ORSAY
N^o D'ORDRE : 8106

UNIVERSITÉ PARIS XI
U.F.R. SCIENTIFIQUE D'ORSAY

THÈSE

présentée pour obtenir le grade de

DOCTEUR EN SCIENCES
DE L'UNIVERSITÉ PARIS XI ORSAY
SPÉCIALITÉ : MATHÉMATIQUES

par

Christine TULEAU

SELECTION DE VARIABLES POUR LA DISCRIMINATION EN GRANDE DIMENSION ET CLASSIFICATION DE DONNEES FONCTIONNELLES

Rapporteurs : M. Philippe BESSE
M. Gérard BIAU

Soutenue le 5 décembre 2005 devant le jury composé de :

M.	Philippe BESSE	Rapporteur
M.	Gérard BIAU	Rapporteur
M.	Jean-Jacques DAUDIN	Examinateur
M.	Pascal MASSART	Président
M.	Jean-Michel POGGI	Directeur de Thèse
Mme	Nadine ANSALDI	Invitée

Remerciements

Tout d'abord, je souhaite sincèrement remercier la personne sans qui cette thèse n'aurait jamais atteint son terme. Jean-Michel, par ton soutien, tu m'as donné le courage d'achever ces trois années de recherche, notamment en me remotivant lorsque j'en éprouvais le besoin. Autour d'un café, tu m'as fait partager tes nombreuses connaissances, tu m'as accompagnée dans mes recherches tout en me permettant de m'épanouir en m'accordant une grande autonomie. En dépit de quelques tensions, j'ai beaucoup appris à ton contact, tant dans le domaine scientifique qu'humain.

Je tiens également à remercier chaleureusement la personne qui, d'une certaine façon, a donné naissance à cette thèse. Pascal, après m'avoir transmis ton savoir statistique et initier à la recherche lors de tes cours de maîtrise et de DEA, tu m'as encouragée à poursuivre dans cette voie et à entreprendre ce doctorat. Ton incroyable capacité à décrypter les relations humaines t'as alors conduit à me mettre en relation avec Jean-Michel, mais également Marie et Magalie. Par ailleurs, ta confiance, ta présence et ton aide inestimable ont contribué à l'aboutissement de ce travail, riche en collaborations fructueuses.

Marie, j'ai pris un grand plaisir à travailler à tes côtés au cours de ces deux dernières années. Ton sens de la précision a beaucoup apporté à notre collaboration, de même que nos nombreuses discussions dans la salle de thé ou dans un train.
Magalie, ton enthousiasme, ton savoir, ta ténacité et ta joie de vivre, tous communicatifs, ont amené rigueur et efficacité à notre travail commun.
J'espère que ces deux collaborations pourront se poursuivre au cours des mois à venir.

Merci à Philippe Besse et Gérard Biau qui ont accepté d'être les rapporteurs de cette thèse, ainsi qu'à l'ensemble du jury.

Je remercie la Direction de la Recherche de Renault pour leur collaboration active, notamment en mettant à ma disposition les données nécessaires à l'étude. Nadine et François, nos discussions ont été très constructives.

Tous mes remerciements à l'Équipe de Probabilité et Statistique de l'Université Paris-Sud Orsay pour son accueil et sa gentillesse. En particulier à Nathalie Cheze, Jean-Michel Loubes, Vincent Rivoirard et Marie-Luce Taupin pour leurs conseils avisés et à Liliane Bel, Jean Coursol Patrick Jakubowicz et Yves Misiti pour leur sympathie, leur disponibilité et leur aide dans le domaine informatique.

Mais aussi à toutes les personnes avec lesquelles j'ai eu l'occasion de discourir de mes travaux scientifiques, tant dans le cadre de discussions informelles que de séminaires : Laurent Rouvière, Gilles Celeux et tous les membres du groupe de travail INAPG-SELECT.

Merci aux doctorants actuels et passés d'Orsay pour nos discussions parfois animées, nos repas conviviaux, votre soutien et vos conseils. Un clin d'oeil tout particulier à mon bureau d'accueil, le bureau 112 dans lequel la bonne humeur était le maître mot, et à Laurent dont la présence et l'aide ont été précieuses tout au long de notre cursus universitaire commun.

Je ne peux oublier la formidable équipe du département Statistique et Traitement Informatique des Données de l'IUT de Paris 5. Elle m'a réservé un charmant accueil et m'a épaulée tout au long de mes trois années de monitorat. De même, merci à l'équipe du département Sciences Économiques, Gestion, Mathématique et Informatique de l'Université Paris X - Nanterre qui m'a accueillie en tant qu'ATER.

Parce que votre amitié à été une source permanente de réconfort, merci à Stéphanie, Sandrine, Marina, Géraldine, Nicolas, Romain, Christopher, Manu et Sauveur.

Merci à ma famille et belle-famille pour tous les instants partagés.
Laurent, Jocelyne et Christian, votre indéfectible soutien et votre amour m'ont portée durant toute cette thèse. Je vous dédie ce travail, ainsi qu'à mon mari Yannick qui a su me supporter et m'accompagner dans les moments de joie et surtout de doute.

Les Chats

*Au commencement, Dieu créa le chat à son image.
Et bien entendu, il trouva que c'était bien. Et c'était bien, d'ailleurs.
Du moins sur le plan de l'esthétisme.
Mais le chat ne voulait rien faire, n'avait rien envie de faire.
Il était paresseux, renfermé, taciturne, économe de ses gestes et, de plus,
extrêmement buté.
C'est alors que Dieu eut l'idée de créer l'homme.
Uniquement dans le but de servir le chat, de lui servir d'esclave
jusqu'à la fin des temps.
Au chat, il avait donné l'indolence, la sagesse, la lucidité,
l'art de faire son temps le plus agréablement possible
en s'économisant le plus possible. A l'homme, il inocula la névrose
de l'agitation, la passion du travail même le plus ingrat,
l'ambition qui allait le pousser à édifier toute une civilisation
fondée sur l'invention et la production, la concurrence
et la consommation. Civilisation fort tapageuse, emphatique,
pléthorique qui n'avait en réalité qu'un seul but secret :
offrir au chat le minimum qu'il exigeait,
soit le confort, le gîte et le couvert.
C'est dire que l'homme inventa des milliers d'objets
bien souvent absurdes, assez vains, tout cela pour produire parallèlement
les quelques éléments indispensables au bien-être du chat :
le coussin, le radiateur, le bol, des centaines de variantes
de préparer la viande, le plat de sciure, le tapis ou la moquette,
le panier d'osier, le pêcheur breton et le vétérinaire,
peut-être aussi la radio puisque les chats aiment bien la musique.
Mais, de tout cela, les hommes ne savent rien. Tout est donc pour le mieux
dans le meilleur monde du chat.*

*Jacques Sternberg
"Dieu, moi et les autres"*

A ma famille,
A mon mari,
A ceux que j'aime.

Résumé

Cette thèse s'inscrit dans le cadre de la statistique non paramétrique et porte sur la classification et la discrimination en grande dimension et plus particulièrement la sélection de variables. Elle comporte à la fois des aspects théoriques et des aspects appliqués.
Une première partie traite du problème de la sélection de variables au moyen de l'algorithme CART, tant dans un contexte de régression que de classification binaire. L'objectif est de fournir une procédure alternative à celle basée sur l'importance des variables, proposée par Breiman *et al.* Cette nouvelle procédure permet de déterminer automatiquement un paquet de variables explicatives qui intervient, de façon essentielle, dans l'explication de la réponse Y. Concrètement, nous fouillons dans une famille finie, mais typiquement grande, de paquets de variables explicatives, et nous déterminons celui qui satisfait "au mieux" notre objectif. Ainsi, nous transformons notre problème de sélection de variables en un problème de sélection de modèle. Afin de procéder à la sélection attendue, nous utilisons d'une part l'algorithme CART et d'autre part, nous nous basons sur la sélection de modèle par pénalisation développée par Birgé et Massart.
Une seconde partie est motivée par un problème réel émanant de la Direction de la Recherche de Renault qui consiste à objectiver la prestation évaluée, en l'occurrence le décollage à plat. Autrement dit, à partir de signaux temporels, mesurés au cours d'essais, nous souhaitons déterminer les signaux pertinents pour expliquer l'agrément de conduite, à savoir le ressenti de confort du conducteur lors de l'évaluation de la prestation. D'autre part, on souhaite identifier les plages temporelles responsables de cette pertinence. Par ailleurs, le caractère fonctionnel des variables explicatives fait que le problème est mal posé dans le sens où le nombre de variables explicatives est nettement supérieur au nombre d'observations. La démarche de résolution s'articule en trois points : un prétraitement des signaux, une réduction de la taille des signaux par compression dans une base d'ondelettes commune et enfin, l'extraction des variables utiles au moyen d'une stratégie incluant des applications successives de la méthode CART.
Enfin, une dernière partie aborde le thème de la classification de données fonctionnelles au moyen de la procédure des k-plus proches voisins, méthode largement étudiée et utilisée dans le cadre de données à valeurs dans un espace fini-dimensionnel. Pour des données de type fonctionnel, on commence par les projeter dans une base de dimension d sur laquelle on utilise alors une procédure des k-plus proches voisins pour sélectionner simultanément la dimension d et la règle de classification. Nous nous intéressons, théoriquement et pratiquement, à cette phase de sélection. Tout d'abord, nous considérons la procédure classique des k-plus proches voisins puis une version légèrement pénalisée, l'idée de la pénalisation ayant été introduite par Biau *et al.*

Table des matières

Chapitre 1

Présentation générale

Sommaire

Les travaux présentés dans cette thèse s'inscrivent dans le cadre de la statistique non paramétrique des problèmes de classification et de régression. En effet, dans chacun des chapitres, nous disposons des réalisations de n copies indépendantes et identiquement distribuées d'un couple de variables aléatoires (X, Y) avec $X = (X^1, ..., X^J)$. La variable Y est appelée variable à expliquer ou encore réponse et, suivant le cadre d'étude, cette variable est à valeurs dans $\mathbb{R}$ ou dans $\{1, ..., K\}$ avec $K \in \mathbb{N}$. Les variables $X^1, ..., X^J$ sont, quant à elles, les variables explicatives et suivant les chapitres, elles appartiennent à $\mathbb{R}$ ou à un espace fonctionnel $\mathcal{F}$ supposé séparable.

