\frac{n_1}{p}\right)\right)+\rho R\right)\left(1+(|M|+1)\left(1+\log\left(\frac{n_1}{|M|+1}\right)\right)\right)\frac{|T|}{n_1}$$
$$+\beta\left(\sigma^2\left(1+\frac{\rho^4}{\sigma^4}\log^2\left(\frac{n_1}{p}\right)\right)+\rho R\right)\frac{|M|}{n_1}\left(1+\log\left(\frac{p}{|M|}\right)\right).$$

If $p \leq \log n_1$, $\alpha > \alpha_0$ and $\beta > \beta_0$ where α_0 and β_0 are two theoretical constants, then there exists three positive constants $C_1 > 2$, C_2 and Σ which only depend on α and β, such that:

$\forall \xi > 0$, with probability $\geq 1 - \mathrm{e}^{-\xi}\Sigma - \frac{c}{n_1 \log n_1}\mathbb{1}_{\rho\neq 0}$,

$$\|s-\tilde{s}\|_{n_1}^2 \leq C_1 \inf_{(M,T),\ M\in\mathcal{P}(\Lambda) \text{ and } T\preceq T_{\max}^{(M)}}\left\{\inf_{u\in S_{M,T}}\|s-u\|_{n_1}^2 + pen_r(M,T,\alpha,\beta,\rho,R)\right\}$$
$$+\frac{C_2}{n_1}\left(\left(1+\frac{\rho^4}{\sigma^4}\log^2\left(\frac{n_1}{p}\right)\right)\sigma^2+\rho R\right)\xi$$

where $\|\ .\ \|_{n_1}$ denotes the empirical norm on $\{X_i;\ (X_i,Y_i) \in \mathcal{L}_1\}$ and c is a constant which depends on ρ and σ.

Like in the $(M1)$ case, for a given $|M|$, we find a penalty proportional to $\frac{|T|}{n_1}$ as proposed by Breiman *et al.* and validated by Gey and Nédélec in the Gaussian regression framework. So here again, we validate the CART pruning penalty in a more general regression framework.

Unlike the $(M1)$ case, the multiplicative factor of $\frac{|T|}{n_1}$, in the penalty function, depends on M and n_1. Moreover, in the method $(M2)$, the inequality is obtained only with high probability.

Remark 4.4. If $\rho = 0$, the form of the penalty is

$$pen_r(M,T,\alpha,\beta,\rho,R) = \alpha\sigma^2\left[1+(|M|+1)\left(1+\log\left(\frac{n_1}{|M|+1}\right)\right)\right]\frac{|T|}{n_1} + \beta\sigma^2\frac{|M|}{n_1}\left(1+\log\left(\frac{p}{|M|}\right)\right),$$

the oracle bound is $\forall\ \xi > 0$, with probability $\geq 1 - \mathrm{e}^{-\xi}\Sigma$,

$$\|\tilde{s}-s\|_{n_1}^2 \leq C_1 \inf_{(M,T),\ M\in\mathcal{P}(\Lambda) \text{ and } T\preceq T_{\max}^{(M)}}\left\{\inf_{u\in S_{M,T}}\|s-u\|_{n_1}^2 + pen_r(M,T,\alpha,\beta,\rho,R)\right\} + C_2\frac{\sigma^2}{n_1}\xi$$

and the assumptions on $\|s\|_\infty$ and p are no longer required. Moreover, we see that we can take $\alpha_0 = \beta_0 = 3$.

4.2. Final selection

The next result validates the final selection.

Proposition 4.5. *Let $\tilde{\tilde{s}}$ the estimator defined in Definition 1.1 and with the same notations as the one used in Propositions 4.1 and 4.3, we have:*

- *In the $(M1)$ situation, taking $p \leq \log n_2$ and $N_{\min} \geq 4\frac{\sigma^2+\rho R}{R^2}\log n_2$, we have: for any $\xi > 0$, with probability $\geq 1 - \mathrm{e}^{-\xi} - \mathbb{1}_{\rho\neq 0}\frac{R^2}{2(\sigma^2+\rho R)}\frac{1}{n_2^{1-\log 2}}$, $\forall \eta \in (0,1)$,*

$$\left\|s - \tilde{\tilde{s}}\right\|_{n_3}^2 \leq \frac{(1+\eta^{-1}-\eta)}{\eta^2}\inf_{\tilde{s}(\alpha,\beta)\in\mathcal{G}}\|s-\tilde{s}(\alpha,\beta)\|_{n_3}^2 + \frac{1}{\eta^2}\left(\frac{2}{1-\eta}\sigma^2 + 8\rho R\right)\frac{(2\log\mathcal{K}+\xi)}{n_3}.$$

- *In the $(M2)$ situation, denoting $\epsilon(n_1) = 2\mathbb{1}_{\rho\neq 0}n_1\exp\left(-\frac{9\rho^2\log^2 n_1}{2(\sigma^2+3\rho^2\log n_1)}\right)$, we have: for any $\xi > 0$, with probability $\geq 1 - \mathrm{e}^{-\xi} - \epsilon(n_1)$, $\forall \eta \in (0,1)$,*

$$\left\|s - \tilde{\tilde{s}}\right\|_{n_3}^2 \leq \frac{(1+\eta^{-1}-\eta)}{\eta^2}\inf_{\tilde{s}(\alpha,\beta)\in\mathcal{G}}\|s-\tilde{s}(\alpha,\beta)\|_{n_3}^2 + \frac{1}{\eta^2}\left(\frac{2}{1-\eta}\sigma^2 + 4\rho R + 12\rho^2\log n_1\right)\frac{(2\log\mathcal{K}+\xi)}{n_3}.$$

Remark 4.6. If $\rho = 0$, by integrating with respect to ξ, we get for the two methods $(M1)$ and $(M2)$ that: for any $\eta \in (0,1)$,

$$\mathbb{E}\left[\left\|s - \tilde{\tilde{s}}\right\|_{n_3}^2 \;\middle|\mathcal{L}_1,\ \mathcal{L}_2\right] \leq \frac{1+\eta^{-1}-\eta}{\eta^2}\inf_{\tilde{s}(\alpha,\beta)\in\mathcal{G}}\left\{\mathbb{E}\left[\|s-\tilde{s}(\alpha,\beta)\|_{n_3}^2 \;\middle|\mathcal{L}_1,\ \mathcal{L}_2\right]\right\} + \frac{2}{\eta^2(1-\eta)}\frac{\sigma^2}{n_3}\left(2\log\mathcal{K}+1\right).$$

The conditional risk of the final estimator $\tilde{\tilde{s}}$ with respect to $\|\ \|_{n_3}$ is controlled by the minimum of the errors made by $\tilde{s}(\alpha,\beta)$. Thus the test sample selection does not alter so much the accuracy of the final estimator. Now we can conclude that theoretically our procedure is valid.

Unlike the classification framework, we are not able, even when $\rho = 0$, to compare the final estimator $\tilde{\tilde{s}}$ with the entire collection of models since the different inequalities involve empirical norms that can not be compared.

5. Simulations

The aim of this section is twice. On the one hand, we illustrate by an example the theoretical procedure, described in the Section 1. On the other hand, we compare the results of the theoretical procedure with those obtained when we consider the procedure restricted to a family $\mathcal{P}^*$ constructed thanks to Breiman's Variable Importance.

5.1. In practice

The procedure described above is an exhaustive selection that is fully demonstrated. However, completeness makes the procedure difficult to apply in practice when the number of variables p becomes too large. Indeed, as described, the procedure is very time consuming. For proof, one can view the approximate number of operations required for the procedure, or the computation time required by a computer.

TABLE 1. Computational time of the complete procedure according to the number of variables.

p	2	3	4	5	6	7	8	9	10	11	12
time (in seconds)	6	16	45	152	444	1342	4016	11 917	36 882	112 969	346 599

Regarding the number of operations, an approximation is given by:

$$\underbrace{\sum_{M\in\mathcal{P}(\Lambda)}}_{\text{cardinality } 2^p}\left(\underbrace{(2^{n-1}*(\underbrace{|\tilde{T}^{(M)}_{\max}|}_{\approx n}-1)}_{\text{construction of the maximal tree}}+\sum_{T\preceq T^{(M)}_{\max}}\left(\underbrace{\sum_{X_i\in\mathcal{L}_2}(l(X_i,T)+1)+\underbrace{c}_{\text{computation of the penality}}}_{\text{computation of the criterion}}\right)\right)$$

where $l(X_i,T)$ denotes the length of the path from the root to the leaf associated to X_i, with respect to the tree T.

From a practical point of view, the Table 1 shows the computation time required to the complete procedure, and this according to the number of variables p. It should be noted that the programs were executed on a cluster that is currently the most important resource of the laboratory.

According to the Table 1, it is clear that in practice it seems unreasonable to use the procedure in the state. This is the main reason why we propose to slightly modify the procedure by replacing the family $\mathcal{P}(\Lambda)$ by a family better suited. Indeed, if one refers to the proofs, we find that if a penalty function is valid for the family $\mathcal{P}(\Lambda)$, it is also the case for a subfamily. Therefore, as the results are conditional to $\mathcal{L}_1$, if $\mathcal{P}^*$ is built using $\mathcal{L}_1$ or prior informations as it is the case in our studies, the results remain the same! But, if on believes the approximate number of operations, it is reasonable to expect that to replace $\mathcal{P}(\Lambda)$ by a smaller family $\mathcal{P}^*$ will reduce computation time. Regarding the construction of $\mathcal{P}^*$, different strategies are possible. A response will be presented a little later.

5.2. Theoretical *versus* practical procedures

The simulated example, also used by Breiman *et al.* (see [5], p. 237), is composed of $p = 10$ explanatory variables $X^1,\dots,X^{10}$ such that:

$$\begin{cases} \mathbb{P}\left(X^1=-1\right)=\mathbb{P}(X^1=1)=\frac{1}{2} \\ \forall i\in\{2,\dots,10\},\quad \mathbb{P}(X^i=-1)=\mathbb{P}(X^i=0)=\mathbb{P}(X^i=1)=\frac{1}{3} \end{cases}$$

and of the explained variable Y given by:

$$Y=s\left(X^1,\dots,X^{10}\right)+\varepsilon=\begin{cases} 3+3X^2+2X^3+X^4+\varepsilon & \text{if} \quad X^1=1\ , \\ -3+3X^5+2X^6+X^7+\varepsilon & \text{if} \quad X^1=-1\ . \end{cases}$$

where the unobservable random variable ε is independent of $X^1,\dots,X^{10}$ and normally distributed with mean 0 and variance 2.

