Variable selection through CART

Marie Sauvé, Christine Tuleau

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Variable selection through CART

Marie Sauvé — Christine Tuleau

N° 5912

Mai 2006

Thème COG

ISSN 0249-6399 ISRN INRIA/RR--5912--FR+ENG

Variable selection through CART

Marie Sauvé* † , Christine Tuleau‡†

Thème COG — Systèmes cognitifs
Projet SELECT

Rapport de recherche n° 5912 — Mai 2006 — 30 pages

Abstract: This paper deals with variable selection in the regression or 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 to get "oracle type inequalities" justifying the performances of the proposed procedure. A simulation study completes the theoretical results.

Key-words: variable selection, regression, binary classification, CART

* email : marie.sauve@math.u-psud.fr
† Département de Mathématiques, Université Paris-Sud, 91405 Orsay Cedex
‡ email : christine.tuleau@math.u-psud.fr

Unité de recherche INRIA Futurs
Parc Club Orsay Université, ZAC des Vignes,
4, rue Jacques Monod, 91893 ORSAY Cedex (France)
Téléphone : +33 1 72 92 59 00 — Télécopie : +33 1 72 92 59 ??

Sélection de variables à travers CART

Résumé : Ce papier aborde le thème de la sélection de variables en proposant une procédure automatique et exhaustive qui repose d'une part sur l'utilisation de CART et d'autre part sur la sélection de modèle par minimisation d'un contraste empirique pénalisé.
L'objet de ce travail, de nature théorique, consiste à déterminer, dans la cadre de la régression et de la classification binaire, les fonctions de pénalités adaptés au problème, autrement dit qui permettent d'obtenir des inégalités de type "oracle" et ainsi de justifier de l'efficacité de la procédure proposée. Par ailleurs, un travail de simulation complète ces éléments théoriques.

Mots-clés : sélection de variables, régression, classification binaire, CART

1 Introduction

This paper deals with variable selection in nonlinear regression and classification using CART estimation and model selection approach.

Let us begin this introduction with some basic ideas focusing on linear regression models of the form:

$$Y = \sum_{j=1}^{p} \beta_j X^j + \varepsilon = X\beta + \varepsilon$$

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 Square (OLS) estimator provides an useful way to estimate the vector β but it suffers from a main drawback: it is not adapted to variable selection since, when p is large, many components of β are not equal to zero.
However, if OLS is not a convenient method to perform variable selection, the least squares is a criterion which often appears in model selection.
For example, Ridge Regression and Lasso are penalized versions of OLS. Ridge Regression (see [7]) involves a L_2 penalization which produces the shrinkage of β but does not put any coefficients of β to zero. So, Ridge Regression is better than OLS, but it is not a variable selection method unlike Lasso. Lasso (see Tibshirani [11]) uses the least squares criterion penalized by a L_1 penalty term. By this way, Lasso shrinks some coefficients and puts the others to zero. Thus, this last method performs variable selection but computationally, its implementation needs quadratic programming techniques.

Penalization is not the only way to perform variable or model selection. For example, we can cite the Subset Selection (see Hastie [7]) which provides, for each $k \in \{1, \ldots, n\}$, the best subset of size k, i.e. the subset of size k which gives smallest residual sum of squares. Then, by cross validation, the final subset is selected. This method is exhaustive, and so it is difficult to use in practice when p is large. Often, Forward or Backward Stepwise Selection (see Hastie [7]) are preferred since they are computationally efficient methods. But, they perhaps eliminate useful predictors. Since they are not exhaustive methods they may not reach the global optimal model. In the regression framework, there exists an efficient algorithm developped by Furnival and Wilson [5] which arrises the optimal model, for a small number of explanatory variables, without exploring all the models.

At present, the most promising method seems to be the method called Least Angle Regression (LARS) due to Efron *et al.* [4].
Let $\mu = x\beta$ where $x = (X_1^T, \ldots, X_n^T)$. LARS builds an estimate of μ by successive steps. It proceeds by adding, at each step, one covariate to the model, as Forward Selection.
At the begining, $\mu = \mu_0 = 0$. At 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 much correlation with the current residuals. So, 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 equiangulary strategy. For example, at the second step, LARS proceeds equiangulary between X^{j_1} and X^{j_2} until another explanatory variable enters.
This method is computationally efficient and gives good results in practice. However, a complete theoretical elucidation needs further investigation.

This paper proposes first a theoretical variable selection procedure for nonlinear models and gives also some practical indications.
The purpose is to propose, for regression and classification frameworks, a method consisting of applying the CART algorithm to each subset of variables. Then, considering model selection via penalization (cf. Birgé and Massart [2]), it selects the set which minimizes a penalized criterion. In the regression and classification frameworks, we determine via oracle bounds, the expressions of this penalized criterion.

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 $\mathbb{R}^p$ 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. $\Lambda = \{X^1, X^2, \ldots, X^p\}$. The explained variable Y is called the response.
Our purpose is to find a subset M of Λ, as small as possible, such that the variables in M enable to predict the response Y.

To achieve this objective, we split the sample $\mathcal{L}$ in three subsamples $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$ of size n_1, n_2 and n_3 respectively and we apply the CART algorithm to all the subsets of Λ. More precisely, for any $M \in \mathcal{P}(\Lambda)$, we build the maximal tree by the CART growing procedure using the subsample $\mathcal{L}_1$. This tree, denoted $T_{max}^{(M)}$, is constructed thanks to the class of admissible splits $\mathcal{S}p_M$ which involves only the variables of M.
Then, for any $M \in \mathcal{P}(\Lambda)$ and any $T \preceq T_{max}^{(M)}$, 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

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

which depends only on $\mathcal{L}_1$.
Then, for any (M, T), we denote $\hat{s}_{M,T}$ the $\mathcal{L}_2$ empirical risk minimizer on $S_{M,T}$.