A partir de ces données, nous cherchons à mettre en évidence ou à étudier les liens existant entre Y et les X^i. Plus précisément, nous souhaitons répondre aux questions suivantes : au sein du paquet de variables $\{X^1, ..., X^J\}$, existe-t-il un petit paquet de variables qui, à lui seul, contienne toute l'information utile pour expliquer la variable Y ? En présence de variables fonctionnelles, peut-on adapter des méthodes classiques de discrimination telles que CART et les k-plus proches voisins, méthodes couramment utilisées lorsque l'on dispose de données à valeurs dans un espace fini-dimensionnel ?
Les **Chapitres 2, 3 et 4** apportent des réponses à ces interrogations dans des contextes variés que nous allons maintenant présenter.

Le **Chapitre 2** est un travail théorique réalisé en collaboration avec Marie Sauvé, (en thèse à Orsay, sous la direction de Pascal Massart). Il aborde le thème de la sélection de variables au moyen de l'algorithme CART, tant dans un contexte de régression que de classification binaire. L'objectif est de fournir une procédure alternative à celle basée sur l'importance des variables, proposée par Breiman *et al.*. Cette nouvelle procédure permet de déterminer automatiquement un paquet de variables explicatives qui intervient, de façon essentielle, dans l'explication de la réponse Y. Concrètement, nous fouillons une famille finie de paquets de variables explicatives, et nous déterminons celui qui satisfait "au mieux" notre objectif. Ainsi, nous transformons notre problème de sélection de variables en un problème de sélection de modèle. Afin de procéder à la sélection attendue, nous utilisons d'une part l'algorithme CART et d'autre part, nous recourons à la théorie de la sélection de modèle par pénalisation développée par Birgé et Massart dans [16], [15] et [7].

Le **Chapitre 3** est consacré à une application émanant de la Direction de la Recherche de Renault et est mené conjointement avec Jean-Michel Poggi. L'objectif de cette application est d'objectiver la prestation évaluée, en l'occurrence le décollage à plat. Autrement dit, à partir de signaux temporels, mesurés au cours d'essais, nous souhaitons déterminer les signaux permettant d'expliquer l'agrément de conduite, à savoir le ressenti de confort du conducteur lors de l'évaluation de la prestation. D'autre part, on souhaite identifier les plages temporelles responsables de cette pertinence. Par ailleurs, le caractère fonctionnel des variables explicatives fait que le problème est mal posé dans le sens où le nombre de variables explicatives est nettement supérieur au nombre d'observations. La démarche de résolution s'articule en trois points : un prétraitement des signaux, une réduction de la taille des signaux par compression dans une base d'ondelettes commune et enfin, l'extraction des variables utiles au moyen d'une stratégie incluant des applications successives de la méthode CART.

Le **Chapitre 4** est le fruit d'une collaboration avec Magalie Fromont (maître de conférences à Rennes 2). Il aborde le thème de la classification de données fonctionnelles au moyen de la procédure des k-plus proches voisins, méthode largement étudiée et utilisée dans le cadre de données à valeurs dans un espace fini-dimensionnel. Pour des données de type fonctionnel, on commence par les projeter dans une base de dimension d sur laquelle on utilise alors une procédure des k-plus proches voisins pour sélectionner simultanément la dimension d et la règle de classification. Nous nous intéressons, théoriquement et pratiquement, à cette phase de sélection. Tout d'abord, nous considérons la procédure classique des k-plus proches voisins puis une version légèrement pénalisée, l'idée de la pénalisation ayant été introduite par Biau *et al.* dans [12].

Dans cette introduction, avant de présenter plus en détail chacun des trois chapitres, nous rappelons quelques éléments concernant deux outils cruciaux et récurrents de la thèse à savoir CART d'une part et la sélection de modèle “à la Birgé-Massart” d'autre part.

1.1 A propos de CART

Soit $\mathcal{L} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, n réalisations indépendantes du couple de variables aléatoires $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, où $\mathcal{X} = \mathbb{R}^d$.
Dans le cadre de la régression, on a $\mathcal{Y} = \mathbb{R}$ et $Y = s(X) + \varepsilon$ où ε est un bruit additif centré conditionnellement à X et s, définie sur $\mathcal{X}$ par $s(x) = \mathbb{E}[Y|X = x]$ est la fonction de régression.
Dans le cadre de la classification, on a $\mathcal{Y} = \{1, \ldots, J\}$ et s est le classifieur de Bayes, défini sur $\mathcal{X}$ par $s(x) = \underset{j \in \mathcal{Y}}{argmax}\ P(Y = j|X = x)$, c'est-à-dire la meilleure règle de classification possible.
Dans chacun de ces deux contextes, le but est le même, il s'agit d'estimer la fonction s généralement inconnue.

Remarque 1.1.1
Dans le contexte de la classification binaire, on a usuellement $\mathcal{Y} = \{0; 1\}$ et le classifieur de Bayes s'écrit $s(x) = \mathbb{1}_{\eta(x) \geq 1/2}$ avec $\eta(x) = P(Y = 1|X = x)$.

1.1.1 Utilisation et interprétation d'un arbre CART

L'algorithme CART (Classification And Regression Trees), proposé par Breiman *et al.* [20], permet d'obtenir rapidement et facilement des estimateurs par histogramme de la fonction s. Cette méthode repose sur le partitionnement récursif et dyadique de l'espace des observations $\mathcal{X}$, ce qui se représente par un arbre binaire de décision encore appelé arbre de segmentation.

Remarque 1.1.2
Cette représentation par arbre fait que, communément, on procède à l'identification des arbres, des partitions et des estimateurs associés, alors que ce sont des notions de nature différente.

La **Figure 1.1** donne une illustration d'un arbre de classification noté T. Une interprétation en est donnée ci-dessous.

A chaque nœud non terminal t de l'arbre T est associée une division, à savoir une question qui vise à scinder en deux les observations contenues dans le dit nœud t. Les divisions utilisées appartiennent à une famille pré-définie $\mathcal{S}p$. Ainsi, dans le cadre de la régression et de la classification, les divisions retenues sont de la forme $(X^i \leq a)$ ou $(X^i > a)$ et sont donc associés à la question “les observations appartiennent-elles au demi-espace engendré par l'inégalité considérée ?”. Par exemple, la division associée au nœud 2 de l'arbre est $(X2 > 5.7)$.
Une fois une division associée au nœud t, les observations de t satisfaisant la division sont

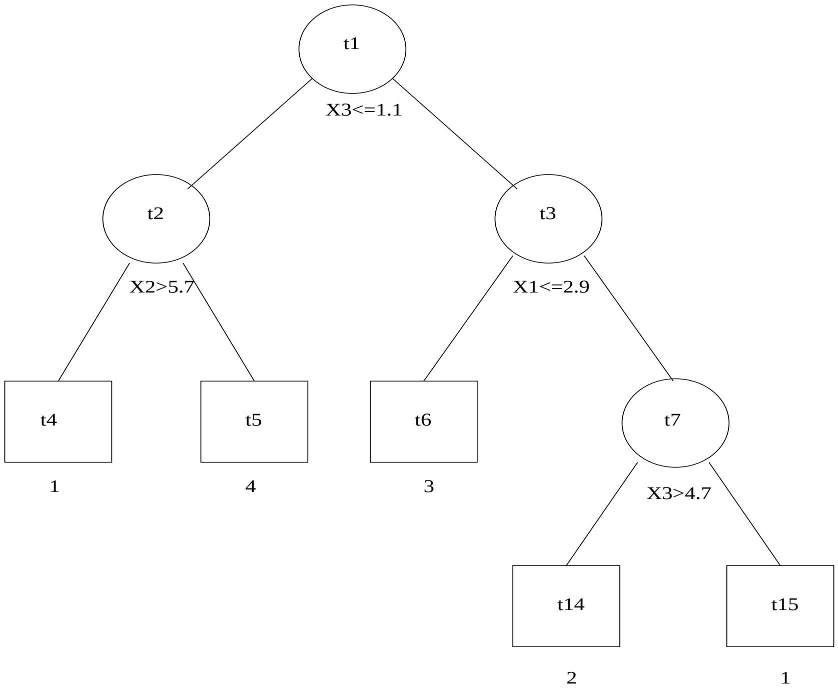

FIG. 1.1 – Illustration d'un arbre binaire de classification T.