The variables X^8, X^9 and X^{10} do not appear in the definition of the explained variable Y, they can be considered as observable noise.

The Table 2 contains the Breiman's Variable Importance. The first row presents the explanatory variables ordered from the most influential to the less influential, whereas the second one contains the Breiman's Variable Importance Ranking.

We note that the Variable Importance Ranking is consistent with the simulated model since the two orders coincide. In fact, in the model, the variables X^3 and X^6 (respectively, X^4 and X^7) have the same effect on the response variable Y.

TABLE 2. Variable Importance Ranking for the considered simulated example.

Variable	X^1	X^2	X^5	X^3	X^6	X^4	X^7	X^8	X^9	X^{10}
Rank	1	2	3	5	4	7	6	8	9	10

TABLE 3. In this table appears the set associated with the estimator $\tilde{s}$ for some values of the parameters α and β which appear in the penalty function *pen*.

β \ α	$\alpha \leq 0.05$	$0.05 < \alpha \leq 0.1$	$0.1 < \alpha \leq 2$	$2 < \alpha \leq 12$	$12 < \alpha \leq 60$	$60 \leq \alpha$
$\beta \leq 100$	$\{1,2,5,6,3,7,4,8,9,10\}$	$\{1,2,5,6,3,7,4\}$	$\{1,2,5,6,3,7,4\}$	$\{1,2,5,6,3\}$	$\{1,2,5\}$	$\{1\}$
$100 < \beta \leq 700$	$\{1,2,5,6,3,7,4\}$	$\{1,2,5,6,3,7,4\}$	$\{1,2,5,6,3\}$	$\{1,2,5,6,3\}$	$\{1,2,5\}$	$\{1\}$
$700 < \beta \leq 1300$	$\{1,2,5,6,3\}$	$\{1,2,5,6,3\}$	$\{1,2,5,6,3\}$	$\{1,2,5,6,3\}$	$\{1,2,5\}$	$\{1\}$
$1300 < \beta \leq 1700$	$\{1,2,5\}$	$\{1,2,5\}$	$\{1,2,5\}$	$\{1,2,5\}$	$\{1\}$	$\{1\}$
$1900 < \beta$	$\{1\}$	$\{1\}$	$\{1\}$	$\{1\}$	$\{1\}$	$\{1\}$

To make in use our procedure, we consider a training sample $\mathcal{L} = \mathcal{L}_1 = \mathcal{L}_2$ which consists of the realization of 1000 independent copies of the pair of random variables (X, Y) where $X = (X^1, \ldots, X^{10})$.

The first results are related to the behaviour of the set of variables associated with the estimator $\tilde{s}$. More precisely, for given values of the parameters α and β of the penalty function, we look at the selected set of variables.

According to the model definition and the Variable Importance Ranking, the expected results are the following ones:

- the size of the selected set should belong to $\{1, 3, 5, 7, 10\}$. As the variables X^2 and X^5 (respectively X^3 and X^6, X^4 and X^7 or X^8, X^9 and X^{10}) have the same effect on the response variable, the other sizes could not appear, theoretically;
- the set of size k, $k \in \{1, 3, 5, 7, 10\}$, should contain the k most important variables since Variable Importance Ranking and model definition coincide;
- the final selected set should be $\{1, 2, 5, 3, 6, 4, 7\}$.

The behaviour of the set associated with the estimator $\tilde{s}$, when we apply the theoretical procedure, is summarized by the Table 3.

At the intersection of the row β and the column α appears the set of variables associated with $\tilde{s}(\alpha, \beta)$.

First, we notice that those results are the expected ones. Then, we see that for a fixed value of the parameter α (respectively β), the increasing of β (resp. α) results in the decreasing of the size of the selected set, as expected. Therefore, this decreasing is related to Breiman's Variable Importance since the explanatory variables disappear according to the Variable Importance Ranking (see Tab. 2).

This table summarizes the results obtained for one simulated data set. To see the finite sample performance, we consider seven others simulated data sets and we see that the table is quite the same for all simulated data sets. The bin borders are not exactly the same, but the orders of magnitude are similar (see Figs. 1 and 2). We were not able to consider more data sets because even if the number of variables is small, we have to consider for all couples (α, β), 1023 $(=2^{10} - 1)$ sets of variables, and the computations are really time consuming.

To perform the final step of the procedure, we consider a new data set $\mathcal{L}_3$, composed of 500 observations. The final selected set is the $\{1, 2, 5, 3, 6, 4, 7\}$ (the expected one) or $\{1, 2, 5, 3, 6\}$, in quite the same proportion. One explanation of this phenomenon is the choice of the grid for α ans β. Indeed, if the grid is finer than the

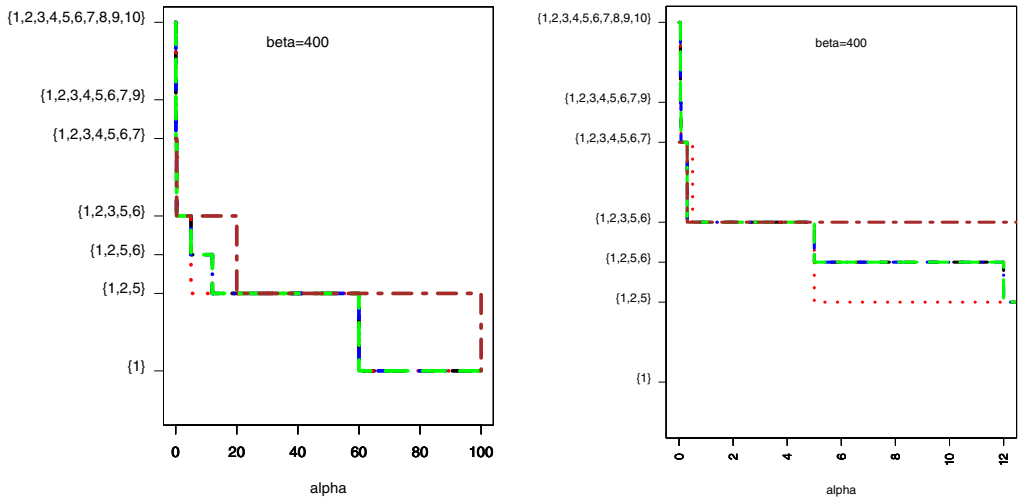

FIGURE 1. Behaviour of the selected set according different values of α for $\beta = 400$. The graph on the right side is just a zoom of the other graph. We see that for the different simulations, the behaviour are quite the same.

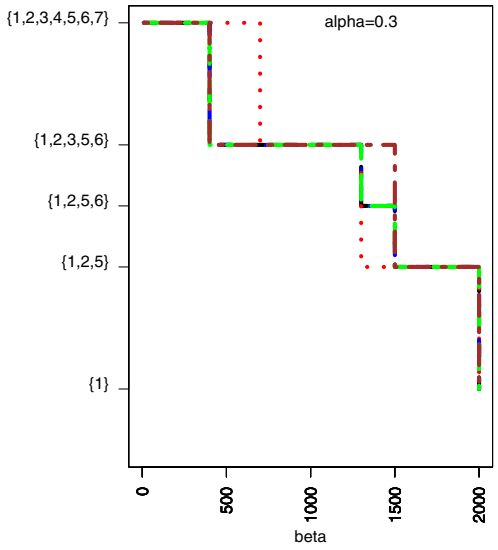

FIGURE 2. Behaviour of the selected set according different values of β for $\alpha = 0.3$. As previously, we get the same results for the different simulations.

one chosen, the procedure does not give the same result. A solution to this should be the calibration of α and β. The results of the procedure show that α and β vary in small range but are data dependent. Thus the results lead to the conclusion that a data-driven determination of the parameters α and β of the penalty function may be possible and that further investigations are needed, since with this determination the test sample in the final step could disappear.

Another explanation is due to the model. Indeed, if we consider the linear regression model $Y = X\theta + \varepsilon$ on one hand on the observations associated to $(X^1 = 1)$ and on the other hand on the observations associated to

($X^1 = -1$), the estimations of θ have a smaller coefficient than the expected one for X^4 (and respectively X^7) since it is quite zero. Thus, the solution of our procedure seems to be correct.

As the theoretical procedure is validated on the simulated example, we consider now a more realistic procedure when the number of explanatory variables is large. It involves a smaller family $\mathcal{P}^*$ of sets of variables. To determine this family, we use an idea introduced by Poggi and Tuleau in [25] which associates Forward Selection and variable importance (VI) and whose principle is the following one. The sets of $\mathcal{P}^*$ are constructed by invoking and testing the explanatory variables according to Breiman's Variable Importance ranking. More precisely, the first set is composed of the most important variable according to VI. To construct the second one, we consider the two most important variables and we test if the addition of the second most important variable has a significant incremental influence on the response variable. If the influence is significant, the second set of $\mathcal{P}^*$ is composed of the two most importance variables. If not, we drop the second most important variable and we consider the first and the third most important variables and so on. So, at each step, we add an explanatory variable to the preceding set which is less important than the preceding ones.

For the simulated example, the corresponding family $\mathcal{P}^*$ is:

$$\mathcal{P}^* = \{\{1\}; \{1,2\}; \{1,2,5\}; \{1,2,5,6\}; \{1,2,5,6,3\}; \{1,2,5,6,3,7\}; \{1,2,5,6,3,7,4\}\}$$

In this family, the variables X^8, X^9 and X^{10} do not appear. This is consistent with the model definition and Breiman's VI ranking.