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

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

$$\widehat{(M, T)} = \underset{(M,T)}{argmin} \{\gamma_{n_2}(\hat{s}_{M,T}) + pen(M, T)\}$$

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

Our purpose is to determine the penalty function *pen* such that the model $\widehat{(M, T)}$ is close to the optimal one, i.e.:

$$\mathbb{E}\left[l(s, \tilde{s}) \ |\mathcal{L}_1\right] \leq C \inf_{(M,T)} \left\{ \mathbb{E}\left[l(s, \hat{s}_{M,T}) \ |\mathcal{L}_1\right] \right\}, \quad C \text{ close to } 1$$

where l denotes the loss function.

The described procedure is, of course, a theoretical one since, when p is too large, it may be impossible, in practice, to take into account all the 2^p sets of variables. A solution consists of determining, at first, few data-driven subsets of variables which are adapted to perform variable selection and then applying our procedure to those subsets.
As this family of subsets, denoted $\mathcal{P}^*$, is constructed thanks to the data, the theoretical penalty, determined when the procedure involves the 2^p sets, is still adapted for the procedure restricted to $\mathcal{P}^*$.

The paper is organized as follows. After this introduction, the **Section 2** recalls the different steps of the CART algorithm and defines some notations. The **Sections 3** and **4** present the results obtain in the regression and classification frameworks. In the **Section 5**, we apply our procedure to a simulated example and we compare the results of the procedure when on the one hand we consider all sets of variables and on the other hand we take into account only a subset determined thanks to the Variable Importance. **Sections 6** and **7** collect lemmas and proofs.

2 Preliminaries

2.1 Overview of CART

In the regression and classification frameworks and thanks to a training set, CART splits recursively the observations space $\mathcal{X}$ and defines a piecewise constant function on this partition which is called a predictor or a classifier according to the case. CART proceeds in three steps: the construction of a maximal tree, the construction of nested models by pruning and a final model selection.

The first step consists of the construction of a nested sequence of partitions of the observations space using binary splits. Each split involves only one original explanatory variable and is determined by maximizing a quality criterion. A useful representation of this construction is a tree of maximal depth, called maximal tree.
The principle of the pruning step is to extract, from the maximal tree, a sequence of nested subtrees which minimize a penalized criterion. This penalized criterion realizes a tradeoff between the goodness of fit and the complexity of the tree or the model.
At last, via a test sample or cross validation, a subtree is selected among the preceding sequence.

The penalized criterion which appears in the pruning step was proposed by Breiman *et al.* [3]. It is composed of two parts:

- an empirical contrast which quantifies the goodness of fit,
- a penalty proportional to the complexity of the model which is measured by the number of leaves of the associated tree. So, if T denotes a tree and S_T the associated model, i.e. the linear subspace of $\mathbb{L}^2(\mathcal{X})$ composed of the piecewise constant functions defined on the leaves of T, the penalty is proportional to $|T|$, the number of leaves of T.

In the gaussian or bounded regression, Gey and Nédélec [6] proved some oracle inequalities for the well-known penalty term $\left(\frac{\alpha|T|}{n}\right)$. They consider two situations that we used too in this article:

- $(M1)$: 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 subsample $\mathcal{L}_1$ is used for the construction of the maximal tree, $\mathcal{L}_2$ for its pruning and $\mathcal{L}_3$ for 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.

Remark 2.1
The $(M1)$ situation is easier since all the subsamples are independent. But, often it is difficult to split the data in three parts because the number of data is too small. That is why we also consider the more realistic situation $(M2)$. □

CART is an algorithm which builds binary decision tree. A first idea is to perform variable selection by retaining the variables appearing in the tree. This has many drawbacks since on the one hand, the number of selected variables may be to large, and on the other hand, some really important variables could be hidden by the selected ones.
Another approach is based on the Variable Importance (VI) introduced by Breiman *et al.* [3]. This criterion, calculated with respect to a given tree (typically coming from the procedure CART), quantifies the contribution of each variable by awarding it a note between 0 and 100. The variable selection consists of keeping the variables whose notes are greater than an arbitrary threshold. But, there is, at present, no way to automatically determine the threshold and such a method does not allow to suppress highly dependent influent variables.

2.2 The context

The paper deals with two frameworks: the regression and the binary classification.
In both cases, the function s is defined by

$$s = \underset{u:\mathbb{R}^p \to \mathcal{Y}}{argmin}\ \mathbb{E}\left[\gamma(u,(X,Y))\right] \text{ with } \gamma(u,(x,y)) = (y-u(x))^2. \tag{1}$$

Since the distribution P is unknown, s is unknown too. Thus, in the regression and classification frameworks, we use $(X_1,Y_1),...,(X_n,Y_n)$, independent copies of (X,Y), to construct an estimator of s. The quality of this one is measured by the loss function l

$$l(s,u) = \mathbb{E}[\gamma(u,.)] - \mathbb{E}[\gamma(s,.)]. \tag{2}$$

In the regression case, the expression of s defined in (1) is

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

the loss function l given by (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 $(\varepsilon_1,...,\varepsilon_n)$ 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.