envoyées dans le nœud descendant gauche tandis que les autres sont acheminées vers le nœud descendant droit.
La connaissance de cette règle d'acheminement permet de faire descendre, le long de l'arbre T, une donnée x de l'espace des observations $\mathcal{X}$ jusque dans une feuille (un nœud terminal) t_s de T. On lui associe alors une réponse $\hat{Y}_{t_s}$ définie par :

- dans le cadre de la régression, la moyenne empirique :

$$\hat{Y}_{t_s} = \frac{1}{\#\{(X_i, Y_i) \in \mathcal{L};\; X_i \in t_s\}} \sum_{\{X_i;\; X_i \in t_s\}} Y_i$$

- dans le cadre de la classification, la modalité majoritaire :

$$\hat{Y}_{t_s} = \underset{k \in \mathcal{Y}}{argmax}\ \#\{(X_i, Y_i) \in \mathcal{L};\; X_i \in t_s\ et\ Y_i = k\}$$

Puisque l'ensemble $\tilde{T}$ des feuilles de T induit une partition de $\mathcal{Y}$, un estimateur associé à T est alors :

$$\hat{s}_T = \sum_{t \in \tilde{T}} \hat{Y}_t . 1\!\!1_t$$

Par exemple l'estimateur associé à l'arbre représenté dans la **Figure 1.1** est :

$$\begin{aligned}
\hat{s}_T &= (1\!\!1_{t_4} + 1\!\!1_{t_{15}}) + 2.1\!\!1_{t_{14}} + 3.1\!\!1_{t_6} + 4.1\!\!1_{t_5} \\
&= 1.(1\!\!1_{\{X^3 \leq 1.1; X^2 > 5.7\}} + 1\!\!1_{\{X^1 > 2.9; 1.1 < X^3 \leq 4.7\}}) + 2.1\!\!1_{\{X^1 > 2.9; X^3 > 4.7\}} \\
&\quad + 3.1\!\!1_{\{X^3 > 1.1; X^1 \leq 4.9\}} + 4.1\!\!1_{\{X^3 \leq 1.1; X^2 \leq 5.7\}}
\end{aligned}$$

1.1.2 Construction d'un arbre CART

Sans se concentrer sur la manière dont l'algorithme CART procède, on peut la résumer de la façon suivante.
Considérons pour chaque $k \in \{1, \ldots, n\}$, le "meilleur" arbre de taille k, autrement dit l'arbre à k feuilles qui minimise l'erreur de classification ou les résidus suivant le cadre d'étude. CART construit cette suite seulement implicitement, en fait, une sous-suite est astucieusement choisie.
Ensuite, à l'aide d'un critère pénalisé, sont comparés ces différents arbres et on en sélectionne un, le critère pénalisé mis en œuvre servant à réaliser un compromis entre la taille des arbres et leur qualité d'ajustement. En effet, l'objectif est de trouver un compromis entre l'erreur d'apprentissage et l'erreur de généralisation. Or, un arbre de grande taille est trop fidèle aux données d'apprentissage ce qui a pour conséquence d'obtenir une estimation de l'erreur de généralisation trop optimiste car trop faible. A l'inverse, un arbre de petite taille engendre une perte de précision et une estimation de l'erreur de généralisation trop grande. Ainsi, l'arbre sélectionné a, en général, une taille ni trop grande ni trop petite.

Dans la pratique, afin de déterminer le "meilleur" arbre, CART procède en trois étapes que sont :

- <u>étape 1</u> : Construction de l'arbre maximal T_{max}
  De manière récursive et dyadique, on construit une suite de partitions de plus en plus fines de l'espace des observations $\mathcal{X}$ jusqu'à ce que chacun des éléments de la partition ne contienne qu'une seule observation ou des observations de même réponse. Cela revient à construire un arbre en le développant au maximum, autrement dit jusqu'à ce que les nœuds terminaux ne contiennent que des observations de même réponse.
  A chacune de ces partitions est associé un arbre. L'arbre maximal, noté T_{max}, est celui associé à la partition la plus fine.

  Remarque 1.1.3
  Développer l'arbre consiste à diviser en 2 chaque nœud t à l'aide de la division optimale $\delta^*(t)$ qui maximise le gain en homogénéité et qui est définie par :

  $$\delta^*(t) = \begin{cases} \underset{\delta \in \mathcal{S}_p}{argmax}(H(t) - H(t_g) - H(t_d)) & \text{pour la régression,} \\ \underset{\delta \in \mathcal{S}_p}{argmax}(H(t) - p_g.H(t_g) - p_d.H(t_d)) & \text{pour la classification.} \end{cases}$$

  où :

  – t_g (resp. t_d) est le nœud descendant gauche (resp. droit) induit par la division δ du nœud t ;

- p_g (resp. p_d) est la proportion des observations de t envoyées dans le nœud t_g (resp. t_d) par la division δ ;
- H est une fonction d'hétérogénéité : la variance dans le cadre de la régression et dans celui de la classification, des fonctions basées sur l'indice de Gini ou l'entropie de Shannon.

□

- étape 2 : Élagage de T_{max}
  De la suite d'arbres précédemment obtenue, on extrait une sous-suite par minimisation, pour $\alpha \geq 0$, du critère pénalisé

$$crit_\alpha(T) = \gamma_n(\hat{s}_T) + \alpha\frac{|T|}{n},$$

  où γ_n est l'erreur quadratique définie par

$$\gamma_n(u) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - u(X_i))^2$$

  et $|T|$ le nombre de feuilles de l'arbre T.
  Lorsque α augmente, la minimisation de $crit_\alpha$ conduit à des arbres dont la taille décroît.

  Breiman *et al.* ont montré qu'il existe une suite finie $\alpha_0 = 0 < \alpha_1 < \ldots < \alpha_K$ et la suite associée de sous-arbres emboîtés $\{T_k\}_{0\leq k\leq K}$ qui résument toute l'information et telles que pour $\alpha \in [\alpha_k, \alpha_{k+1}[,\ \underset{T\preceq T_{max}}{argmin}\ crit_\alpha(T) = T_k$.
  La suite obtenue à l'issue de la phase d'élagage est la suite $\{T_k\}_{0\leq k\leq K}$.

- étape 3 : Sélection finale
  Cette dernière phase sélectionne, à l'aide d'un échantillon témoin ou par validation croisée, le "meilleur" arbre dans la suite précédemment construite.

Remarque 1.1.4
Avant CART, d'autres méthodes de construction d'arbres de segmentation étaient disponibles comme la méthode CHAID (cf. Nakache et Confais [74]).
Dans ces méthodes, la construction fait intervenir une règle d'arrêt dépendant d'un seuil arbitraire α. L'originalité de CART est de supprimer cette règle d'arrêt au profit d'une procédure entièrement automatique. □

1.1.3 Importance des variables

La notion d'importance des variables, définie par Breiman *et al.* [20], permet, relativement à un arbre CART donné T, de hiérarchiser les variables explicatives $\{X^1, \ldots, X^p\}$ en leur attribuant une note comprise entre 0 et 100. En voici le principe.

Pour tout nœud t de l'arbre T et toute variable explicative X^m, on détermine la division de substitution $\delta_m(t)$, portant sur la variable X^m, qui se rapproche le plus de la division optimale $\delta^*(t)$, au sens où :

$$\delta_m(t) = \underset{\delta \in \{(X^m \leq a) \cup (X^m > a)\}}{argmax} (p_{gg}(\delta, \delta^*(t)) + p_{dd}(\delta, \delta^*(t)).$$

où $p_{gg}(\delta, \delta^*(t))$ (resp. $p_{dd}(\delta, \delta^*(t))$) est un estimateur de la probabilité que les divisions $\delta^*(t)$ et δ envoient une observation du nœud t dans son descendant gauche (resp. droit).

On mesure alors le gain d'homogénéité apporté par la division $\delta_m(t)$ par :

$$\Delta H(\delta_m(t), t) = \begin{cases} H(t) - H(t_g^m) - H(t_d^m) & \text{pour la régression,} \\ H(t) - p_g^m.H(t_g^m) - p_d^m.H(t_d^m) & \text{pour la classification.} \end{cases}$$

où :

- t_g^m (resp. t_d^m) est le nœud descendant gauche (resp. droit) de t induit par la division de substitution $\delta_m(t)$.
- p_g^m (resp. p_d^m) est la proportion des observations de t envoyées dans le nœud t_g^m (resp. t_d^m) par la division $\delta_m(t)$;

On définit alors l'importance de la variable X^m par :

$$I(X^m) = \sum_{t \in T - \tilde{T}} \Delta H(\delta_m(t), t)$$

Ainsi, l'importance de la variable X^m est la somme, sur tous les nœuds t de l'arbre, des gains en homogénéité si l'on remplaçait à chaque nœud t la division optimale par la division de substitution $\delta_m(t)$.