The first advantage of this family $\mathcal{P}^*$ is that it involves, at the most p sets of variables instead of 2^p. A consequence is the possibility to perform the procedure with more simulated data sets than the theoretical procedure. Indeed, in this part, we consider $K = 100$ data sets. The second advantage is that, if we perform our procedure restricted to the family $\mathcal{P}^*$, we obtain nearly the same results for the behaviour of the set associated with $\tilde{s}$ than the one obtained with all the $2^p - 1$ sets of variables (see Tab. 3). The only difference is that, since $\mathcal{P}^*$ does not contain the set of size 10, in the Table 3, the set $\{1,2,3,4,5,6,7,8,9,10\}$ is replaced by $\{1,2,5,6,3,7,4\}$. And, the conclusions are the same for the set associated to $\tilde{\tilde{s}}$. Thus in practice, to avoid a very time consuming procedure, it is better to consider this practical procedure than the theoretical one,which has almost the same performance.

5.3. Influence of correlated variables

The previous example shows the pertinence of our theoretical procedure and of our practical on an simple model. One can also wonder what happens when some correlated variables are added. To answer this problem, we consider the following simulated example composed of $p = 10$ explanatory variables $X^1, \ldots, X^{10}$ such that:

$$\begin{cases} \mathbb{P}(X^1 = -1) = \mathbb{P}(X^1 = 1) = \frac{1}{2} \\ \forall i \in \{2, \ldots, 8\}, \quad \mathbb{P}(X^i = -1) = \mathbb{P}(X^i = 0) = \mathbb{P}(X^i = 1) = \frac{1}{3} \\ X^9 \text{ and } X^{10} \text{ are two variables correlated to } X^2 \text{ with a correlation equal to } \rho \end{cases}$$

and of the explained variable Y given by:

$$Y = s\left(X^1, \ldots, X^{10}\right) + \varepsilon = \begin{cases} 3 + 3X^2 + 2X^3 + \varepsilon & \text{if} \quad X^1 = 1\ , \\ -3 + 3X^4 + 2X^5 + \varepsilon & \text{if} \quad X^1 = -1\ . \end{cases}$$

where the unobservable random variable ε is independent of $X^1, \ldots, X^{10}$ and normally distributed with mean 0 and variance 2.

The second model is a simplified version of the previous one just justified by computational costs.

To see the influence of correlated variables, for the theoretical procedure, we compare the results obtained when we consider the data set without the correlated variables with the results obtained with the full data set when $\rho = 0.9$, $\rho = 0.5$ and $\rho = 0.2$ and we perform 5 repetitions of each case.

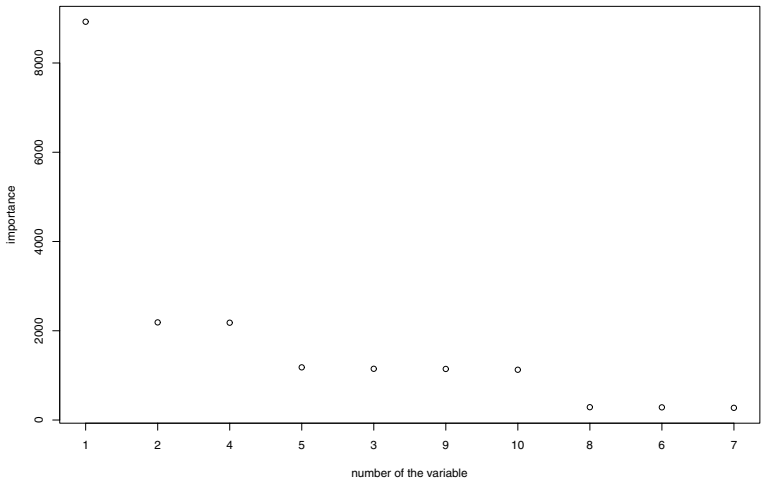

FIGURE 3. Variable importance of the 10 variables.

For all the simulations, the selected final set of variables is always $\{1, 2, 3, 4, 5\}$ which is the expected one. Thus, it seems that the introduction of correlated variables does not deteriorate the performance of our theoretical procedure.

The low number of repetitions is only due to considerations of feasibility (computing time). Thus, under this small number of repetitions, it is hard to really conclude about the robustness of the procedure with respect to the correlations, although the results seem to go in the right direction. Indeed, if one looks at the penalized criterion involved in the variable selection, there is a bias term and a term of variance. The bias term will not be greatly disturbed by the addition of correlated variables, in the sense that the minimum value will not decrease drastically. By cons, regarding the term of variance, which is a function of the number of variables ($|M|$), it may vary in proportions far greater.

To support all this into practice, we consider a family $\mathcal{P}^{**}$ intermediary between $\mathcal{P}(\Lambda)$ and $\mathcal{P}^{*}$. $\mathcal{P}^{**}$ is defined by:

$$\mathcal{P}^{**} = \left\{ \bigcup_{j\in\{1,\dots,4\}} c_j \;\; c_1 = \{1\},\; c_2 = \mathcal{P}(\{2,4\}),\; c_3 = \mathcal{P}(\{3,5,9,10\}),\; c_4 = \mathcal{P}(\{6,7,8\}) \right\}$$

This family allows to include correlated variables while taking into consideration the variable importance. Indeed, as shown in Figure 3, there are four groups of variables clearly identified by their importance, that are $\{X^1\}$, $\{X^2, X^4\}$, $\{X^3, X^5, X^9, X^{10}\}$ and $\{X^6, X^7, X^8\}$. For practical reasons, the variable importance is the one associated to a Random Forest (see Breiman [3]), random forest composed of one tree and whose construction involved the p explanatory variables for each split.

By considering the family $\mathcal{P}^{**}$, we reduce heavily the computational cost since instead of considering 2^{10} subsets of variables, we consider only 364 subsets. This allows to make more repetitions.

To have a better idea of the robustness of the proposed applied procedure, we repeat 100 times, for $\rho = 0.2$, $\rho = 0.5$ and $\rho = 0.9$, our applied procedure involving the family $\mathcal{P}^{**}$. In all the cases, the final subset is $\{X^1, X^2, X^3, X^4, X^5\}$, *i.e.* the expected set with respect to the model.

6. Appendix

This section presents some lemmas which are useful in the proofs of the propositions of Sections 4 and 3. The Lemmas 6.1 to 6.4 are known results. We just give the statements and references for the proofs. The Lemma 6.5

is a variation of Lemma 6.4. The remaining lemmas are intermediate results which we prove to obtain both the propositions and their proofs.

The Lemma 6.1 is a concentration inequality due to Talagrand. This type of inequality allows to know how a random variable behaves around its expectation.

Lemma 6.1 (Talagrand)**.**
Consider n independent random variables $\xi_1, \ldots, \xi_n$ with values in some measurable space Θ. Let $\mathcal{F}$ be some countable family of real valued measurable functions on Θ, such that $\|f\|_\infty \leq b < \infty$ for every $f \in \mathcal{F}$.

Let $Z = \sup_{f\in\mathcal{F}} |\sum_{i=1}^n (f(\xi_i) - \mathbb{E}[f(\xi_i)])|$ and $\sigma^2 = \sup_{f\in\mathcal{F}} \{\sum_{i=1}^n \mathrm{Var}[f(\xi_i)]\}$

Then, there exists K_1 and K_2 two universal constants such that for any positive real number x,

$$\mathbb{P}\left(Z \geq K_1\mathbb{E}[Z] + K_2\left(\sigma\sqrt{2x} + bx\right)\right) \leq \exp(-x).$$

Proof. (See Massart [22]). □

The Lemma 6.2 allows to pass from local maximal inequalities to a global one.

Lemma 6.2 (Maximal inequality)**.**
Let $(\mathcal{S}, d)$ be some countable set.

Let Z be some process indexed by $\mathcal{S}$ such that $\sup_{t\in B(u,\sigma)} |Z(t) - Z(u)|$ has finite expectation for any positive real σ, with $B(u,\sigma) = \left\{t \in \mathcal{S} \text{ such that } d(t,u) \leq \sigma\right\}$.

Then, for all $\Phi : \mathbb{R} \to \mathbb{R}^+$ such that:

$$- \quad x \to \frac{\Phi(x)}{x} \text{ is non increasing,}$$

$$- \quad \forall \sigma \geq \sigma_* \quad \mathbb{E}\left[\sup_{t\in B(u,\sigma)} |Z(t) - Z(u)|\right] \leq \Phi(\sigma),$$

we have:

$$\forall x \geq \sigma_* \quad \mathbb{E}\left[\sup_{t\in\mathcal{S}} \frac{|Z(t) - Z(u)|}{d^2(t,u) + x^2}\right] \leq \frac{4}{x^2}\Phi(x).$$

Proof. (See Massart and Nédélec [24], Sect. "Appendix: Maximal inequalities", Lem. 5.5). □

Thanks to the Lemma 6.3, we see that the Hold-Out is an adaptive selection procedure for classification.

Lemma 6.3 (Hold-Out)**.**
Assume that we observe $N + n$ independent random variables with common distribution P depending on some parameter s to be estimated. The first N observations $X' = (X'_1, \ldots, X'_N)$ are used to build some preliminary collection of estimators $(\hat{s}_m)_{m\in\mathcal{M}}$ and we use the remaining observations $(X_1, \ldots, X_n)$ to select some estimator $\hat{s}_{\hat{m}}$ among the collection defined before by minimizing the empirical contrast.

Suppose that $\mathcal{M}$ is finite with cardinal K.

If there exists a function w such that:

$$- \quad w : \mathbb{R}^+ \to \mathbb{R}^+,$$

$$- \quad x \to \frac{w(x)}{x} \text{ is non increasing,}$$

$$- \quad \forall \epsilon > 0, \quad \sup_{l(s,t)\leq\epsilon^2} Var_P\left(\gamma(t,.) - \gamma(s,.)\right) \leq w^2(\epsilon)$$

Then, for all $\theta \in (0,1)$*, one has:*

$$(1-\theta)\,\mathbb{E}\left[l(s,\hat{s}_{\hat{m}})|X'\right] \leq (1+\theta)\inf_{m\in\mathcal{M}} l(s,\hat{s}_m) + \delta_*^2\left(2\theta + (1+\log K)\left(\frac{1}{3}+\frac{1}{\theta}\right)\right)$$

where δ_*^2 *satisfies to* $\sqrt{n}\delta_*^2 = w(\delta_*)$.