In the classification case, the Bayes classifier s, given by (1), is defined by:

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

As Y and the predictors u take their values in $\{0; 1\}$, we have

$$\begin{aligned} \gamma(u,(x,y)) &= \mathbb{1}_{u(x)\neq y}, \\ l(s,u) &= \mathbb{P}(Y \neq u(X)) - \mathbb{P}(Y \neq s(X)), \\ &= \mathbb{E}\left[|s(X) - u(X)||2\eta(X) - 1|\right]. \end{aligned}$$

3 Regression

Let us consider the regression framework where the ε_i are supposed to have exponential moments around 0 conditionally to X_i. As explained in [10], this assumption can be expressed by the existence of two constants $\sigma \in \mathbb{R}_+^*$ and $\rho \in \mathbb{R}_+$ such that

$$\text{for any } \lambda \in (-1/\rho, 1/\rho), \; log\mathbb{E}\left[e^{\lambda\varepsilon_i}\middle|X_i\right] \leq \frac{\sigma^2\lambda^2}{2\left(1-\rho|\lambda|\right)} \tag{3}$$

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

Remark 3.1
If $\rho = 0$ in (3), the random variables ε_i are said to be sub-gaussian conditionally to X_i. □

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.

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

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 subsample $\mathcal{L}_2$ to select a model $\widehat{(M,T)}$ which is close to the optimal one. To do this, we minimize a penalized criterion

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

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 suppose that $\|s\|_\infty \leq R$, *with* R *a positive constant.*
Let consider a penalty function of the form:
$\forall\ M \in \mathcal{P}(\Lambda)$ *and* $\forall\ T \preceq T_{max}^{(M)}$

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

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

$$\begin{aligned}\mathbb{E}\left[\|s - \tilde{s}\|_{n_2}^2 \ |\mathcal{L}_1\right] \quad \leq \quad & C_1 \inf_{(M,T)}\left\{\inf_{u \in S_{M,T}}\|s-u\|_\mu^2 + pen(M,T)\right\} + C_2\frac{(\sigma^2+\rho R)}{n_2} \\ & + C(\rho,\sigma,R)\frac{\mathbb{1}_{\rho\neq 0}}{n_2 log(n_2)}\end{aligned}$$

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. □

The penalty function is the sum of two terms. The first one $\alpha\left(\sigma^2 + \rho R\right)\frac{|T|}{n_2}$ is the penalty proposed by Breiman *et al.* [3] in their pruning algorithm and validated by Gey and Nédélec [6] for the Gaussian regression case. This proposition validates the CART pruning penalty proposed by Breiman *et al.* [3] in a more general regression framework than the Gaussian one. The second one is due to the variable selection. It penalizes models that are based on too much explanatory variables.

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

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

$$\hat{T}_M = \underset{T \preceq T_{max}^{(M)}}{argmin}\left\{\gamma_{n_2}(\hat{s}_{M,T}) + \alpha\left(\sigma^2 + \rho R\right)\frac{|T|}{n_2}\right\}.$$

- Then we choose a set $\hat{M}$ by

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

Remark 3.2
If $\rho = 0$, the form of the penalty is

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

the oracle bound is

$$\mathbb{E}\left[\|s-\tilde{s}\|_{n_2}^2 \ \ |\mathcal{L}_1\right] \leq C_1 \inf_{(M,T)} \left\{ \inf_{u\in S_{M,T}} \|s-u\|_\mu^2 + pen(M,T) \right\} + C_2 \frac{\sigma^2}{n_2},$$

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

$$\alpha_0 = 2(1+3log2) \ \text{ and } \ \beta_0 = 3.$$

□

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 may not always have enough observations to split the training sample $\mathcal{L}$ in three subsamples. This is the reason why we study now the $(M2)$ situation.

• $(M2)$ case :

In this situation, the same subsample $\mathcal{L}_1$ 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 3.2
Let suppose that $\|s\|_\infty \leq R$, with R a positive constant.
Let consider a penalty function of the form:
$\forall\ M \in \mathcal{P}(\Lambda)$ *and* $\forall\ T \preceq T_{max}^{(M)}$

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

If $p \leq log(n_1)$, $\alpha > \alpha_0$ and $\beta > \beta_0$,
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* $\geq 1 - e^{-\xi}\Sigma - \frac{c}{n_1 log(n_1)} 1\!\!1_{\rho \neq 0}$,

$$\|s-\tilde{s}\|_{n_1}^2 \ \leq \ C_1 \inf_{(M,T)} \left\{ \inf_{u\in S_{M,T}} \|s-u\|_{n_1}^2 + pen(M,T) \right\} + \frac{C_2}{n_1} \left(\left(1 + \frac{\rho^4}{\sigma^4} log^2 \left(\frac{n_1}{p} \right) \right) \sigma^2 + \rho R \right) \xi$$

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 3.3
If $\rho = 0$, the form of the penalty is

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

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

$$\|\tilde{s} - s\|_{n_1}^2 \leq C_1 \underset{(M,T)}{inf}\left\{\underset{u \in S_{M,T}}{inf}\|s - u\|_{n_1}^2 + pen(M,T)\right\} + C_2\frac{\sigma^2}{n_1}\xi$$

and the assumptions on $\|s\|_\infty$ and p are no longer useful. Moreover, if we look at the proof more closely, we see that we can take $\alpha_0 = \beta_0 = 3$.

□

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 $\hat{M}$.
Now, we choose the final predictor using test sample and we deduce the corresponding set of selected variables.

3.2 Final selection

Now, we have 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 K of $\mathcal{G}$ is finite and

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

where K_M is the number of subtrees of $T_{max}^{(M)}$ obtained by the pruning algorithm defined by Breiman *et al.* [3].

Given the subsample $\mathcal{L}_3$, we choose the final estimator $\tilde{\tilde{s}}$ by minimizing the empirical contrast γ_{n_3} on $\mathcal{G}$.