Généralement, on ramène cette importance sur une échelle comprise entre 0 et 100 par :

$$\tilde{I}(X^m) = 100.\frac{I(X^m)}{\underset{k \in \{1,\ldots,p\}}{max} I(X^k)}$$

Ceci induit un ordre sur les variables explicatives. On considère comme importantes les variables dont l'importance est supérieure à un seuil convenablement choisi.

1.2 La sélection de variables et la sélection de modèle

1.2.1 Sélection de variables

La thématique de la sélection de variables, que l'on retrouve d'un point de vue théorique dans le **Chapitre 2** et d'un point de vue pratique dans le **Chapitre 3**, consiste à extraire de l'ensemble des variables explicatives disponibles un paquet de variables, le plus petit possible, tel que les variables retenues dans ce dernier soient capables d'expliquer la réponse.

Si la sélection de variables est d'un intérêt évident pour toute valeur de p, le nombre de variables explicatives, elle est particulièrement cruciale lorsque p est très grand. Or, avec le développement des outils informatiques qui permettent de stocker et de traiter toujours davantage de données, ce type de situation se rencontre fréquemment. Ceci explique en partie l'intérêt actuellement porté au thème de la sélection de variables.
Une autre explication est apportée par le fait que, si aujourd'hui le nombre p a tendance à être grand, le nombre d'observations n reste faible dans bien des situations. Par exemple, dans le domaine biomédical, il est courant de disposer d'une multitude de variables explicatives, cependant en raison du faible nombre de malades soumis au protocole, n est généralement petit. Ceci se retrouve également dans le contexte industriel en raison du coût élevé des essais.
Ceci donne alors naissance à des problèmes que l'on qualifie de mal posé, dans le sens où $n << p$, et pour lesquels, il faut déterminer les variables explicatives influentes avant de pouvoir construire de "bons" classifieurs ou prédicteurs.

En ce qui nous concerne, nous nous sommes intéressés à la sélection de variables pour p a priori quelconque et nous avons privilégié une approche par CART.

Une idée naïve pour procéder à de la sélection de variables au travers de CART consiste à construire l'arbre maximal avec toutes les variables explicatives et à ne retenir que celles qui définissent les premiers nœuds de cet arbre ou les variables qui étiquettent l'arbre optimal après élagage et sélection. Cette idée intuitive possède de nombreux inconvénients dont celui d'omettre certaines variables influentes.
Une autre idée, mise en œuvre dans le **Chapitre 2**, consiste à adopter une approche par sélection de modèle. Ceci est motivé par le fait que l'étape d'élagage de CART peut s'interpréter en tant que phase de sélection de modèle.

Dans la suite de ce paragraphe, nous détaillons le principe de la sélection de modèle, et nous focalisons plus particulièrement notre attention sur la sélection de modèle par pénalisation, méthode utilisée dans le **Chapitre 2**.

1.2.2 Sélection de modèle

On observe n couples de variables aléatoires indépendantes $(X_1, Y_1), ..., (X_n, Y_n)$ à valeurs dans un espace mesurable $\Xi = \mathcal{X} \times \mathcal{Y}$. On suppose, par ailleurs, que ces n couples sont des copies du couple de variables aléatoires (X, Y) dont la loi jointe est notée P.
On considère la fonction de régression η définie sur $\mathcal{X}$ par

$$\eta(x) = \mathbb{E}[Y|X = x].$$

Dans le cadre de la régression gaussienne, η est le meilleur ajustement de Y par X au sens des moindres carrés.
Dans le contexte de la classification binaire, la meilleure règle de discrimination ou de décision

est réalisée par le classifieur de Bayes s défini sur $\mathcal{X}$ par

$$s(x) = 1\!\!\mathrm{I}_{\eta(x) \geq 1/2}.$$

Lorsque la distribution P est inconnue, les fonctions η et s le sont également. On souhaite alors estimer l'une ou l'autre de ces fonctions selon le contexte.
Une méthode usuelle pour estimer la fonction η ou s qui dépend d'une distribution inconnue P consiste à procéder par "minimisation du contraste empirique".

Minimisation du contraste empirique

Soit $\mathcal{S}$ un espace contenant la fonction à estimer f, avec $f = \eta$ dans le cadre de la régression et $f = s$ dans le contexte de la classification. Par exemple, dans le cadre de la régression (resp. classification binaire), $\mathcal{S}$ est l'ensemble des fonctions mesurables à valeurs dans $\mathbb{R}$ (resp. $\{0; 1\}$). On considère un contraste γ défini sur $\mathcal{S} \times \Xi$, à valeurs dans $[0, 1]$ et tel que

$$f = \underset{t \,\in\, \mathcal{S}}{argmin}\ \mathbb{E}[\gamma(t, (X, Y))].$$

Un estimateur $\hat{f}$ de la fonction f est alors obtenu en minimisant, sur un sous-ensemble S de $\mathcal{S}$, appelé modèle, le contraste empirique γ_n associé à γ. Autrement dit

$$\hat{f} = \underset{t \,\in\, S}{argmin}\ \gamma_n(t)$$

avec

$$\gamma_n(t) = P_n[\gamma(t, .)] = \frac{1}{n} \sum_{i=1}^{n} \gamma(t, (X_i, Y_i)).$$

Afin d'évaluer la performance d'un estimateur t de la fonction f, on introduit la fonction de perte l définie sur $\mathcal{S} \times \mathcal{S}$ par

$$l(f, t) = \mathbb{E}[\gamma_n(t) - \gamma_n(f)] \geq 0,$$

et on considère ensuite le risque associé

$$R(f, t) = \mathbb{E}[l(f, t)].$$

Soit $\bar{f}$ un minimiseur de $l(f, .)$ sur S, une décomposition du risque de $\hat{f}$ est alors

$$R(f, \hat{f}) = l(f, \bar{f}) + \mathbb{E}[l(\bar{f}, \hat{f})]. \tag{1.1}$$

Le terme non aléatoire $l(f, \bar{f})$ est le terme de "biais" ; il correspond à l'erreur induite par le choix du modèle S puisqu'il est analogue à la distance entre la fonction f et le modèle S. Le terme $\mathbb{E}[l(\bar{f},\hat{f})]$ est, quant à lui, un terme de "variance" qui quantifie l'erreur d'approximation dans le modèle S.

L'objectif est alors de déterminer un modèle S qui minimise le risque défini par (1.1). Ceci s'avère d'autant plus difficile que les deux termes intervenant dans la décomposition de ce risque n'évoluent pas dans le même sens. Plus précisément, lorsque la taille du modèle S augmente, le terme de "biais" devient faible alors que celui de "variance" croît. A l'inverse, lorsque la taille du modèle S diminue, le terme de "biais" s'accroît tandis que celui de "variance" s'amenuise. Par conséquent, considérer un modèle S trop petit ou trop grand conduit à un risque, évalué sur S, trop important.
Par conséquent, au lieu de travailler sur un unique modèle S, déterminé a priori en espérant que son risque associé soit faible, on considère une collection de modèles $\{S_m\}_{m\in\mathcal{M}}$ au sein de laquelle on va sélectionner un modèle.

L'idéal serait de pouvoir déterminer le modèle dit "oracle" et noté $S_{m(f)}$ qui minimise sur $\mathcal{M}$ le risque défini par (1.1), à savoir tel que $m(f) = \underset{m\in\mathcal{M}}{argmin}\ R(f,\hat{f}_m)$. Cependant, le risque est difficilement évaluable puisque le terme de "biais" dépend de la fonction inconnue f, ce qui conduit à l'impossibilité de déterminer le modèle $S_{m(f)}$. L'idée consiste alors à mettre en œuvre une procédure de sélection de modèle basée sur les données et d'en évaluer les performances en comparant son risque au risque du modèle "oracle" qui constitue une référence. Plus précisément, on souhaite que la procédure de sélection de modèle fournisse un modèle $S_{\tilde{m}}$ et un estimateur associé $\hat{f}_{\tilde{m}}$ tel que :

$$R(f,\hat{f}_{\tilde{m}}) \leq C.\underset{m\in\mathcal{M}}{inf}\ R(f,\hat{f}_m)$$

avec C proche de 1.

Un type de procédure satisfaisant est, par exemple, la méthode de sélection de modèle par pénalisation qui procède en minimisant une perte empirique à laquelle on adjoint un terme de pénalité.
On trouve déjà ce type de procédure dans les années 70 : Akaike [2], Mallows [67], Schwarz [83] et Rissanen [80] ont respectivement défini les critères pénalisés AIC, C_p, BIC et MDL. Plus récemment, on peut citer les travaux de Baron et Cover [8], de Polyac et Tsybakov [77] et de Birgé et Massart ([7], [15], [16]).

Dans la suite de ce paragraphe, nous nous focalisons sur la méthode de sélection de modèle par minimisation du contraste pénalisé, méthode développée par Birgé et Massart.