Proof. (See [23], Chap. "Statistical Learning", Sect. "Advanced model selection problems"). □

The Lemmas 6.4 and 6.5 are concentration inequalities for a sum of squared random variables whose Laplace transform are controlled. The Lemma 6.4 is due to Sauvé [28] and allows to generalize the model selection result of Birgé and Massart [2] for histogram models without assuming the observations to be Gaussian. In the first lemma, we consider only partitions m of $\{1,\dots,n\}$ constructed from an initial partition m_0 (*i.e.* for any element J of m, J is the union of elements of m_0), whereas in the second lemma we consider all partitions m of $\{1,\dots,n\}$.

Lemma 6.4. *Let* $\varepsilon_1,\dots,\varepsilon_n$ *n independent and identically distributed random variables satisfying:*

$$\mathbb{E}[\varepsilon_i] = 0 \quad \textit{and} \quad \textit{for any } \lambda \in (-1/\rho, 1/\rho)\,,\ \log\mathbb{E}\left[\mathrm{e}^{\lambda\varepsilon_i}\right] \leq \frac{\sigma^2\lambda^2}{2\,(1-\rho|\lambda|)}$$

Let m_0 *a partition of* $\{1,\dots,n\}$ *such that,* $\forall J \in m_0$, $|J| \geq N_{\min}$.

We consider the collection $\mathcal{M}$ *of all partitions of* $\{1,\dots,n\}$ *constructed from* m_0 *and the statistics*

$$\chi_m^2 = \sum_{J\in m} \frac{\left(\sum_{i\in J}\varepsilon_i\right)^2}{|J|},\ m\in\mathcal{M}$$

Let $\delta > 0$ *and denote* $\Omega_\delta = \left\{\forall J\in m_0;\ \left|\sum_{i\in J}\varepsilon_i\right| \leq \delta\sigma^2|J|\right\}$.
Then for any $m\in\mathcal{M}$ *and any* $x>0$,

$$\mathbb{P}\left(\chi_m^2 \mathbb{1}_{\Omega_\delta} \geq \sigma^2|m| + 4\sigma^2(1+\rho\delta)\sqrt{2|m|x} + 2\sigma^2(1+\rho\delta)x\right) \leq \mathrm{e}^{-x}$$

and

$$\mathbb{P}\left(\Omega_\delta^c\right) \leq 2\frac{n}{N_{\min}}\exp\left(\frac{-\delta^2\sigma^2 N_{\min}}{2(1+\rho\delta)}\right).$$

Proof. (See [28], Lem. 1). □

Lemma 6.5. *Let* $\varepsilon_1,\dots,\varepsilon_n$ *n independent and identically distributed random variables satisfying:*

$$\mathbb{E}[\varepsilon_i] = 0 \quad \textit{and} \quad \textit{for any } \lambda \in (-1/\rho, 1/\rho)\,,\ \log\mathbb{E}\left[\mathrm{e}^{\lambda\varepsilon_i}\right] \leq \frac{\sigma^2\lambda^2}{2\,(1-\rho|\lambda|)}$$

We consider the collection $\mathcal{M}$ *of all partitions of* $\{1,\dots,n\}$ *and the statistics*

$$\chi_m^2 = \sum_{J\in m} \frac{\left(\sum_{i\in J}\varepsilon_i\right)^2}{|J|},\ m\in\mathcal{M}$$

Let $\delta > 0$ *and denote* $\Omega_\delta = \left\{\forall 1\leq i\leq n;\ |\varepsilon_i| \leq \delta\sigma^2\right\}$.

Then for any $m\in\mathcal{M}$ *and any* $x>0$,

$$\mathbb{P}\left(\chi_m^2 \mathbb{1}_{\Omega_\delta} \geq \sigma^2|m| + 4\sigma^2(1+\rho\delta)\sqrt{2|m|x} + 2\sigma^2(1+\rho\delta)x\right) \leq \mathrm{e}^{-x}$$

and

$$\mathbb{P}\left(\Omega_\delta^c\right) \leq 2n\exp\left(\frac{-\delta^2\sigma^2}{2(1+\rho\delta)}\right).$$

Proof. The proof is exactly the same as the preceding one. The only difference is that the set Ω_δ is smaller and $N_{\min} = 1$. □

The Lemmas 6.6 and 6.7 give the expression of the weights needed in the model selection procedure.

Lemma 6.6. *The weights* $x_{M,T} = a|T| + b|M|\left(1 + \log\left(\frac{p}{|M|}\right)\right)$, *with* $a > 2\log 2$ *and* $b > 1$ *two absolute constants, satisfy*

$$\sum_{M\in\mathcal{P}(\Lambda)} \sum_{T\preceq T_{\max}^{(M)}} \mathrm{e}^{-x_{M,T}} \leq \Sigma(a,b) \tag{6.1}$$

with $\Sigma(a,b) = -\log\left(1-\mathrm{e}^{-(a-2\log 2)}\right)\frac{\mathrm{e}^{-(b-1)}}{1-\mathrm{e}^{-(b-1)}} \in \mathbb{R}_+^*$.

Proof. We are looking for weights $x_{M,T}$ such that the sum

$$\Sigma(\mathcal{L}_1) = \sum_{M\in\mathcal{P}(\Lambda)} \sum_{T\preceq T_{\max}^{(M)}} \mathrm{e}^{-x_{M,T}}$$

is lower than an absolute constant.

Taking x as a function of the number of variables $|M|$ and of the number of leaves $|T|$, we have

$$\Sigma(\mathcal{L}_1) = \sum_{k=1}^{p} \sum_{\substack{M\in\mathcal{P}(\Lambda)\\|M|=k}} \sum_{D=1}^{n_1} \left|\left\{T\preceq T_{\max}^{(M)};\ |T| = D\right\}\right| \mathrm{e}^{-x(k,D)}.$$

Since

$$\left|\left\{T\preceq T_{\max}^{(M)};\ |T| = D\right\}\right| \leq \frac{1}{D}\binom{2(D-1)}{D-1} \leq \frac{2^{2D}}{D},$$

we get

$$\Sigma(\mathcal{L}_1) \leq \sum_{k=1}^{p}\left(\frac{ep}{k}\right)^k \sum_{D\geq 1}\frac{1}{D}\mathrm{e}^{-(x(k,D)-(2\log 2)D)}.$$

Taking $x(k,D) = aD + bk\left(1+\log\left(\frac{p}{k}\right)\right)$ with $a > 2\log 2$ and $b > 1$ two absolute constants, we have

$$\Sigma(\mathcal{L}_1) \leq \left(\sum_{k\geq 1}\mathrm{e}^{-(b-1)k}\right)\left(\sum_{D\geq 1}\frac{1}{D}\mathrm{e}^{-(a-(2\log 2))D}\right) = \Sigma(a,b).$$

Thus the weights $x_{M,T} = a|T| + b|M|\left(1+\log\left(\frac{p}{|M|}\right)\right)$, with $a > 2\log 2$ and $b > 1$ two absolute constants, satisfy (6.1). □

Lemma 6.7. *The weights*

$$x_{M,T} = \left(a + (|M|+1)\left(1+\log\left(\frac{n_1}{|M|+1}\right)\right)\right)|T| + b\left(1+\log\left(\frac{p}{|M|}\right)\right)|M|$$

with $a > 0$ *and* $b > 1$ *two absolute constants, satisfy*

$$\sum_{M\in\mathcal{P}(\Lambda)} \sum_{T\in\mathcal{M}_{n_1,M}} \mathrm{e}^{-x_{M,T}} \leq \Sigma'(a,b) \tag{6.2}$$

with $\Sigma'(a,b) = \frac{\mathrm{e}^{-a}}{1-\mathrm{e}^{-a}}\frac{\mathrm{e}^{-(b-1)}}{1-\mathrm{e}^{-(b-1)}}$ *and* $\mathcal{M}_{n_1,M}$ *the set of trees built on the grid* $\{X_i;\ (X_i,Y_i)\in\mathcal{L}_1\}$ *with splits on the variables in* M.

Proof. The proof is quite the same as the preceding one. □

The two last lemmas provide controls in expectation for processes studied in classification.

Lemma 6.8. *Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be n independent observations taking their values in some measurable space $\Theta \times \{0, 1\}$, with common distribution P. We denote d the $L^2(\mu)$ distance where μ is the marginal distribution of X_i.*

Let S_T the set of piecewise constant functions defined on the partition $\tilde{T}$ associated to the leaves of the tree T.

Let suppose that:

$$\exists h > 0,\ \forall x \in \Theta,\ |2\eta(x) - 1| \geq h \quad \text{with} \quad \eta(x) = \mathbb{P}(Y = 1 | X = x)$$

Then:

(i) $\sup_{u \in S_T,\ l(s,u) \leq \varepsilon^2} d(s, u) \leq w(\varepsilon)$ *with* $w(x) = \frac{1}{\sqrt{h}} x$,

(ii) $\exists \phi_T : \mathbb{R}^+ \to \mathbb{R}^+$ *such that:*

- $\phi_T(0) = 0$,
- $x \to \frac{\phi_T(x)}{x}$ *is non increasing,*
- $\forall \sigma \geq w(\sigma_T),\quad \sqrt{n}\mathbb{E}\left[\sup_{u \in S_T,\ d(u,v) \leq \sigma} |\bar{\gamma}_n(u) - \bar{\gamma}_n(v)|\right] \leq \phi_T(\sigma)$,

with σ_T the positive solution of $\phi_T(w(x)) = \sqrt{n}x^2$ and with $\bar{\gamma}_n$ the centered empirical process (for a more detailed definition see Massart and Nédélec [24]).

(iii) $\sigma_T^2 \leq \frac{K_3^2 |T|}{nh}$.