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

The next result validates this selection.

Proposition 3.3

- *In the* $(M1)$ *situation, taking* $p \leq logn_2$ *and* $N_{min} \geq 4\frac{\sigma^2+\rho R}{R^2} logn_2$, *we have: for any* $\xi > 0$, *with probability* $\geq 1 - e^{-\xi} - 1\!\!1_{\rho\neq 0}\frac{R^2}{2(\sigma^2+\rho R)}\frac{1}{n_2^{1-log2}}$,

  $\forall \eta \in (0,1)$,

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

- *In the* $(M2)$ *situation, denoting* $\epsilon(n_1) = 2\,1\!\!1_{\rho\neq 0} n_1 exp\left(-\frac{9\rho^2 log^2 n_1}{2(\sigma^2+3\rho^2 logn_1)}\right)$, *we have: for any* $\xi > 0$, *with probability* $\geq 1 - e^{-\xi} - \epsilon(n_1)$,

  $\forall \eta \in (0,1)$,

$$\begin{aligned}\left\|s - \tilde{\tilde{s}}\right\|_{n_3}^2 \quad \leq \quad & \frac{(1+\eta^{-1}-\eta)}{\eta^2} \underset{\tilde{s}(\alpha,\beta)\in\mathcal{G}}{inf} \left\|s - \tilde{s}(\alpha,\beta)\right\|_{n_3}^2 \\ & + \frac{1}{\eta^2}\left(\frac{2}{1-\eta}\sigma^2 + 4\rho R + 12\rho^2 logn_1\right)\frac{(2logK+\xi)}{n_3}.\end{aligned}$$

□

Remark 3.4

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

$$\begin{aligned}\mathbb{E}\left[\left\|s - \tilde{\tilde{s}}\right\|_{n_3}^2 \ \middle|\mathcal{L}_1,\ \mathcal{L}_2\right] \quad \leq \quad & \frac{1+\eta^{-1}-\eta}{\eta^2} \underset{\tilde{s}(\alpha,\beta)\in\mathcal{G}}{inf} \left\{\mathbb{E}\left[\left\|s - \tilde{s}(\alpha,\beta)\right\|_{n_3}^2 \ \middle|\mathcal{L}_1,\ \mathcal{L}_2\right]\right\} \\ & + \frac{2}{\eta^2(1-\eta)}\frac{\sigma^2}{n_3}\left(2logK+1\right).\end{aligned}$$

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 alterate so much the accuracy of the final estimator. Now we can conclude that theoretically our procedure is valid. □

4 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 [12]:

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

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

• $(M1)$ case :

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

Proposition 4.1
Let suppose the existence of $h > 0$ such that:

$$\forall x \in \mathbb{R}^p, \quad |2\eta(x) - 1| \geq h$$

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

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

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:

$$\mathbb{E}\left[l\left(s, \tilde{s}\right) | \mathcal{L}_1\right] \leq C_1 \inf_{(M,T)} \left\{ l\left(s, S_{M,T}\right) + pen\left(M,T\right) \right\} + C_2 \frac{1}{n_2 h}$$

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

Like in the regression case, for a given value of $|M|$, the penalty is proportional to $\frac{|T|}{n_2}$. This validates the CART pruning algorithm in the binary classification framework.
Unfortunately, the multiplicative factor of $\frac{|T|}{n_2}$ depends on the margin h which is difficult to estimate.

A main difference between regression and classification is that, in the first case, we overestimate the expectation of the empirical loss, whereas in classification we control the real risk.

• $(M2)$ case :

Like in the regression case, we manage to extend our result for only one subsample $\mathcal{L}_1$. 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:

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

Proposition 4.2
Let suppose the existence of $h > 0$ such that:

$$\forall x \in \mathbb{R}^p, \quad |2\eta(x) - 1| \geq h$$

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

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

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 $\geq 1 - e^{-\xi}\Sigma^2$:

$$l_1(s, \tilde{s}) \leq C_1 \inf_{(M,T)} \left\{ l_1\left(s, S_{M,T}\right) + pen_n(M,T) \right\} + \frac{C_2}{n_1 h}(1 + \xi)$$

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

Like in the regression case, when we consider the $(M2)$ situation instead of the $(M1)$ one, we obtain only an inequality with high probability instead of a result in expectation.

4.2 Final selection

With the same notations as in the **Subsection 3.2**, we validate the final selection for the two methods.
The following proposition is expressed for the $(M1)$ method.

Proposition 4.3
For any $\eta \in (0,1)$, we have:

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

□

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

Unlike the regression case, for the $(M1)$ method in the classification framework, since the results in expectation of the **Propositions 4.1** and **4.3** involve the same expected loss, we can compare the final estimator $\tilde{\tilde{s}}$ with the entire collection of models:

$$\mathbb{E}\Big[l(s,\tilde{\tilde{s}})\ |\mathcal{L}_1,\mathcal{L}_2\Big] \leq \tilde{C}_1 \underset{(M,T)}{inf} \Big\{l(s,S_{M,T}) + pen(M,T)\Big\} + \frac{C_2}{n_2 h} + \frac{C_3}{n_3 h}\Big(1 + log(K)\Big).$$

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.

The simulated example, also used by Breiman *et al.* (see [3] p. 237), is composed of $p = 10$ explanatory variables $X^1, \ldots, X^{10}$ such that:

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

and of the explained variable Y given by:

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

where the unobservable random variable ε is independent of $X^1, \ldots, X^{10}$ 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 1** 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.

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 1: Variable Importance Ranking for the considered simulated example.