Sélection de modèle par pénalisation

On considère une famille dénombrable de modèles $\{S_m\}_{m\in\mathcal{M}}$ de dimensions respectives D_m et la collection d'estimateurs associée $\{\hat{f}_m\}_{m\in\mathcal{M}}$ obtenue par minimisation du contraste empirique γ_n sur chacun des modèles S_m. On souhaite alors choisir un estimateur parmi ceux proposés.
Puisque le risque défini par (1.1) dépend de la fonction inconnue f que l'on souhaite estimer, on ne peut pas utiliser ce critère en l'état.

On considère alors un critère pénalisé, dépendant uniquement des données, défini par

$$crit(m) = \gamma_n(\hat{f}_m) + pen(m)$$

où $pen : \mathcal{M} \to \mathcal{R}$ est la fonction de pénalité à déterminer.

On choisit alors $\hat{m}$ comme étant le minimiseur sur $\mathcal{M}$ de la fonction $crit$, et on définit l'estimateur final $\tilde{f}$ par

$$\tilde{f} = \hat{f}_{\hat{m}}.$$

En conséquence, le problème est ramené à la détermination d'une fonction de pénalité pen adaptée au problème, à savoir telle que l'estimateur final $\tilde{f}$ vérifie des inégalités de type "oracle" comme

$$\mathbb{E}[l(f,\tilde{f})] \leq C \underset{m\in\mathcal{M}}{inf} \mathbb{E}[l(f,\hat{f}_m)]$$

avec C une constante, si possible universelle, proche de 1.

Le choix de la fonction de pénalité dépend essentiellement de la complexité de la collection $\mathcal{M}$.

La détermination de la fonction de pénalité a fait l'objet d'une multitude d'articles. On citera par exemple ceux de Mallows [67], Birgé et Massart ([16], [15]) ou Barron, Birgé et Massart [7] dans le cadre de la régression et ceux de Lugosi et Zeger [65], Massart [69] ou Bartlett, Boucheron et Lugosi [9] dans le cadre de la classification binaire.

Dans le cadre de la régression gaussienne, voici à titre d'exemple typique, un des résultats obtenus par Birgé et Massart [16].

Soit $\{L_m\}_{m\in\mathcal{M}}$ une famille de poids vérifiant

$$\sum_{m\in\mathcal{M};D_m>0} e^{-L_m D_m} \leq \Sigma \leq +\infty.$$

Alors, si la fonction de pénalité satisfait, pour $K > 1$ et pour tout $m \in \mathcal{M}$,

$$pen(m) \geq K\sigma^2 \frac{D_m}{n}\left(1 + 2L_m + 2\sqrt{L_m}\right),$$

le risque de l'estimateur final $\tilde{f}$ vérifie

$$R(f,\tilde{f}) \leq C(K) \underset{m\in\mathcal{M}}{inf} \left(\underset{t\in S_m}{inf}\ l(f,t) + pen(m)\right) + C'(K)\sigma^2\frac{\Sigma}{n}.$$

où σ^2 est la variance du bruit.

1.3 Méthode de Sélection de Variables utilisant CART

Ce travail est le fruit d'une collaboration avec Marie Sauvé (en thèse à Orsay, sous la direction de Pascal Massart), et aborde le thème de la sélection de variables.

1.3.1 Le problème

Soit $\mathcal{L} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, n copies indépendantes d'un couple de variables aléatoires (X, Y) où Y est la variable réponse et $X = (X^1, \ldots, X^p)$ un vecteur de p variables explicatives. On souhaite déterminer, parmi ces p variables, le plus petit paquet de variables capable, à lui seul, d'expliquer la variable Y. Autrement dit, on cherche à mettre en œuvre de la sélection de variables.

De nombreuses méthodes de sélection de variables existent, notamment dans le cadre des modèles linéaires. On peut, par exemple, citer la "Subset Selection", Lasso ou encore LARS qui sont des méthodes exhaustives ou pénalisées qui font chacune intervenir le critère des moindres carrés.

Pour notre part, nous privilégions CART et une méthode pénalisée en recourant à une approche de sélection de modèle par minimisation d'un contraste empirique pénalisé. Voici succinctement une description de notre procédure.
Soit $\Lambda = \{X^1, \ldots, X^p\}$. Pour tout sous-ensemble ou paquet M de Λ, on construit l'arbre CART maximal $T_{max}^{(M)}$ en ne faisant intervenir dans les divisions de l'arbre que les variables du paquet M. Ensuite, pour tout sous-arbre M de $T_{max}^{(M)}$ noté $T \preceq T_{max}^{(M)}$, on considère le modèle $S_{M,T}$ constitué des fonctions constantes par morceaux sur la partition induite par T. Pour finir, on procède à la sélection de modèle dans la collection $\{S_{M,T}, M \in \mathcal{P}(\Lambda), T \preceq T_{max}^{(M)}\}$, en minimisant un contraste empirique pénalisé.

La question naturelle qui se pose est : Comment choisir le terme de pénalité afin que la procédure soit théoriquement valide ?
Ceci constitue le but majeur de notre travail.

Dans un second temps, nous nous sommes intéressées à l'application de la procédure proposée lorsque la valeur de p est grande.
La question qui survient est : Comment déterminer une famille $\mathcal{P}^*$ qui serait plus petite que $\mathcal{P}(\Lambda)$ et qui pourrait se substituer à $\mathcal{P}(\Lambda)$ dans la procédure ?

1.3.2 Le cadre

Afin d'apporter une réponse à la première interrogation, nous considérons les deux cadres d'étude que sont :

- la régression définie par $Y = s(X) + \varepsilon$ avec :
  - $\mathbb{E}[\varepsilon|X] = 0$;

– il existe $\rho \geq 0$ et $\sigma > 0$ tels que pour tout $\lambda \in (-1/\rho, 1/\rho)$, $log\mathbb{E}\left[e^{\lambda\varepsilon_i}|X_i\right] \leq \frac{\sigma^2\lambda^2}{2(1-\rho|\lambda|)}$, avec la convention $1/0 = \infty$;
– $\|s\|_\infty \leq R$ avec $R > 0$.

Remarque 1.3.1
Ces hypothèses permettent de prendre en considération les modèles gaussiens, mais également des modèles plus généraux. □

• la classification binaire, $Y \in \{0; 1\}$ avec :

– s est le classifieur de Bayes défini par $s(x) = 1\!\mathrm{I}_{\eta(x)\geq 1/2}$ où $\eta(x) = P(Y = 1|X = x)$;

– une hypothèse de marge du type : $\exists h > 0,\ \forall x \in \chi,\ |2\eta(x) - 1| > h$.

Remarque 1.3.2
L'hypothèse sur la *log* transformée de Laplace des erreurs implique que le moment d'ordre 2 des erreurs est majoré par σ^2, soit :

$$V(\varepsilon_i|X_i) \leq \sigma^2$$

□

Par ailleurs, dans chacun de ces deux contextes, on considère deux situations :

$(M1)$: L'échantillon $\mathcal{L}$ est scindé en trois parties indépendantes $\mathcal{L}_1$, $\mathcal{L}_2$ et $\mathcal{L}_3$ de tailles respectives n_1, n_2 et n_3 et respectivement appelées échantillon d'apprentissage, échantillon de validation et échantillon test.
L'échantillon $\mathcal{L}_1$ sert à la construction d'un arbre maximal tandis que $\mathcal{L}_2$ est utilisé dans la phase de sélection de modèle qui produit une suite de sous-arbres $\{(T_k)_{1\geq k\geq K}\}$.
L'échantillon $\mathcal{L}_3$ permet de procéder à la phase de sélection finale d'un paquet ainsi que d'un estimateur.

$(M2)$: L'échantillon $\mathcal{L}$ est divisé en deux parties indépendantes $\mathcal{L}_1$ et $\mathcal{L}_3$. La construction des arbres maximaux et la phase de sélection de modèle s'opèrent toutes les deux avec $\mathcal{L}_1$ tandis que $\mathcal{L}_3$ est utilisé, ici encore, comme échantillon témoin pour la phase de sélection finale.

Afin de déterminer les fonctions de pénalité bien adaptées à chacune des situations considérées, nous utilisons les méthodes décrites dans le paragraphe 1.2.2 et plus précisément les résultats de Birgé et Massart ([15],[16]) dans le cadre de la régression et de Massart et Nédélec [71] dans celui de la classification.