Proof. The first point (i) is easy to obtain from the following expression of l:

$$l(s, u) = \mathbb{E}\left(|s(X) - u(X)|\,|2\eta(X) - 1|\right)$$

The existence of the function ϕ_T has been proved by Massart and Nédélec [24]. They also give an upper bound of σ_T^2 based on Sauer's lemma. The upper bound of σ_T^2 is better than the one of [24] because it has been adapted to the structure of S_T. □

Thanks to Lemmas 6.8 and 6.2, we deduce the next one.

Lemma 6.9. *Let $(X_1, Y_1), \dots, (X_n, Y_n)$ a sample taking its values in some measurable space $\Theta \times \{0, 1\}$, with common distribution P. Let T a tree, S_T the space associated, h the margin and K_3 the universal constant which appear in the Lemma 6.8. If $2x \geq \frac{K_3\sqrt{|T|}}{\sqrt{n}h}$, then:*

$$\mathbb{E}\left[\sup_{u \in S_T} \frac{|\bar{\gamma}_n(u) - \bar{\gamma}_n(v)|}{d^2(u, v) + (2x)^2}\right] \leq \frac{2K_3\sqrt{|T|}}{x\sqrt{n}}$$

7. Proofs

In this paper, the proofs are not fully detailed. All the details can be found in [29].

7.1. Classification

In the sequel, to simplify the notations, we note $pen\,(M, T)$ the function $pen_c\,(M, T, \alpha, \beta, h)$.

7.1.1. Proof of the Proposition 3.1:

Let M $\in \mathcal{P}(\Lambda)$, $T \preceq T_{\max}^{(M)}$ and $s_{M,T} \in S_{M,T}$. We let

- $w_{M',T'}(u) = (d(s, s_{M,T}) + d(s,u))^2 + y_{M',T'}^2$
- $V_{M',T'} = \sup_{u \in S_{M',T'}} \dfrac{|\gamma_{n_2}^-(u) - \gamma_{n_2}^-(s_{M,T})|}{w_{M',T'}(u)}$

where $y_{M',T'}$ is a parameter that will be chosen later.

Following the proof of Theorem 4.2 in [22], we get

$$l(s,\tilde{s}) \leq l(s, s_{M,T}) + w_{\widehat{M,T}}(\tilde{s}) \times V_{\widehat{M,T}} + pen(M,T) - pen(\widehat{M,T}) \tag{7.1}$$

To control $V_{\widehat{M,T}}$, we check a uniform overestimation of $V_{M',T'}$. To do this, we apply the Talagrand's concentration inequality, written in Lemma 6.1, to $V_{M',T'}$. So we obtain that for any (M',T'), and for any $x > 0$

$$\mathbb{P}\left(V_{M',T'} \geq K_1 \mathbb{E}\left[V_{M',T'}\right] + K_2\left(\sqrt{\frac{x}{2n_2}} y_{M',T'}^{-1} + \frac{x}{n_2} y_{M',T'}^{-2}\right)\right) \leq \mathrm{e}^{-x}$$

where K_1 and K_2 are universal positive constants.

Setting $x = x_{M',T'} + \xi$, with $\xi > 0$ and the weights $x_{M',T'} = a|T'| + b|M'|\left(1 + \log\left(\frac{p}{|M'|}\right)\right)$, as defined in Lemma 6.6, and summing all those inequalities with respect to (M',T'), we derive a set $\Omega_{\xi,(M,T)}$ such that:

- $\mathbb{P}\left(\Omega_{\xi,(M,T)}^c | \mathcal{L}_1 \text{ and } \{X_i,\ (X_i,Y_i) \in \mathcal{L}_2\}\right) \leq \mathrm{e}^{-\xi}\Sigma(a,b)$
- on $\Omega_{\xi,(M,T)}$, $\forall (M',T')$,

$$V_{M',T'} \leq K_1 \mathbb{E}\left[V_{M',T'}\right] + K_2\left(\sqrt{\frac{x_{M',T'} + \xi}{2n_2}} {y_{M',T'}}^{-1} + \frac{x_{M',T'} + \xi}{n_2} {y_{M',T'}}^{-2}\right) \tag{7.2}$$

Now we overestimate $\mathbb{E}\left[V_{M',T'}\right]$.

Let $u_{M',T'} \in S_{M',T'}$ such that $d(s, u_{M',T'}) \leq \inf_{u \in S_{M',T'}} d(s,u)$.

Then

$$\mathbb{E}\left[V_{M',T'}\right] \leq \mathbb{E}\left[\frac{|\gamma_{n_2}^-(u_{M',T'}) - \gamma_{n_2}^-(s_{M,T})|}{\inf_{u \in S_{M',T'}}(w_{M',T'}(u))}\right] + \mathbb{E}\left[\sup_{u \in S_{M',T'}}\left(\frac{|\gamma_{n_2}^-(u) - \gamma_{n_2}^-(u_{M',T'})|}{w_{M',T'}(u)}\right)\right]$$

We prove:

$$\mathbb{E}\left[\frac{|\gamma_{n_2}^-(u_{M',T'}) - \gamma_{n_2}^-(s_{M,T})|}{\inf_{u \in S_{M',T'}}(w_{M',T'}(u))}\right] \leq \frac{1}{\sqrt{n_2} y_{M',T'}}$$

For the second term, thanks to the Lemma 6.9, we have for $2y_{M',T'} \geq \frac{K_3\sqrt{|T'|}}{\sqrt{n_2}h}$,

$$\mathbb{E}\left[\sup_{u \in S_{M',T'}}\left(\frac{|\gamma_{n_2}^-(u) - \gamma_{n_2}^-(u_{M',T'})|}{w_{M',T'}(u)}\right)\right] \leq \frac{8K_3\sqrt{|T'|}}{\sqrt{n_2} y_{M',T'}}$$

Thus from (7.2), we know that on $\Omega_{\xi,(M,T)}$ and $\forall (M',T')$

$$V_{M',T'} \leq \frac{K_1}{\sqrt{n_2} y_{M',T'}}\left(8K_3\sqrt{|T'|} + 1\right) + K_2\left(\sqrt{\frac{x_{M',T'} + \xi}{2n_2}} {y_{M',T'}}^{-1} + \frac{x_{M',T'} + \xi}{n_2} {y_{M',T'}}^{-2}\right)$$

For $y_{M',T'} = 3K\left(\frac{K_1}{\sqrt{n_2}}\left(8K_3\sqrt{|T'|}+1\right)+K_2\sqrt{\frac{x_{M',T'}+\xi}{2n_2}}+\frac{1}{\sqrt{3K}}\sqrt{K_2\frac{x_{M',T'}+\xi}{n_2}}\right)$ with $K \geq \frac{1}{48K_1h}$, we get:

$$V_{M',T'} \leq \frac{1}{K}$$

By overestimating $w_{\widehat{M,T}}(\tilde{s})$, $y^2_{\widehat{M,T}}$ and replacing all of those results in (7.1), we get

$$\begin{aligned}\left(1-\frac{2}{Kh}\right)l\left(s,\tilde{s}\right) \leq \left(1+\frac{2}{Kh}\right)l\left(s,s_{M,T}\right) - pen(\widehat{M,T}) + pen(M,T)\\ + 18K\left(\frac{64{K_1}^2{K_3}^2}{n_2}|\hat{T}| + 2K_2\frac{x_{\widehat{M,T}}}{n_2}\left(\sqrt{\frac{K_2}{2}}+\frac{1}{\sqrt{3K}}\right)^2\right)\\ + 18K\left(\frac{2K_1^2}{n_2}+2K_2\frac{\xi}{n_2}\left(\sqrt{\frac{K_2}{2}}+\frac{1}{\sqrt{3K}}\right)^2\right)\end{aligned}$$

We let $K = \frac{2}{h}\frac{C_1+1}{C_1-1}$ with $C_1 > 1$.

Taking a penalty $pen(\widehat{M,T})$ which balances all the terms in $\left(\widehat{M,T}\right)$, *i.e.*

$$pen(M,T) \geq \frac{36(C_1+1)}{h(C_1-1)}\left(\frac{64K_1^2K_3^2}{n_2}|T| + 2K_2\frac{x_{M,T}}{n_2}\left(\sqrt{\frac{K_2}{2}}+\sqrt{\frac{C_1-1}{6(C_1+1)}}\right)^2\right)$$

We obtain that on $\Omega_{\xi,(M,T)}$

$$l(s,\tilde{s}) \leq C_1\left(l(s,s_{M,T})+pen(M,T)\right)+\frac{C}{n_2h}\xi$$

Integrating with respect to ξ and by minimizing, we get

$$\mathbb{E}\left[l(s,\tilde{s})|\mathcal{L}_1\right] \leq C_1 \inf_{M,T}\left\{l(s,S_{M,T})+pen(M,T)\right\}+\frac{C}{n_2h}\Sigma(a,b)$$

The two constants α_0 and β_0, which appear in the proposition 3.1, are defined by

$$\alpha_0 = 36\left(64K_1^2K_3^2+4\log 2K_2\left(\sqrt{\frac{K_2}{2}}+\frac{1}{\sqrt{6}}\right)^2\right) \quad \text{and} \quad \beta_0 = 72K_2\left(\sqrt{\frac{K_2}{2}}+\frac{1}{\sqrt{6}}\right)^2$$ □

7.1.2. Proof of the Proposition 3.2

This proof is quite similar to the previous one. We just need to replace $w_{M',T'}(u)$ and $V_{M',T'}$ by

- $w_{(M',T'),(M,T)}(u) = (d(s,s_{M,T})+d(s,u))^2+(y_{M',T'}+y_{M,T})^2$
- $V_{(M',T'),(M,T)} = \sup_{u\in S_{M',T'}}\frac{|\gamma_{n_1}^-(u)-\gamma_{n_1}^-(s_{M,T})|}{w_{(M',T'),(M,T)}(u)}$

And like the proof of Proposition 4.3, we change the conditioning. □

7.1.3. Proof of the Proposition 3.3

This result is obtained by a direct application of the Lemma 6.3 which appears in the Section 6. □

7.2. Regression

In the sequel, to simplify the notations, we note $pen(M,T)$ the function $pen_r(M,T,\alpha,\beta,\rho,R)$.

7.2.1. Proof of the Proposition 4.1

Let $a > 2\log 2$, $b > 1$, $\theta \in (0,1)$ and $K > 2-\theta$ four constants.