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.

To make in use our procedure, we consider a training sample $\mathcal{L}$ 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 behavior 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 behavior of the set associated with the estimator $\tilde{s}$, when we apply the theoretical procedure, is summarized by the **Table 2**.
At the intersection of the row β and the column α appears the set of variables associated with $\tilde{s}$.

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}

Table 2: 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*.

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 **Table 1**).
As the expected final set $\{1, 2, 5, 3, 6, 4, 7\}$ appears in the **Table 2**, obviously, the final step of the procedure, whose results are given by the **Table 3**, returns the "good" set.

$\hat{\alpha}$	$\hat{\beta}$	selected set
0.3	$\rightarrow 100$	$\{1, 2, 3, 4, 5, 6, 7\}$

Table 3: In this table, we see the results of the final model selection.

The **Table 3** provides some other informations.
At present, we do not know how to choose the parameters α and β of the penalty function. This is the reason why the theoretical procedure includes a final selection by test sample. But, if we are able to determine, thanks to the data, the value of those parameters, this final step would disappear.
If we analyse the **Table 3**, we see that the "best" parameter $\hat{\alpha}$ takes only one value and that $\hat{\beta}$ belongs to a "small" range. So, those 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.

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 [9] 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:

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

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. The second one is that, if we perform our procedure restricted to the family $\mathcal{P}^*$, we obtain nearly the same results for the behavior of the set associated with $\tilde{s}$. The only difference is that, since $\mathcal{P}^*$ does not contain the set of size 10, in the **Table 2**, the set $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ is replaced by $\{1, 2, 5, 6, 3, 7, 4\}$.

6 Appendix

This section presents some lemmas which are useful in the proofs of the propositions of the **Sections 3** and **4**. The lemmas 6.1 to 6.4 are known results. We just give the statements and references for the proofs. 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 $\xi_1, ..., \xi_n$ 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

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

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 of the lemma 6.1: see Massart [1] □

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:
$\forall \Phi : \mathbb{R} \to \mathbb{R}^+$ *such that :*

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

we have:
$\forall x \geq \sigma_*$

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

□

PROOF OF THE LEMMA 6.2: see Massart and Nédélec [6], section: "Appendix: Maximal inequalities", lemma 5.5. □

Thanks to the lemma 6.3, we see that the Hold-Out is an adaptative 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 $X' = (X'_1, ..., X'_N)$ are used to build some preliminary collection of estimators $(\hat{s}_m)_{m\in\mathcal{M}}$ and we use the remaining observations $X_1, ..., X_n$ 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 cardinality K.
If there exists a function w such that:

$$\begin{aligned} &- \quad w \quad : \quad \mathbb{R}^+ \to \mathbb{R}^+, \\ &- \quad x \to \frac{w(x)}{x} \quad \textit{is nonincreasing}, \\ &- \quad \forall \epsilon > 0, \quad \sup_{l(s,t)\leq\epsilon^2} Var_P\left(\gamma(t,.) - \gamma(s,.)\right) \leq w^2(\epsilon) \end{aligned}$$

Then, $\forall \theta \in (0,1)$, 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 OF THE LEMMA 6.3: see [8], Chapter: "Statistical Learning", Section: "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. In the first lemma, we consider only partitions m of $\{1,\ldots,n\}$ 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 $\{1,\ldots,n\}$.

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

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

Let m_0 a partition of $\{1,\ldots,n\}$ such that, $\forall J \in m_0$, $|J| \geq N_{min}$.
We consider the collection $\mathcal{M}$ of all partitions of $\{1,\ldots,n\}$ constructed from m_0 and the statistics

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

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 $x > 0$,

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

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 of the lemma 6.4: see Sauvé [10], lemma 1.

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

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

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

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

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 $x > 0$,

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

and

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

□

Proof of the lemma 6.5: The proof is exactly the same as the preceding one. The only difference is that the set Ω_δ is smaller. □

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 $a > 2log(2)$ and $b > 1$ two absolute constants, satisfy

$$\sum_{M\in\mathcal{P}(\Lambda)} \sum_{T\preceq T^{(M)}_{max}} e^{-x_{M,T}} \leq \Sigma(a,b) \tag{4}$$

with $\Sigma(a,b) = -log\left(1 - e^{-(a-2log2)}\right)\frac{e^{-(b-1)}}{1-e^{-(b-1)}} \in \mathbb{R}^*_+$. □

Proof of the lemma 6.6:
We are looking for weights $x_{M,T}$ such that the sum

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

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

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

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

$$\Sigma(\mathcal{L}_1) \quad \leq \quad \sum_{k=1}^{p} \left(\frac{ep}{k} \right)^k \sum_{D \geq 1} \frac{1}{D} e^{-(x(k,D) - (2log2)D)}.$$

Taking $x(k,D) = aD + bk\left(1 + log\left(\frac{p}{k}\right)\right)$ with $a > 2log2$ and $b > 1$ two absolute constants, we have

$$\Sigma(\mathcal{L}_1) \leq \left(\sum_{k \geq 1} e^{-(b-1)k} \right) \left(\sum_{D \geq 1} \frac{1}{D} e^{-(a-(2log2))D} \right) = \Sigma(a,b).$$

Thus the weights $x_{M,T} = a|T| + b|M| \left(1 + log\left(\frac{p}{|M|}\right)\right)$, with $a > 2log(2)$ and $b > 1$ two absolute constants, satisfy (4). □