1.3.3 Les résultats

Dans ce paragraphe, nous utilisons les notations suivantes.
Soit M un élément de $\mathcal{P}(\Lambda)$, autrement dit un paquet de variables explicatives, de cardinal noté $|M|$. A chaque sous-arbre T de $T_{max}^{(M)}$, ce que l'on note $T \preceq T_{max}^{(M)}$, on associe le sous-espace $S_{M,T}$ des fonctions constantes par morceaux sur la partition $\tilde{T}$ où $\tilde{T}$ représente l'ensemble des feuilles de l'arbre T ; $|T|$ désigne le cardinal de $\tilde{T}$. Pour finir, $\hat{s}_{M,T}$ désigne l'estimateur par histogramme associé à l'arbre T.
Lors de la phase de sélection de modèle, on procède à la minimisation du critère

$$crit_{\alpha,\beta}(M,T) = \gamma_{n_2}(\hat{s}_{M,T}) + pen(M,T) \tag{1.2}$$

à α et β fixés, α et β étant des paramètres intervenant dans l'expression de la fonction de pénalité pen.
On obtient alors, pour chaque valeur du couple (α, β) un estimateur $\tilde{s} = \hat{s}_{\widehat{M,T}}$ où

$$(\widehat{M,T}) = \underset{M\in\mathcal{P}(\Lambda);T\preceq T_{max}^{(M)}}{argmin} \; crit_{\alpha,\beta}(M,T).$$

A l'aide de l'échantillon témoin, on sélectionne parmi la collection d'estimateurs $\{\tilde{s}; \alpha > \alpha_0 \text{ et } \beta > \beta_0\}$ l'estimateur $\tilde{\tilde{s}}$, ainsi que le paquet optimal associé.

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

Le but est de déterminer une fonction de pénalité pen pour laquelle on puisse obtenir des inégalités de type “oracle”.

Les résultats sont obtenus conditionnellement à la construction des arbres maximaux $T_{max}^{(M)}$ et les performances des différents estimateurs sont évaluées par leurs risques ou leurs normes empiriques conditionnels.

Dans le cadre de la régression, les résultats font intervenir les normes $\|.\|_{n_1}$, $\|.\|_{n_2}$ et $\|.\|_{n_3}$ qui sont respectivement les normes empiriques sur les grilles $\{X_i; (X_i, Y_i) \in \mathcal{L}_1\}$, $\{X_i; (X_i, Y_i) \in \mathcal{L}_2\}$ et $\{X_i; (X_i, Y_i) \in \mathcal{L}_3\}$.

Voici, deux résultats obtenus dans le cadre de la régression. Ils font intervenir les notations introduites dans le paragraphe 1.3.2. Le premier concerne l'estimateur $\tilde{s}$ et le second l'estimateur final $\tilde{\tilde{s}}$.

Si $\tilde{s}$ est obtenu par $(M1)$:

Si la fonction de pénalité est telle que :
$\forall\ M \in \mathcal{P}(\Lambda)$ et $\forall\ T \preceq T_{max}^{(M)}$

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

alors, pour α et β suffisamment grands et sous certaines conditions, on a

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

Ce premier résultat valide le choix des fonctions de pénalisation puisqu'une inégalité de type "oracle" est obtenue.

Quand on analyse ce résultat, on constate que la fonction de pénalité *pen* est la somme d'un terme proportionnel à $|T|$, le nombre de feuilles et d'un terme proportionnel à $|M|$ la taille du paquet considéré. Le premier terme n'est autre que la pénalité proposée par Breiman *et al.* [20], il sert à pénaliser les arbres de trop grande taille. Le second terme est propre à la sélection de variables, en effet il pénalise les modèles impliquant un trop grand nombre de variables.

Des résultats similaires sont obtenus dans la situation $(M2)$ ou dans le contexte de la classification, seule la forme des pénalités se trouve modifiée.

Si $\tilde{\tilde{s}}$ est obtenu par (M1) :

Pour $\xi > 0$, avec probabilité $\geq 1-h(\xi)$,
$\forall \eta \in (0,1)$,

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

L'intérêt de ce résultat réside dans le fait qu'il permet de valider la phase de sélection finale de notre procédure. En effet, il montre, avec grande probabilité, que la sélection finale par échantillon test, n'altère pas, de façon notable la qualité de l'estimateur final. Dans la situation $(M2)$, un résultat identique est obtenu.
En ce qui concerne le contexte de la classification, le résultat obtenu est analogue sauf que l'inégalité en grande probabilité est remplacée par une inégalité en espérance qui autorise la comparaison entre l'estimateur final et tout autre.

Il apparaît donc que dans chacune des situations envisagées, la procédure est validée théoriquement par l'obtention de fonctions de pénalité conduisant à des estimateurs convenables en termes de performance.

En ce qui concerne la seconde interrogation, à savoir la mise en œuvre pratique de cette procédure, nous proposons d'appliquer la procédure à une famille restreinte de paquets de variables, déterminée à partir des données et d'une idée initiée par Poggi et Tuleau dans [76] qui consiste à associer l'importance des variables, définie par Breiman *et al.* [20] et une procédure de sélection ascendante, classique en régression linéaire.

La mise en œuvre, sur un exemple simulé, de la procédure pratique, autrement dit restreinte à une famille convenablement choisie, nous permet de constater que nous obtenons alors une procédure efficace en termes de sélection de variables et de temps de calculs.

Par ailleurs, le fait que la famille soit déterminée à l'aide d'une stratégie basée sur les données, et non de façon déterministe, justifie l'emploi de la pénalité théorique définie en considérant l'ensemble des 2^p paquets possibles.

1.4 L'objectivation de l'agrément de conduite automobile

Le Chapitre 3 est dévolu à la présentation d'un travail effectué dans le cadre d'un contrat de collaboration de recherche avec la Direction de la Recherche de Renault. Il a été mené avec Jean-Michel Poggi et porte sur un problème pratique de sélection de variables.

1.4.1 Le problème

L'industrie automobile, comme par exemple Renault, souhaite satisfaire sa clientèle. Dans ce but, des sondages sont réalisés afin de déterminer les prestations à améliorer. Une fois ces dernières identifiées, il s'agit de les quantifier ou objectiver afin pouvoir intégrer les résultats dans le cahier des charges relatif à la conception du véhicule.
Concrètement, cela signifie qu'il faut déterminer des critères véhicule, encore appelés critères "physiques", responsables de la satisfaction du conducteur (ou agrément de conduite) liée à la prestation évaluée.
L'étude qui nous occupe est relative à la boîte de vitesses et au confort ressenti par le conducteur lors de la mise en mouvement du véhicule.

1.4.2 Le cadre

Afin de pouvoir mener à bien cette étude, une campagne d'essais a été réalisée. Celle-ci a requis 7 pilotes essayeurs et a mêlé différentes conditions de roulage afin de traduire diverses situations (route, autoroute, ...) et façons de conduire (brusque, douce, ...) ainsi que la charge du véhicule. De même, elle a impliqué différents réglages de la boîte de vitesses, organe évalué lors de la prestation considérée, en l'occurrence le décollage à plat pour un groupe motopropulseur à boîte de vitesses robotisée.

En raison du caractère purement subjectif de la prestation, puisqu'il s'agit du confort ressenti, et donc de la difficulté inhérente à l'évaluation des essais, ces derniers ont été réalisés par des comparaisons par paires. Plus précisément, les pilotes testaient successivement deux produits distincts, un produit étant une combinaison d'un réglage de boîte et de conditions de roulage spécifiques. A l'issue de chaque paire testée, le pilote donnait alors l'essai préféré, sachant que, bien entendu, il n'avait pas connaissance des réglages de la boîte de vitesses afin de ne pas influencer son jugement.

Par ailleurs, lors de chacun des essais, divers signaux "physiques", tels les accélérations ou les couples, ont été relevés à l'aide de capteurs disposés dans le véhicule utilisé pour tous les essais.

Après l'élimination des essais associés à des mesures erronées, la sélection d'un ensemble de signaux mesurés et le traitement de la réponse pilote, conduisant à l'obtention d'un ordre sur les produits qui est à la fois total et indépendant des pilotes, les données qui nous ont été communiquées, et qui constituent nos données d'étude, sont les suivantes :

$$\left\{\begin{array}{l} X_i = (X_i^1, \ldots, X_i^{21}) \text{ avec } X_i^j = X_i^j(t) \text{ la } j^{ème} \text{ variable fonctionnelle (signal)} \\ \qquad \text{mesurée lors de l'essai } i, \\ \\ Y_i = \text{ le rang, dans l'ordre total, attribué au produit testé au cours de l'essai } i. \end{array}\right.$$

Une étude menée par Ansaldi [4] établit une méthodologie d'objectivation, autrement dit une méthodologie à même de définir les grandeurs physiques représentatives de la prestation étudiée et leurs plages de valeurs optimales. Elle peut se résumer en trois grandes étapes que sont :

- l'association d'un agrément à chacun des produits :
  A l'issue des essais, on obtient une préférence individuelle (pilote par pilote) donnée sur des paires de produits. A partir de ces données "locales", un ordre "global" est obtenu sur les produits. Il permet leur classement, du plus au moins apprécié, indépendamment du pilote.

- l'extraction de critères et l'Analyse discriminante :
  Après la génération d'une liste très importante de critères candidats grâce notamment à des règles d'expertise, ceux utiles à l'explication de l'agrément sont identifiés par une méthode d'analyse discriminante arborescente par moindres écarts. Cette étape vise à extraire des signaux la quantité minimum d'information suffisante pour expliquer l'agrément.

- le calcul d'intervalles de tolérance :
  Les critères pertinents étant déterminés, on cherche pour chacun d'eux la plage de valeurs la plus grande compatible avec la maximisation de l'agrément sous contraintes.