Let us denote

$$s_{M,T} = \underset{u\in S_{M,T}}{\operatorname{argmin}} \|s-u\|_{n_2}^2 \quad \text{and} \quad \varepsilon_{M,T} = \underset{u\in S_{M,T}}{\operatorname{argmin}} \|\varepsilon-u\|_{n_2}^2$$

Following the proof of Theorem 1 in [2], we get

$$(1-\theta)\|s-\tilde{s}\|_{n_2}^2 = \Delta_{\widehat{M,T}} + \inf_{(M,T)} R_{M,T} \tag{7.3}$$

where

$$\Delta_{M,T} = (2-\theta)\|\varepsilon_{M,T}\|_{n_2}^2 - 2\langle \varepsilon, s-s_{M,T}\rangle_{n_2} - \theta\|s-s_{M,T}\|_{n_2}^2 - pen(M,T)$$
$$R_{M,T} = \|s-s_{M,T}\|_{n_2}^2 - \|\varepsilon_{M,T}\|_{n_2}^2 + 2\langle \varepsilon, s-s_{M,T}\rangle_{n_2} + pen(M,T)$$

We are going first to control $\Delta_{\widehat{M,T}}$ by using concentration inequalities of $\|\varepsilon_{M,T}\|_{n_2}^2$ and $-\langle \varepsilon, s-s_{M,T}\rangle_{n_2}$.

For any M, we denote

$$\Omega_M = \left\{ \forall t \in \widetilde{T_{\max}^{(M)}} \ \left| \sum_{X_i\in t} \varepsilon_i \right| \le \frac{\sigma^2}{\rho}|X_i \in t| \right\}$$

Thanks to Lemma 6.4, we get that for any (M,T) and any $x>0$

$$\mathbb{P}\left(\|\varepsilon_{M,T}\|_{n_2}^2 \mathbb{1}_{\Omega_M} \ge \frac{\sigma^2}{n_2}|T| + 8\frac{\sigma^2}{n_2}\sqrt{2|T|x} + 4\frac{\sigma^2}{n_2}x \ \middle| \mathcal{L}_1 \text{ and } \{X_i;\ (X_i,Y_i)\in\mathcal{L}_2\} \right) \le \mathrm{e}^{-x} \tag{7.4}$$

and

$$\mathbb{P}\left(\Omega_M^c \ \middle| \mathcal{L}_1 \text{ and } \{X_i;\ (X_i,Y_i)\in\mathcal{L}_2\} \right) \le 2\frac{n_2}{N_{\min}} \exp\left(\frac{-\sigma^2 N_{\min}}{4\rho^2} \right)$$

Denoting $\Omega = \bigcap_M \Omega_M$, we have

$$\mathbb{P}\left(\Omega^c \ \middle| \mathcal{L}_1 \text{ and } \{X_i;\ (X_i,Y_i)\in\mathcal{L}_2\} \right) \le 2^{p+1}\frac{n_2}{N_{\min}} \exp\left(\frac{-\sigma^2 N_{\min}}{4\rho^2} \right)$$

Thanks to assumption (2.3) and $\|s\|_\infty \le R$, we easily obtain for any (M,T) and any $x>0$

$$\mathbb{P}\Big(-\langle \varepsilon, s-s_{M,T}\rangle_{n_2} \ge \frac{\sigma}{\sqrt{n_2}}\|s-s_{M,T}\|_{n_2}\sqrt{2x} + \frac{2\rho R}{n_2}x \ \Big| \mathcal{L}_1 \text{ and } \{X_i;\ (X_i,Y_i)\in\mathcal{L}_2\} \Big) \le \mathrm{e}^{-x} \tag{7.5}$$

Setting $x = x_{M,T} + \xi$ with $\xi > 0$ and the weights $x_{M,T} = a|T| + b|M|\left(1+\log\left(\frac{p}{|M|}\right)\right)$ as defined in Lemma 6.6, and summing all inequalities (7.4) and (7.5) with respect to (M,T), we derive a set E_ξ such that

- $\mathbb{P}\left(E_\xi^c \ | \mathcal{L}_1 \text{ and } \{X_i;\ (X_i,Y_i)\in\mathcal{L}_2\} \right) \le 2\mathrm{e}^{-\xi}\Sigma(a,b)$

- on the set $E_\xi \bigcap \Omega$, for any (M,T),

$$\begin{aligned}\Delta_{M,T} \leq (2-\theta)\frac{\sigma^2}{n_2}|T| + 8(2-\theta)\frac{\sigma^2}{n_2}\sqrt{2|T|(x_{M,T}+\xi)} + 4(2-\theta)\frac{\sigma^2}{n_2}(x_{M,T}+\xi)\\ + 2\frac{\sigma}{\sqrt{n_2}}\|s-s_{M,T}\|_{n_2}\sqrt{2(x_{M,T}+\xi)} + 4\frac{\rho R}{n_2}(x_{M,T}+\xi)\\ - \theta\|s-s_{M,T}\|_{n_2}^2 - pen(M,T)\end{aligned}$$

where $\Sigma(a,b) = -\log\left(1-\mathrm{e}^{-(a-2\log 2)}\right)\frac{\mathrm{e}^{-(b-1)}}{1-\mathrm{e}^{-(b-1)}}$.

Using the two following inequalities

$2\frac{\sigma}{\sqrt{n_2}}\|s-s_{M,T}\|_{n_2}\sqrt{2(x_{M,T}+\xi)} \leq \theta\|s-s_{M,T}\|_{n_2}^2 + \frac{2}{\theta}\frac{\sigma^2}{n_2}(x_{M,T}+\xi)$,

$2\sqrt{|T|(x_{M,T}+\xi)} \leq \eta|T| + \eta^{-1}(x_{M,T}+\xi)$

with $\eta = \frac{K+\theta-2}{2-\theta}\frac{1}{4\sqrt{2}} > 0$, we derive that on the set $E_\xi \bigcap \Omega$, for any (M,T),

$$\Delta_{M,T} \leq K\frac{\sigma^2}{n_2}|T| + \left(4(2-\theta)\left(1+\frac{8(2-\theta)}{K+\theta-2}\right) + \frac{2}{\theta} + 4\frac{\rho}{\sigma^2}R\right)\frac{\sigma^2}{n_2}(x_{M,T}+\xi) - pen(M,T)$$

Taking a penalty $pen(M,T)$ which compensates for all the other terms in (M,T), *i.e.*

$$pen(M,T) \geq K\frac{\sigma^2}{n_2}|T| + \left[4(2-\theta)\left(1+\frac{8(2-\theta)}{K+\theta-2}\right) + \frac{2}{\theta} + 4\frac{\rho}{\sigma^2}R\right]\frac{\sigma^2}{n_2}x_{M,T} \tag{7.6}$$

we get that, on the set E_ξ

$$\Delta_{\widehat{M,T}}\mathbb{1}_\Omega \leq \left(4(2-\theta)\left(1+\frac{8(2-\theta)}{K+\theta-2}\right) + \frac{2}{\theta} + 4\frac{\rho}{\sigma^2}R\right)\frac{\sigma^2}{n_2}\xi$$

Integrating with respect to ξ, we derive

$$\mathbb{E}\left[\Delta_{\widehat{M,T}}\mathbb{1}_\Omega\middle|\mathcal{L}_1\right] \leq 2\left(4(2-\theta)\left(1+\tfrac{8(2-\theta)}{K+\theta-2}\right) + \tfrac{2}{\theta} + 4\tfrac{\rho}{\sigma^2}R\right)\tfrac{\sigma^2}{n_2}\Sigma(a,b) \tag{7.7}$$

We are going now to control $\mathbb{E}\left[\inf_{(M,T)} R_{M,T}\mathbb{1}_\Omega\middle|\mathcal{L}_1\right]$.

In the same way we deduced (7.5) from assumption (2.3), we get that for any (M,T) and any $x>0$

$$\begin{aligned}\mathbb{P}\Big(\langle\varepsilon, s-s_{M,T}\rangle_{n_2} \geq \frac{\sigma}{\sqrt{n_2}}\|s-s_{M,T}\|_{n_2}\sqrt{2x} + \frac{2\rho R}{n_2}x \,\Big|\mathcal{L}_1 \text{ and } \{X_i;\ (X_i,Y_i)\in\mathcal{L}_2\}\Big)\\ \leq \mathrm{e}^{-x}\end{aligned}$$

Thus we derive a set F_ξ such that

- $\mathbb{P}\left(F_\xi^c\ |\mathcal{L}_1 \text{ and } \{X_i;\ (X_i,Y_i)\in\mathcal{L}_2\}\right) \leq \mathrm{e}^{-\xi}\Sigma(a,b)$
- on the set F_ξ, for any (M,T),

$$\langle\varepsilon, s-s_{M,T}\rangle_{n_2} \leq \frac{\sigma}{\sqrt{n_2}}\|s-s_{M,T}\|_{n_2}\sqrt{2\left(x_{M,T}+\xi\right)} + \frac{2\rho R}{n_2}\left(x_{M,T}+\xi\right)$$

It follows from definition of $R_{M,T}$ and inequality (7.6) on the penalty that

$$\mathbb{E}\left[\inf_{(M,T)} R_{M,T}\mathbb{1}_{\Omega}\middle|\mathcal{L}_1\right] \leq 2\inf_{(M,T)}\left\{\mathbb{E}\left[\|s-s_{M,T}\|_{n_2}^2\middle|\mathcal{L}_1\right]+pen(M,T)\right\} + \left(\frac{2}{\theta}+4\frac{\rho}{\sigma^2}R\right)\frac{\sigma^2}{n_2}\Sigma(a,b) \tag{7.8}$$

We conclude from (7.3), (7.7) and (7.8) that

$$(1-\theta)\mathbb{E}\left[\|s-\tilde{s}\|_{n_2}^2\mathbb{1}_{\Omega}\middle|\mathcal{L}_1\right] \leq 2\inf_{(M,T)}\left\{\mathbb{E}\left[\|s-s_{M,T}\|_{n_2}^2\middle|\mathcal{L}_1\right]+pen(M,T)\right\} + \left(8(2-\theta)\left(1+\frac{8(2-\theta)}{K+\theta-2}\right)+\frac{6}{\theta}+12\frac{\rho}{\sigma^2}R\right)\frac{\sigma^2}{n_2}\Sigma(a,b)$$

It remains to control $\mathbb{E}\left[\|s-\tilde{s}\|_{n_2}^2\mathbb{1}_{\Omega^c}\middle|\mathcal{L}_1\right]$, except if $\rho=0$ in which case it is finished.