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*

$$\sum_{M \in \mathcal{P}(\Lambda)} \sum_{T \in \mathcal{M}_{n_1,M}} e^{-x_{M,T}} \leq \Sigma^{'}(a,b) \tag{5}$$

with $\Sigma^{'}(a,b) = \frac{e^{-a}}{1-e^{-a}} \frac{e^{-(b-1)}}{1-e^{-(b-1)}}$ *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 of the lemma 6.7:
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 $(X_1, Y_1), ..., (X_n, Y_n)$ be n independent observations taking their values in some measurable space $\Theta \times \{0, 1\}$, with common distribution P.
Let $S_T = \{$piecewise constant functions, defined on $\tilde{T}\}$, with T a tree.
Let suppose that:

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

Then:

- $\sup_{u \in S_T,\ ls, u \leq \varepsilon^2} d(s, u) \leq w(\varepsilon)$ *with* $w(x) = \frac{1}{\sqrt{h}} x$,
- $\exists \phi_T : \mathbb{R}^+ \to \mathbb{R}^+$ *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 $\phi_T(w(x)) = \sqrt{n}x^2$.
- $\sigma_T^2 \leq \frac{K_3^2 |T|}{nh}$.

□

Thanks to lemma (6.8) and (6.2), we deduce the next one.

Lemma 6.9
Let $(X_1, Y_1), ..., (X_n, Y_n)$ 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 lemme 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

7.1 Regression

Proof of the proposition 3.1:
Let $a > 2log2$, $b > 1$, $\theta \in (0, 1)$ and $K > 2 - \theta$ four constants.
Let us denote

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

Following the proof of theorem 2 in [2], we get

$$(1 - \theta)\|s - \tilde{s}\|_{n_2}^2 = \Delta_{\widehat{M,T}} + \underset{(M,T)}{inf}\ R_{M,T} \tag{6}$$

where

$$\begin{aligned}\Delta_{M,T} &= (2-\theta)\|\varepsilon_{M,T}\|_{n_2}^2 - 2 < \varepsilon, s - s_{M,T} >_{n_2} -\theta\|s - s_{M,T}\|_{n_2}^2 - pen(M,T)\\ R_{M,T} &= \|s - s_{M,T}\|_{n_2}^2 - \|\varepsilon_{M,T}\|_{n_2}^2 + 2 < \varepsilon, s - s_{M,T} >_{n_2} + pen(M,T)\end{aligned}$$

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

For any M, we denote

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

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

$$\begin{aligned}\mathbb{P}\Bigg(\|\varepsilon_{M,T}\|_{n_2}^2 \mathbb{1}_{\Omega_{\delta,M}} &\geq \frac{\sigma^2}{n_2}|T| + 8\frac{\sigma^2}{n_2}\sqrt{2|T|x} + 4\frac{\sigma^2}{n_2}x \ \Big| \mathcal{L}_1 \text{ and } \{X_i;\ (X_i,Y_i) \in \mathcal{L}_2\} \Bigg)\\ &\leq e^{-x}\end{aligned} \tag{7}$$

and

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

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

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

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

$$\begin{aligned}\mathbb{P}\Big(- < \varepsilon, s - s_{M,T} >_{n_2} &\geq \frac{\sigma}{\sqrt{n_2}}\|s - s_{M,T}\|_{n_2}\sqrt{2x} + \frac{2\rho R}{n_2}x \ \Big| \mathcal{L}_1 \text{ and } \{X_i;\ (X_i,Y_i) \in \mathcal{L}_2\} \Big)\\ &\leq e^{-x}\end{aligned} \tag{8}$$

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) and (8) with respect to (M,T), we derive a set E_ξ such that

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

- 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 $\Sigma(a,b) = -log\left(1-e^{-(a-2log2)}\right)\frac{e^{-(b-1)}}{1-e^{-(b-1)}}$.

Using the inequalities $2\frac{\sigma}{\sqrt{n_2}}\|s-s_{M,T}\|_{n_2}\sqrt{2(x_{M,T}+\xi)} \leq \theta\|s-s_{M,T}\|_{n_2}^2 + \frac{2}{\theta}\frac{\sigma^2}{n_2}(x_{M,T}+\xi)$ and $2\sqrt{|T|(x_{M,T}+\xi)} \leq \eta|T| + \eta^{-1}(x_{M,T}+\xi)$ 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.

$$pen(M,T) \geq K\frac{\sigma^2}{n_2}|T| + \left[4(2-\theta)\left(1+\frac{8(2-\theta)}{K+\theta-2}\right)+\frac{2}{\theta}+4\frac{\rho}{\sigma^2}R\right]\frac{\sigma^2}{n_2}x_{M,T} \tag{9}$$

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

$$\mathbb{E}\left[\Delta_{\widehat{M,T}}\mathbb{1}_\Omega\middle|\mathcal{L}_1\right] \leq 2\left(4(2-\theta)\left(1+\frac{8(2-\theta)}{K+\theta-2}\right)+\frac{2}{\theta}+4\frac{\rho}{\sigma^2}R\right)\frac{\sigma^2}{n_2}\Sigma(a,b) \tag{10}$$

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 (8) from assumption (A), we get that for any (M,T) and any $x>0$

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

Thus we derive a set F_ξ such that

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

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

It follows from definition of $R_{M,T}$ and inequality (9) on the penalty that

$$\begin{aligned}\mathbb{E}\left[\inf_{(M,T)} R_{M,T}\mathbb{1}_\Omega\middle|\mathcal{L}_1\right] \leq\ & 2\inf_{(M,T)}\left\{\mathbb{E}\left[\|s-s_{M,T}\|_{n_2}^2\middle|\mathcal{L}_1\right]+pen(M,T)\right\}\\ &+\left(2+4\frac{\rho}{\sigma^2}R\right)\frac{\sigma^2}{n_2}\Sigma(a,b)\end{aligned} \tag{11}$$