Pour notre part, nous nous sommes intéressés à la seconde phase en essayant de s'affranchir au maximum des connaissances "métier" et en tenant davantage compte de l'aspect fonctionnel du problème.

1.4.3 Les résultats

Afin de déterminer les critères "physiques" pertinents pour l'explication de l'agrément de conduite automobile, nous devons réaliser une double phase de sélection de variables. La première consiste à identifier les variables fonctionnelles pertinentes et la seconde, à déterminer pour chacune d'elles les plages temporelles responsables de cette pertinence.
En raison de la nécessité d'interpréter physiquement les résultats, chacune des deux sélections

de variables doit s'effectuer dans le paquet des variables d'origine et non au sein d'un ensemble de variables construit à partir des données d'origine. Par exemple, toute combinaison linéaire est proscrite. Ceci exclut notamment des approches de type PLS ou de régression sur composantes principales.
Par ailleurs, on note que le caractère fonctionnel des données nous empêche d'appliquer la procédure de sélection de variables décrite au **Chapitre 2**.

Alors, afin de procéder à la double phase de sélection, nous avons adopté une démarche essentiellement basée sur deux outils : les ondelettes d'une part et CART d'autre part.
Les ondelettes permettent de concentrer, en un petit nombre de coefficients, l'information contenue dans les signaux, tout en préservant une interprétation dans la grille temporelle d'origine.
L'algorithme CART intervient quant à lui dans la double phase de sélection puisque par son intermédiaire et l'importance des variables qui lui est étroitement associée, nous déterminons non seulement les variables fonctionnelles pertinentes, mais également leurs plages temporelles.

La démarche est donnée constitué des trois phases :

- prétraitements :
  Cette phase applique aux signaux mesurés des étapes de synchronisation, de débruitage par ondelettes et de recalage, visant à rendre les signaux "homogènes".

- compression :
  En raison de la dimension élevée (chacun des 114 essais est résumé par 21 signaux d'environ 1000 points), on souhaite mettre à profit la redondance de l'information dans une courbe et réduire la dimension de l'espace des variables explicative afin de réduire le fléau de la dimension. La stratégie adoptée consiste à travailler variable fonctionnelle par variable fonctionnelle, et à projeter, dans un espace d'approximation commun, chacun des signaux.

- sélection :
  Après compression, chacun des signaux est résumé par un petit nombre de coefficients d'approximation. Cependant, cette réduction de la dimension s'avère encore insuffisante. On procède alors à une sélection pas à pas qui s'articule en cinq étapes. La première consiste à sélectionner, variable fonctionnelle par variable fonctionnelle, les coefficients discriminants en recourant à l'importance des variables. La seconde étape hiérarchise les variables fonctionnelles au moyen du coût de fausse classification. Les deux étapes suivantes procèdent à une première sélection de variables fonctionnelles en construisant une suite de modèles emboîtés et en choisissant celui de coût minimal. L'ultime étape réalise la sélection finale à l'aide de l'importance des variables.

Les résultats obtenus, à l'issue de cette stratégie, recoupent pour l'essentiel ceux de l'étude menée par Ansaldi [4] et en termes d'erreur les résultats sont comparables, ce qui est intéressant en raison du caractère totalement "non informé" de notre démarche. Mais ils apportent également des informations complémentaires en retenant d'autres variables fonctionnelles cohérentes avec l'application.

1.5 Les k-plus proches voisins pour des données fonctionnelles

Ce travail est le fruit d'une collaboration avec Magalie Fromont (maître de conférence à l'Université de Rennes 2) et porte sur le thème de la classification de données fonctionnelles.

1.5.1 Le problème

La classification binaire consiste à déterminer, au moyen d'un jeu de données $\mathcal{L} = \{(X_i, Y_i)_{1 \leq i \leq n}\}$ où $(X_i, Y_i) \in \mathcal{X} \times \{0; 1\}$, une fonction appelée classifieur qui permet d'associer à chaque observation de l'espace $\mathcal{X}$ une réponse dans $\{0; 1\}$.
Une méthode classique et usuelle pour déterminer des classifieurs consiste à utiliser la méthode des k-plus proches voisins. Cette méthode a largement été étudiée dans le cas de données multivariées, autrement dit lorsque $\mathcal{X} = \mathbb{R}^d$.
Cependant, aujourd'hui de nombreuses applications font appel à des données de type fonctionnel auxquelles on souhaite pouvoir appliquer la méthode des k-plus proches voisins. Comment faire pour obtenir une procédure efficace ?

1.5.2 Le cadre

On dispose d'un échantillon $\mathcal{L} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ tel que les observations X_i appartiennent à un espace fonctionnel $\mathcal{X}$ supposé séparable. Le but est de procéder à la classification supervisée de ces données fonctionnelles.
Notre travail repose sur l'approche développée par Biau, Bunea et Wegkamp dans [12]. Il s'agit de projeter ces données sur une base de l'espace $\mathcal{X}$ et, pour $d \in \mathbb{N}^*$, à considérer les variables explicatives X_i^d qui sont les d premiers coefficients de la projection de la variables X_i.
Par ce biais, nous nous ramenons au cadre multivarié, dans lequel il est alors envisageable de procéder à une classification par la règle des k-plus proches voisins qui doit intégrer la sélection simultanée de la dimension d de l'espace de projection et du nombre k de voisins.

Afin de procéder à cette double phase de sélection et lutter contre le "fléau de la dimension", Biau *et al.* ont proposé de pénaliser la procédure des k-plus proches voisins par un terme en $log(d)/m$ où m est le nombre d'observations utilisées lors de la phase de validation de la procédure des k-plus proches voisins.
Ils ont alors obtenu une inégalité de type "oracle". Cependant, dans la pratique, ils ont utilisé une version non pénalisée de cette procédure.

Notre but est de justifier théoriquement le fait que considérer une version non pénalisée est efficace et que l'ajout d'un léger terme de pénalité, qui permet de stabiliser la procédure, n'altère pas les performances. En outre, un travail sur des données réelles et simulées permet de donner un ordre de grandeur de la pénalisation à mettre en œuvre.

1.5.3 Les résultats

Dans ce paragraphe, nous utilisons les notations suivantes.
L'échantillon $\mathcal{L}$ est scindé en un échantillon d'apprentissage $\mathcal{L}_a = \{(X_i, Y_i),\ i \in \mathcal{T}_l\}$ de taille

l et un échantillon de validation $\mathcal{L}_v = \{(X_i, Y_i),\ i \in \mathcal{V}_m\}$ de taille m tel que $m + l = n$. Pour chaque $k \in \{1, \dots, l\}$ et chaque $d \in \mathcal{D} \subset \mathbb{N}^*$, on note $\hat{p}_{l,k,d}$ la règle des k-plus proches voisins construite à partir de l'échantillon $\{(X_i^d, Y_i),\ i \in \mathcal{T}_l\}$. Autrement dit, $\hat{p}_{l,k,d}$ est défini, pour $x \in \mathbb{R}^d$ par :

$$\hat{p}_{l,k,d}(x) = \begin{cases} 0 \text{ si } \sum_{i=1}^{k} 1\!\!\mathrm{I}_{Y_{(i)}(x)=0} \geq \sum_{i=1}^{k} 1\!\!\mathrm{I}_{Y_{(i)}(x)=1}, \\ \\ 1 \text{ sinon} \end{cases}$$

avec $(X_{(1)}^d, Y_{(1)}), \dots, (X_{(l)}^d, Y_{(l)})$ les variables de l'échantillon $\mathcal{L}_a$ réordonnées selon l'ordre croissant des distances euclidiennes $\|X_i^d - x\|$.
Par ailleurs, pour une observation x de l'espace fonctionnel $\mathcal{X}$, on introduit le classifieur $\hat{\phi}_{l,k,d}(x) = \hat{p}_{l,k,d}(x^d)$.
On définit alors la dimension finale $\hat{d}$ et le nombre de voisins associé $\hat{k}$ par :

$$(\hat{d}, \hat{k}) = \underset{k \in \{1,\dots,l\},\ d \in \mathcal{D}}{argmin} \left(\frac{1}{m} \sum_{i \in \mathcal{T}_m} 1\!\!\mathrm{I}_{\hat{\phi}_{l,k,d}(X_i) \neq Y_i} + pen(d) \right),$$

où pen est une fonction de pénalité, positive ou nulle, à ajuster.

L'estimateur final est alors pour $x \in \mathcal{X}$, $\hat{\phi}_n(x) = \hat{\phi}_{l,\hat{k},\hat{d}}(x)$.

On obtient un résultat théorique sur les performances de l'estimateur final $\hat{\phi}_n$ relativement au classifieur de Bayes noté ϕ^* :
Soit $n \geq 2$, $\theta \geq 1$ et $pen(d)$ un terme de pénalité positif ou nul, alors pour $\beta > 0$, et sous une hypothèse de marge, on a :

$$(1-\beta)\mathbb{E}[L(\hat{\phi}_n) - L(\phi^*)|\mathcal{L}_a] \leq (1+\beta) \inf_{k \in \{1,\dots,l\},\ d \in \mathcal{D}} \left\{ \left(L(\hat{\phi}_{l,k,d}) - L(\phi^*) \right) + \mathrm{pen}(d) \right\} + R. \tag{1.3}$$

où R est un terme de reste.