After some calculations (see the proof of Theorem 1 in [28] for more details), we get

$$\mathbb{E}\left[\|s-\tilde{s}\|_{n_2}^2\mathbb{1}_{\Omega^c}\middle|\mathcal{L}_1\right] \leq R^2\mathbb{P}\left(\Omega^c\middle|\mathcal{L}_1\right)+\sum_{M}\sqrt{\mathbb{E}\left[\|\varepsilon_{M,T_{\max}^{(M)}}\|_{n_2}^4\middle|\mathcal{L}_1\right]}\sqrt{\mathbb{P}\left(\Omega^c\middle|\mathcal{L}_1\right)}$$

and

$$\mathbb{E}\left[\|\varepsilon_{M,T_{\max}^{(M)}}\|_{n_2}^4\middle|\mathcal{L}_1\right] \leq \frac{C^2(\rho,\sigma)}{N_{\min}^2}$$

where $C(\rho,\sigma)$ is a constant which depends only on ρ and σ.

Thus we have

$$\mathbb{E}\left[\|s-\tilde{s}\|_{n_2}^2\mathbb{1}_{\Omega^c}\middle|\mathcal{L}_1\right] \leq R^2\mathbb{P}\left(\Omega^c\middle|\mathcal{L}_1\right)+2^p\frac{C(\rho,\sigma)}{N_{\min}}\sqrt{\mathbb{P}\left(\Omega^c\middle|\mathcal{L}_1\right)}$$

Let us recall that

$$\mathbb{P}\left(\Omega^c\middle|\mathcal{L}_1\right) \leq 2^{p+1}\frac{n_2}{N_{\min}}\exp\left(\frac{-\sigma^2 N_{\min}}{4\rho^2}\right)$$

For $p \leq \log n_2$ and $N_{\min} \geq \frac{24\rho^2}{\sigma^2}\log n_2$,

- $2^p\sqrt{\mathbb{P}\left(\Omega_\delta^c\middle|\mathcal{L}_1\right)} \leq \frac{\sigma}{\sqrt{12}\rho}\frac{1}{n_2\sqrt{\log n_2}}$
- $\mathbb{P}\left(\Omega_\delta^c\middle|\mathcal{L}_1\right) \leq \frac{\sigma^2}{12\rho^2}\frac{1}{n_2^4\log n_2}$

It follows that

$$\mathbb{E}\left[\|s-\tilde{s}\|_{n_2}^2\mathbb{1}_{\Omega^c}\middle|\mathcal{L}_1\right] \leq C'(\rho,\sigma,R)\frac{1}{n_2(\log n_2)^{3/2}}$$

Finally, we have the following result:

Denoting by $\Upsilon=\left[4(2-\theta)\left(1+\frac{8(2-\theta)}{K+\theta-2}\right)+\frac{2}{\theta}\right]$ and taking a penalty which satisfies $\forall\, M\in\mathcal{P}(\Lambda)\ \forall\, T\preceq T_{\max}^{(M)}$

$$pen(M,T) \geq \left((K+a\Upsilon)\sigma^2+4a\rho R\right)\frac{|T|}{n_2}+\left(b\Upsilon\sigma^2+4b\rho R\right)\frac{|M|}{n_2}\left(1+\log\left(\frac{p}{|M|}\right)\right)$$

if $p \leq \log n_2$ and $N_{\min} \geq \frac{24\rho^2}{\sigma^2} \log n_2$, we have,

$$(1-\theta)\mathbb{E}\left[\|s-\tilde{s}\|_{n_2}^2 \,|\mathcal{L}_1\right] \leq 2 \inf_{(M,T)} \left\{ \inf_{u \in S_{M,T}} \|s-u\|_\mu^2 + pen(M,T) \right\} + \left(2\Upsilon + 2 + 12\frac{\rho}{\sigma^2}R\right)\frac{\sigma^2}{n_2}\Sigma(a,b) + (1-\theta)C'(\rho,\sigma,R)\frac{1}{n_2(\log n_2)^{3/2}}$$

We deduce the proposition by taking $K=2$, $\theta \to 1$, $a \to 2\log 2$ and $b \to 1$. □

7.2.2. Proof of the Proposition 4.3

To follow the preceding proof, we have to consider the "deterministic" bigger collection of models:

$$\{S_{M,T};\ T \in \mathcal{M}_{n_1,M} \text{ and } M \in \mathcal{P}(\Lambda)\}$$

where $\mathcal{M}_{n_1,M}$ denote the set of trees built on the grid $\{X_i;\ (X_i,Y_i) \in \mathcal{L}_1\}$ with splits on the variables in M. By considering this bigger collection of models, we no longer have partitions built from an initial one. So, we use Lemma 6.5 $\left(\text{with } \delta = 5\frac{\rho}{\sigma^2}\log\left(\frac{n_1}{p}\right)\right)$ instead of Lemma 6.4. The steps of the proof are the same as before. The main difference is that, the quantities are now conditioned by $\{X_i;\ (X_i,Y_i) \in \mathcal{L}_1\}$ instead of $\mathcal{L}_1$ and $\{X_i;\ (X_i,Y_i) \in \mathcal{L}_2\}$. □

7.2.3. Proof of the Proposition 4.5

It follows from the definition of $\tilde{\tilde{s}}$ that for any $\tilde{s}(\alpha,\beta) \in \mathcal{G}$

$$\left\|s-\tilde{\tilde{s}}\right\|_{n_3}^2 \leq \|s-\tilde{s}(\alpha,\beta)\|_{n_3}^2 + 2\left\langle \varepsilon, \tilde{\tilde{s}} - \tilde{s}(\alpha,\beta)\right\rangle_{n_3} \tag{7.9}$$

Denoting $M_{\alpha,\beta,\alpha',\beta'} = \max\{|\tilde{s}(\alpha',\beta')(X_i) - \tilde{s}(\alpha,\beta)(X_i)|\,;\ (X_i,Y_i) \in \mathcal{L}_3\}$, and thanks to assumption (2.3) we get that for any $\tilde{s}(\alpha,\beta)$, $\tilde{s}(\alpha',\beta') \in \mathcal{G}$ and any $x>0$

$$\mathbb{P}\left(\langle \varepsilon, \tilde{s}(\alpha',\beta') - \tilde{s}(\alpha,\beta)\rangle_{n_3} \geq \tfrac{\sigma}{\sqrt{n_3}}\|\tilde{s}(\alpha',\beta') - \tilde{s}(\alpha,\beta)\|_{n_3}\sqrt{2x} + M_{\alpha,\beta,\alpha',\beta'}\tfrac{\rho}{n_3}x \,\middle|\, \mathcal{L}_1,\ \mathcal{L}_2, \{X_i,\ (X_i,Y_i) \in \mathcal{L}_3\}\right) \leq \mathrm{e}^{-x}$$

Setting $x = 2\log\mathcal{K} + \xi$ with $\xi > 0$, and summing all these inequalities with respect to $\tilde{s}(\alpha,\beta)$ and $\tilde{s}(\alpha',\beta') \in \mathcal{G}$, we derive a set E_ξ such that

- $\mathbb{P}\left(E_\xi^c \,|\mathcal{L}_1,\ \mathcal{L}_2,\ \text{and } \{X_i;\ (X_i,Y_i) \in \mathcal{L}_3\}\right) \leq \mathrm{e}^{-\xi}$
- on the set E_ξ, for any $\tilde{s}(\alpha,\beta)$ and $\tilde{s}(\alpha',\beta') \in \mathcal{G}$

$$\langle \varepsilon, \tilde{s}(\alpha',\beta') - \tilde{s}(\alpha,\beta)\rangle_{n_3} \leq \frac{\sigma}{\sqrt{n_3}}\|\tilde{s}(\alpha',\beta') - \tilde{s}(\alpha,\beta)\|_{n_3}\sqrt{2(2\log\mathcal{K}+\xi)} + M_{\alpha,\beta,\alpha',\beta'}\frac{\rho}{n_3}(2\log\mathcal{K}+\xi)$$

It remains to control $M_{\alpha,\beta,\alpha',\beta'}$ in the two situations $(M1)$ and $(M2)$ (except if $\rho=0$).

In the $(M1)$ situation, we consider the set

$$\Omega_1 = \bigcap_{M \in \mathcal{P}(\Lambda)} \left\{ \forall t \in \widetilde{T_{\max}^{(M)}} \left| \sum_{\substack{(X_i,Y_i)\in\mathcal{L}_2 \\ X_i \in t}} \varepsilon_i \right| \leq R\,|\{i;\ (X_i,Y_i) \in \mathcal{L}_2 \text{ and } X_i \in t\}| \right\}$$

Thanks to assumption (2.3), we deduce that for any $x > 0$

$$\mathbb{P}\left(\left|\sum_{\substack{(X_i,Y_i)\in\mathcal{L}_2\\ X_i\in t}}\varepsilon_i\right| \geq x \middle| \mathcal{L}_1 \text{ and } \{X_i;\ (X_i,Y_i)\in\mathcal{L}_2\}\right) \leq 2\mathrm{e}^{\frac{-x^2}{2(\sigma^2|\{i;\ (X_i,Y_i)\in\mathcal{L}_2 \text{ and } X_i\in t\}|+\rho x)}}$$

Taking $x = R\,|\{i;\ (X_i,Y_i)\in\mathcal{L}_2 \text{ and } X_i\in t\}|$ and summing all these inequalities, we get that

$$\mathbb{P}\left(\Omega_1^c\middle|\ \mathcal{L}_1 \text{ and } \{X_i;\ (X_i,Y_i)\in\mathcal{L}_2\}\right) \leq 2^{p+1}\frac{n_1}{N_{\min}}\exp\left(\frac{-R^2N_{\min}}{2(\sigma^2+\rho R)}\right)$$

On the set Ω_1, as for any (M,T), $\|\hat{s}_{M,T}\|_\infty \leq 2R$, we have $M_{\alpha,\beta,\alpha',\beta'} \leq 4R$.