We conclude from (6), (10) and (11) that

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

It remains to control $\mathbb{E}\left[\|s-\tilde{s}\|_{n_2}^2 1\!\!1_{\Omega^c}\middle|\mathcal{L}_1\right]$, except if $\rho=0$ in which case it is finished.
After some calculations (see the proof of theorem 1 in [10] for more details), we get

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

and

$$\mathbb{E}\left[\|\varepsilon_{M,T_{max}^{(M)}}\|_{n_2}^4\middle|\mathcal{L}_1\right] \quad \leq \quad \frac{\sigma^4}{N_{min}^2}+\frac{C^2(\rho,\sigma)}{n_2N_{min}^2}+\frac{3\sigma^4}{n_2N_{min}}$$

where $C^2(\rho,\sigma)$ is a constant which overestimates $\mathbb{E}\left[\varepsilon_i^4\right]$.
Thus we have

$$\mathbb{E}\left[\|s-\tilde{s}\|_{n_2}^2 1\!\!1_{\Omega^c}\middle|\mathcal{L}_1\right] \quad \leq \quad R^2\mathbb{P}\left(\Omega^c\middle|\mathcal{L}_1\right)+2^p\left(\frac{\sigma^2}{N_{min}}+\frac{C(\rho,\sigma)}{\sqrt{n_2}N_{min}}+\frac{\sqrt{3}\sigma^2}{\sqrt{n_2N_{min}}}\right)\sqrt{\mathbb{P}\left(\Omega^c\middle|\mathcal{L}_1\right)}$$

Let us recall that

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

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

- $2^p\sqrt{\mathbb{P}\left(\Omega_\delta^c\middle|\mathcal{L}_1\right)}\leq\frac{\sigma}{\sqrt{12}\rho}\frac{1}{n_2\sqrt{log(n_2)}}$
- $\mathbb{P}\left(\Omega_\delta^c\,\middle|\mathcal{L}_1\right)\leq\frac{\sigma^2}{12\rho^2}\frac{1}{n_2^4log(n_2)}$
- $\frac{\sigma^2}{N_{min}}+\frac{C(\rho,\sigma)}{\sqrt{n_2}N_{min}}+\frac{\sqrt{3}\sigma^2}{\sqrt{n_2N_{min}}}\leq\frac{\sigma^3}{4\rho}\left(\frac{\sigma}{6\rho}+\frac{C(\rho,\sigma)}{6\sigma\rho}+\sqrt{2}\right)\frac{1}{log(n_2)}$

It follows that

$$\mathbb{E}\left[\|s-\tilde{s}\|_{n_2}^2 1\!\!1_{\Omega^c}\middle|\mathcal{L}_1\right] \quad \leq \quad \frac{\sigma^2}{4\rho^2}\left[\frac{R^2}{3}+\frac{\sigma^2}{2\sqrt{3}}\left(\frac{\sigma}{6\rho}+\frac{C(\rho,\sigma)}{6\sigma\rho}+\sqrt{2}\right)\right]\frac{1}{n_2log(n_2)}$$

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 $\forall\ M\in\mathcal{P}(\Lambda)\ \forall\ T\preceq T_{max}^{(M)}$

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

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

$$\begin{aligned}(1-\theta)\mathbb{E}\left[\|s-\tilde{s}\|_{n_2}^2 \,|\mathcal{L}_1\right] &\leq 2 \inf_{(M,T)} \left\{ \inf_{u \in S_{M,T}} \|s-u\|_\mu^2 + pen(M,T) \right\} \\ &\quad + \left(2\Upsilon + 2 + 12\frac{\rho}{\sigma^2}R\right)\frac{\sigma^2}{n_2}\Sigma(a,b) \\ &\quad + \frac{\sigma^2}{4\rho^2}\left[\frac{R^2}{3} + \frac{\sigma^2}{2\sqrt{3}}\left(\frac{\sigma}{6\rho} + \frac{C(\rho,\sigma)}{6\sigma\rho} + \sqrt{2}\right)\right]\frac{1}{n_2 log(n_2)}\end{aligned}$$

We deduce the proposition by taking $K=2$, $\theta \to 1$, $a \to 2log2$ and $b \to 1$. □

Proof of the proposition 3.2:
Let $a>0$, $b>1$, $\theta \in (0,1)$ and $K > 2-\theta$ four constants.

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 instead of lemma 6.4.

To prove the proposition, we follow the same steps as before. The main difference is that, the quantities are now conditionned by $\{X_i;\ (X_i,Y_i) \in \mathcal{L}_1\}$ instead of $\mathcal{L}_1$ and $\{X_i;\ (X_i,Y_i) \in \mathcal{L}_2\}$.

Proof of the proposition 3.3:

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

$$\left\|s-\tilde{\tilde{s}}\right\|_{n_3}^2 \leq \|s-\tilde{s}(\alpha,\beta)\|_{n_3}^2 + 2\left\langle \varepsilon, \tilde{\tilde{s}} - \tilde{s}(\alpha,\beta)\right\rangle_{n_3} \tag{12}$$

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

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

Setting $x = 2logK + \xi$ 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

- $\mathbb{P}\left(E_\xi^c \,|\mathcal{L}_1,\ \mathcal{L}_2,\ \text{and } \{X_i;\ (X_i,Y_i) \in \mathcal{L}_3\}\right) \leq e^{-\xi}$

- on the set E_ξ, for any $\tilde{s}(\alpha,\beta)$ and $\tilde{s}(\alpha',\beta') \in \mathcal{G}$

$$\begin{aligned}\langle \varepsilon, \tilde{s}(\alpha',\beta') - \tilde{s}(\alpha,\beta)\rangle_{n_3} &\leq \frac{\sigma}{\sqrt{n_3}}\|\tilde{s}(\alpha',\beta') - \tilde{s}(\alpha,\beta)\|_{n_3}\sqrt{2(2logK+\xi)} \\ &\quad + M_{\alpha,\beta,\alpha',\beta'}\frac{\rho}{n_3}(2logK+\xi)\end{aligned}$$

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

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

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

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

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

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

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 $\tilde{s}(\alpha,\beta) \in \mathcal{G}$

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

It follows from (12) that, on the set $\Omega_1 \bigcap E_\xi$, for any $\tilde{s}(\alpha,\beta) \in \mathcal{G}$ and any $\eta \in (0;1)$

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

Taking $p \leq logn_2$ and $N_{min} \geq 4 \frac{\sigma^2 + \rho R}{R^2} logn_2$, we have

$$\mathbb{P}\left(\Omega_1^c \right) \leq \frac{R^2}{2(\sigma^2 + \rho R)} \frac{1}{n_2^{1 - log2}}$$

Finally, in the $(M1)$ situation, we have
for any $\xi > 0$, with probability $\geq 1 - e^{-\xi} - \frac{R^2}{2(\sigma^2+\rho R)} \frac{1}{n_2^{1-log2}}$,
$\forall \eta \in (0,1)$,

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

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

$$\Omega_2 = \{ \forall 1 \leq i \leq n_1 \left| \varepsilon_i \right| \leq 3\rho logn_1 \}$$

Thanks to assumption 3, we get that

$$\mathbb{P}\left(\Omega_2^c \middle|\ \{X_i;\ (X_i,Y_i) \in \mathcal{L}_1\} \right) \quad \leq \quad 2n_1 exp\left(-\frac{9\rho^2 log^2 n_1}{2(\sigma^2 + 3\rho^2 logn_1)} \right)$$

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 $\tilde{s}(\alpha,\beta) \in \mathcal{G}$

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

It follows from (12) that, on the set $\Omega_2 \bigcap E_\xi$, for any $\tilde{s}(\alpha,\beta) \in \mathcal{G}$ and any $\eta \in (0;1)$

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

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

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

□

7.2 Classification

Proof of the proposition 4.1:
Let M $\in \mathcal{P}(\Lambda)$, $T \preceq T_{max}^{(M)}$ and $s_{M,T} \in S_{M,T}$. We let

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

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

Following the proof of theorem 4.2 in [1], we get

$$l(s,\tilde{s}) \leq l(s,s_{M,T}) + w_{\widehat{M,T}}(\tilde{s}) \times V_{\widehat{M,T}} + pen(M,T) - pen(\widehat{M,T}) \tag{13}$$

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 $x > 0$

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

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:

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

$$V_{M',T'} \leq K_1 \mathbb{E}\left[V_{M',T'}\right] + K_2 \left(\sqrt{\frac{x_{M',T'}+\xi}{2n_2}} {y_{M',T'}}^{-1} + \frac{x_{M',T'}+\xi}{n_2} {y_{M',T'}}^{-2} \right) \tag{14}$$

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, 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 (14), 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 $K \geq \frac{1}{48K_1 h}$, 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 (13), we get

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

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 $\Omega_{\xi,(M,T)}$

$$l(s,\tilde{s}) \leq C_1\Big(l(s,s_{M,T}) + pen(M,T)\Big) + \frac{C}{n_2 h}\xi$$

Integrating with respect to ξ and by minimizing , we get

$$\mathbb{E}\Big[l(s,\tilde{s})|\mathcal{L}_1\Big] \leq C_1 \inf_{M,T}\Big\{l(s,S_{M,T}) + pen(M,T)\Big\} + \frac{C}{n_2 h}\Sigma(a,b)$$

The two constants α_0 and β_0, which appear in the proposition 4.1, are defined by

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

□

Proof of the proposition 4.2:
This proof is quite simlilar to the preceeding one. We just need to replace $w_{M',T'}(u)$ and $V_{M',T'}$ by

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

And like the proof (3.2), we change the conditionnement.

Proof of the proposition 4.3:
This result is obtained by a direct application of the lemma 6.3 which appears in the subsection 6

□

References

[1] L. Birgé and P. Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268, (2001).

[2] L. Birgé and P. Massart. Minimal penalties for gaussian model selection. To be published in Probability Theory and Related Fields, (2005).

[3] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification And Regression Trees*. Chapman et Hall, (1984).

[4] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, (2004).

[5] G.M. Furnival and R.W. Wilson. Regression by leaps and bounds. *Technometrics*, 16:499–511, (1974).

[6] S. Gey and E. Nédélec. Model Selection for CART Regression Trees. *IEEE Trans. Inf. Theory*, 51(2):658–670, (2005).

[7] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, (2001).

[8] P. Massart. Notes de Saint-Flour. Lecture Notes to be published, (2003).

[9] J.M. Poggi and C. Tuleau. Classification supervisée en grande dimension. Application à l'agrément de conduite automobile. *Preprint Université Paris XI Orsay*, pages 1–16, (2005).

[10] M. Sauvé. Histogram selection in non gaussian regression. *Rapport de recherche INRIA*, (2006).

[11] R. Tibshirani. Regression shrinkage and selection via Lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–288, (1996).

[12] A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32(1):135–166, (2004).

Unité de recherche INRIA Futurs
Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399