Thus, on the set $\Omega_1\bigcap E_\xi$, for any $\tilde{s}(\alpha,\beta)\in\mathcal{G}$

$$\left\langle\varepsilon,\tilde{\tilde{s}}-\tilde{s}(\alpha,\beta)\right\rangle_{n_3} \leq \frac{\sigma}{\sqrt{n_3}}\|\tilde{\tilde{s}}-\tilde{s}(\alpha,\beta)\|_{n_3}\sqrt{2(2\log\mathcal{K}+\xi)} + 4R\frac{\rho}{n_3}(2\log\mathcal{K}+\xi)$$

It follows from (7.9) that, on the set $\Omega_1\bigcap E_\xi$, for any $\tilde{s}(\alpha,\beta)\in\mathcal{G}$ and any $\eta\in(0;1)$

$$\eta^2\left\|s-\tilde{\tilde{s}}\right\|_{n_3}^2 \leq (1+\eta^{-1}-\eta)\left\|s-\tilde{s}(\alpha,\beta)\right\|_{n_3}^2 + \left(\frac{2}{1-\eta}\sigma^2+8\rho R\right)\frac{(2\log\mathcal{K}+\xi)}{n_3}$$

Taking $p\leq\log n_2$ and $N_{\min}\geq 4\frac{\sigma^2+\rho R}{R^2}\log n_2$, we have

$$\mathbb{P}\left(\Omega_1^c\right) \leq \frac{R^2}{2(\sigma^2+\rho R)}\frac{1}{n_2^{1-\log 2}}$$

Finally, in the $(M1)$ situation, we have for any $\xi>0$, with probability $\geq 1-\mathrm{e}^{-\xi}-\frac{R^2}{2(\sigma^2+\rho R)}\frac{1}{n_2^{1-\log 2}}$, $\forall\eta\in(0,1)$,

$$\left\|s-\tilde{\tilde{s}}\right\|_{n_3}^2 \leq \frac{(1+\eta^{-1}-\eta)}{\eta^2}\inf_{\tilde{s}(\alpha,\beta)\in\mathcal{G}}\left\|s-\tilde{s}(\alpha,\beta)\right\|_{n_3}^2 + \frac{1}{\eta^2}\left(\frac{2}{1-\eta}\sigma^2+8\rho R\right)\frac{(2\log\mathcal{K}+\xi)}{n_3}$$

In the $(M2)$ situation, we consider the set

$$\Omega_2 = \{\forall 1\leq i\leq n_1\ |\varepsilon_i| \leq 3\rho\log n_1\}$$

Thanks to assumption (2.3), we get that

$$\mathbb{P}\left(\Omega_2^c\middle|\ \{X_i;\ (X_i,Y_i)\in\mathcal{L}_1\}\right) \leq 2n_1\exp\left(-\frac{9\rho^2\log^2 n_1}{2(\sigma^2+3\rho^2\log n_1)}\right)$$

with $\epsilon(n_1) = 2n_1\exp\left(-\frac{9\rho^2\log^2 n_1}{2(\sigma^2+3\rho^2\log n_1)}\right) \underset{n_1\to+\infty}{\longrightarrow} 0$

On the set Ω_2, as for any (M,T), $\|\hat{s}_{M,T}\|_\infty \leq R+3\rho\log n_1$, we have $M_{\alpha,\beta,\alpha',\beta'} \leq 2(R+3\rho\log n_1)$.

Thus, on the set $\Omega_2\bigcap E_\xi$, for any $\tilde{s}(\alpha,\beta)\in\mathcal{G}$

$$\left\langle\varepsilon,\tilde{\tilde{s}}-\tilde{s}(\alpha,\beta)\right\rangle_{n_3} \leq \frac{\sigma}{\sqrt{n_3}}\|\tilde{\tilde{s}}-\tilde{s}(\alpha,\beta)\|_{n_3}\sqrt{2(2\log\mathcal{K}+\xi)} + 2(R+3\rho\log n_1)\frac{\rho}{n_3}(2\log\mathcal{K}+\xi)$$

It follows from (7.9) that, on the set $\Omega_2\bigcap E_\xi$, for any $\tilde{s}(\alpha,\beta)\in\mathcal{G}$ and any $\eta\in(0;1)$

$$\eta^2\left\|s-\tilde{\tilde{s}}\right\|_{n_3}^2 \leq (1+\eta^{-1}-\eta)\left\|s-\tilde{s}(\alpha,\beta)\right\|_{n_3}^2 + \left(\frac{2}{1-\eta}\sigma^2+4\rho(R+3\rho\log n_1)\right)\frac{(2\log\mathcal{K}+\xi)}{n_3}$$

Finally, in the $(M2)$ situation, we have for any $\xi > 0$, with probability $\geq 1 - \mathrm{e}^{-\xi} - \epsilon(n_1)$, $\forall \eta \in (0,1)$,

$$\left\|s - \tilde{\tilde{s}}\right\|_{n_3}^2 \leq \frac{(1+\eta^{-1}-\eta)}{\eta^2} \inf_{\tilde{s}(\alpha,\beta)\in\mathcal{G}} \left\|s - \tilde{s}(\alpha,\beta)\right\|_{n_3}^2 + \frac{1}{\eta^2}\left(\frac{2}{1-\eta}\sigma^2 + 4\rho R + 12\rho^2 \log n_1\right)\frac{(2\log\mathcal{K} + \xi)}{n_3}$$

□

References

[1] S. Arlot and P. Bartlett, Margin adaptive model selection in statistical learning. *Bernoulli* **17** (2011) 687–713.
[2] L. Birgé and P. Massart, Minimal penalties for gaussian model selection. *Probab. Theory Relat. Fields* **138** (2007) 33–73.
[3] L. Breiman, Random forests. *Mach. Learn.* **45** (2001) 5–32.
[4] L. Breiman and A. Cutler, Random forests. http://www.stat.berkeley.edu/users/breiman/RandomForests/ (2005).
[5] L. Breiman, J. Friedman, R. Olshen and C. Stone, Classification and Regression Trees. Chapman et Hall (1984).
[6] R. Díaz-Uriarte and S. Alvarez de Andrés, Gene selection and classification of microarray data using random forest. *BMC Bioinform.* **7** (2006) 1–13.
[7] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani, Least angle regression. *Ann. Stat.* **32** (2004) 407–499.
[8] J. Fan and J. Lv, A selective overview of variable selection in high dimensional feature space. *Stat. Sin.* **20** (2010) 101–148.
[9] G.M. Furnival and R.W. Wilson, Regression by leaps and bounds. *Technometrics* **16** (1974) 499–511.
[10] R. Genuer, J.M. Poggi and C. Tuleau-Malot, Variable selection using random forests. *Pattern Recognit. Lett.* **31** (2010) 2225–2236.
[11] S. Gey, Margin adaptive risk bounds for classification trees, hal-00362281.
[12] S. Gey and E. Nédélec, Model Selection for CART Regression Trees. *IEEE Trans. Inf. Theory* **51** (2005) 658–670.
[13] B. Ghattas and A. Ben Ishak, Sélection de variables pour la classification binaire en grande dimension: comparaisons et application aux données de biopuces. *Journal de la société française de statistique* **149** (2008) 43–66.
[14] U. Grömping, Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician* **61** (2007) 139–147.
[15] I. Guyon and A. Elisseff, An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3** (2003) 1157–1182.
[16] I. Guyon, J. Weston, S. Barnhill and V.N. Vapnik, Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46** (2002) 389–422.
[17] T. Hastié, R. Tibshirani and J. Friedman, The Elements of Statistical Learning. Springer (2001).
[18] T. Hesterberg, N.H. Choi, L. Meier and C. Fraley, Least angle regresion and l1 penalized regression: A review. *Stat. Surv.* **2** (2008) 61–93.
[19] R. Kohavi and G.H. John, Wrappers for feature subset selection. *Artificial Intelligence* **97** (1997) 273–324.
[20] V. Koltchinskii, Local rademacher complexities and oracle inequalities in risk minimization. *Ann. Stat.* **34** (2004) 2593–2656.
[21] E. Mammen and A. Tsybakov, Smooth discrimination analysis. *Ann. Stat.* **27** (1999) 1808–1829.
[22] P. Massart, Some applications of concentration inequalities to statistics. *Annales de la faculté des sciences de Toulouse* **2** (2000) 245–303.
[23] P. Massart, Concentration Inequlaities and Model Selection. *Lect. Notes Math.* Springer (2003).
[24] P. Massart and E. Nédélec, Risk bounds for statistical learning. *Ann. Stat.* **34** (2006).
[25] J.M. Poggi and C. Tuleau, Classification supervisée en grande dimension. Application à l'agrément de conduite automobile. *Revue de Statistique Appliquée* **LIV** (2006) 41–60.
[26] E. Rio, Une inégalité de bennett pour les maxima de processus empiriques. *Ann. Inst. Henri Poincaré, Probab. Stat.* **38** (2002) 1053–1057.
[27] A. Saltelli, K. Chan and M. Scott, *Sensitivity Analysis*. Wiley (2000).
[28] M. Sauvé, Histogram selection in non gaussian regression. *ESAIM PS* **13** (2009) 70–86.
[29] M. Sauvé and C. Tuleau-Malot, Variable selection through CART, hal-00551375.
[30] I.M. Sobol, Sensitivity estimates for nonlinear mathematical models. *Math. Mod. Comput. Experiment* **1** (1993) 271–280.
[31] R. Tibshirani, Regression shrinkage and selection *via* Lasso. *J. R. Stat. Soc. Ser. B* **58** (1996) 267–288.
[32] A.B. Tsybakov, Optimal aggregation of classifiers in statistical learning. *Ann. Stat.* **32** (2004) 135–166.