Si dans ce résultat on pose $pen(d) = 0$, ceci permet de prouver l'efficacité de la procédure non pénalisée des k-plus proches voisins lorsque l'on dispose de données fonctionnelles.
Par ailleurs, on note que l'ajout d'une pénalité ne vient pas altérer les performances de l'estimateur final dans la mesure où cette dernière ne perturbe pas le terme $L(\hat{\phi}_{l,k,d}) - L(\phi^*)$ dont on ne connaît qu'une évaluation partielle grâce aux travaux de Györfi.

De plus, la procédure des k-plus proches voisins présente une instabilité qui résulte d'une part de la sélection de la dimension $\hat{d}$ et d'autre part du découpage en deux de l'échantillon initial. On envisage ensuite une pénalité qui, correctement calibrée, peut gommer cette variabilité comme l'illustre le travail effectué tant sur des données réelles que sur des données simulées.

Ainsi, bien que théoriquement la pénalisation n'engendre pas d'amélioration en termes de risque, en pratique la prise en compte d'un terme de pénalité peut s'avérer appréciable dans le sens où ce dernier apporte une stabilisation des résultats dans tous les exemples étudiés et en améliore parfois les performances.

1.6 Perspectives

Les travaux menés dans cette thèse peuvent être saisis au travers de deux axes. Le premier est thématique, on peut dégager deux directions :

- la sélection de variables dans un contexte de régression et de classification supervisée ;
- le traitement de données fonctionnelles dans le cadre de la classification supervisée.

D'autre part, un second axe concerne le type de ces travaux qui sont tant théoriques (**Chapitres 2** et **4**) qu'appliqués et méthodologiques (**Chapitre 3**).

Les prolongements et les perspectives liés à ces travaux sont, par conséquent, variés. Quelques-uns sont esquissés ci-dessous :

- Le **Chapitre 2** fait intervenir une fonction de pénalité dépendant de deux paramètres α et β. Ils sont déterminés, dans l'étude proposée, au moyen d'un échantillon test. Une perspective consiste à "calibrer", à partir des données, ces deux paramètres en utilisant une méthode basée sur la détection de changement de pente du contraste empirique, à l'image de l'"heuristique de pente" (cf. Lebarbier [62]).
- Dans le **Chapitre 2**, une stratégie de mise en œuvre pratique de la procédure de sélection de variables, développée dans ce même chapitre, est proposée et son application est illustrée sur un exemple. Cette voie est à explorer plus en profondeur en testant d'autres stratégies ou des variantes en tentant de la justifier théoriquement dans des contextes restreints.
- Le travail mené dans le **Chapitre 3** se situe dans un contexte industriel et concerne le thème de la sélection de variables tout en traitant des données fonctionnelles. Il s'inscrit dans la continuité du travail, mené par Ansaldi [4], sur l'objectivation d'une prestation. Il se poursuit par une thèse CIFRE, qui débute actuellement chez Renault, sur les questions d'objectivation simultanée de plusieurs prestations. Ainsi, le travail réalisé est une contribution qui s'inscrit dans le programme de recherche de l'industriel automobile.
- Dans la **section 3.5** du **Chapitre 3**, on tente de faire un lien entre des parties théoriques et appliquées de l'étude relative au premier thème : la sélection de variables. On peut penser à développer des liens semblables pour le second thème : le traitement de données fonctionnelles, notamment en considérant le problème du choix de la base permettant de représenter les objets d'intérêt. En effet, dans le **Chapitre 3**, pour chacune des variables fonctionnelles, on détermine la base de représentation commune en la choisissant, parmi les bases des espaces d'approximation, à l'aide d'un critère indépendant de l'objectif de discrimination. Dans le **Chapitre 4**, la base de représentation privilégiée est celle de Fourier, mais de récents travaux dûs à Berlinet *et al.* [10] montrent que l'on peut également considérer les bases d'ondelettes. Mais quelle que soit la base considérée, l'introduction d'un "léger" terme de pénalité a le mérite de stabiliser les résultats tout en préservant de bonnes performances. Des prolongements consistent d'une part à affiner la détermination de la fonction de pénalité à mettre en jeu, et d'autre part, à envisager le recours à d'autres types de base.

- Concernant le **Chapitre 4** consacré à la méthode de classification des k-plus proches voisins, voici deux pistes un peu plus spéculatives.
  Sur le plan appliqué, on peut envisager de mettre en compétition diverses méthodes comme les Support Vector Machines (cf. Rossi et Villa [82]) ou les Réseaux de Neurones (cf. Rossi et Conan-Guez [81]), qui sont des méthodes disponibles pour la classification de données fonctionnelles. Bien que plus délicates à étudier théoriquement, on peut également penser à des méthodes basées sur le rééchantillonnage, à l'instar du bagging (cf. Breiman [18]) et des Random Forests (cf. Breiman [19]), par exemple.
  Par ailleurs, sur le plan théorique, on peut s'intéresser à l'étude de l'estimateur des k-plus proches voisins pour des données fonctionnelles, en examinant son optimalité au sens minimax.

La suite de cette thèse est composée de trois chapitres qui font ou feront très prochainement l'objet d'articles. Les chapitres 2 et 4 sont en anglais, le chapitre 3 en français. Chacun d'eux est précédé d'une courte présentation en français.

Chapitre 2

Variable Selection through CART

Sommaire

Ce chapitre présente un travail réalisé en collaboration avec Marie Sauvé. Il fera prochainement l'objet d'un article.

Ce chapitre aborde, de façon théorique, 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 complètent ces éléments théoriques.

2.1 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 :

$$Y = \sum_{i=1}^{p} \beta_i X^i + \varepsilon = X\beta + \varepsilon$$

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 [56]) 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 [85]) 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 [56]) 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 [56]) 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 [42] 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.* [36].
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 [16]), 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 :

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

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 closed to the optimal one, i.e. :

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

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 2.2** recalls the different steps of the CART algorithm and defines some notations. The **Sections 2.3** and **2.4** present the results obtain in the regression and classification frameworks. In the **Section 2.5**, 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 2.6** and **2.7** collect lemmas and proofs.

2.2 Preliminaries

2.2.1 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.* [20]. 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 [44] 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 2.2.1
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 $(M2)$. $\square$

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.* [20]. 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.

2.2.2 The context

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

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

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

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

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

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

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 $(\varepsilon_1,...,\varepsilon_n)$ is a sample such that $\mathbb{E}\left[\varepsilon_i|X_i\right]=0$ and the ε_i satisfy the following assumption on the Laplace transform:

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

with σ and ρ two positive constants.

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

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

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

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

2.3 Regression

Let us consider the regression framework where, for a given $i \in \{1,\ldots,n\}$, ε_i is centered conditionally to X_i and satisfies the assumption (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.

The two following subsections give 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.

2.3.1 Variable selection via $(M1)$

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 closed 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 2.3.1
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:*

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

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.* [20] in their pruning algorithm and validated by Gey and Nédélec [44] for the Gaussian regression case. This proposition validates the CART pruning penalty proposed by Breiman *et al.* [20] 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 2.3.1
If $\rho = 0$, our assumption (2.3) on the Laplace transform becomes :

$$\text{for any } \lambda \in \mathbb{R},\ log\mathbb{E}\left[e^{\lambda\varepsilon_i} \middle| X_i\right] \leq \frac{\sigma^2\lambda^2}{2}.$$

The random variables ε_i are said to be sub-gaussian conditionally to X_i. In this case, 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

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

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.

2.3.2 Variable selection via $(M2)$

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 $\{X_i;\ (X_i,Y_i) \in \mathcal{L}_1\}$. With the same notations as in the **Subsection 2.3.1**, we find nearly the same result as in the $(M1)$ situation.

Proposition 2.3.2
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* $\geq 1 - e^{-\xi}\Sigma - \frac{c}{n_1 log(n_1)} 1\!\!I_{\rho\neq 0}$,

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

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 2.3.2
If $\rho = 0$, the random variables ε_i are sub-gaussian conditionally to X_i. In this case, 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.

2.3.3 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 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.* [20].
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 2.3.3

- *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 $\geq 1 - e^{-\xi} - 1\!\!I_{\rho \neq 0} \frac{R^2}{2(\sigma^2+\rho R)} \frac{1}{n_2^{1-log2}}$,

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

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

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

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

□

Remark 2.3.3
If $\rho = 0$, by integrating with respect to ξ, we get for the two methods $(M1)$ and $(M2)$ that: for any $\eta \in (0,1)$,

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

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. □

2.4 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 closed 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 [86] :

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

This section follows the same organization as **Section 2.3** and uses the same notations.

2.4.1 Variable selection via $(M1)$

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 2.4.1
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:

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

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

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.

2.4.2 Variable selection via $(M2)$

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 2.4.2
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:
$\forall M \in \mathcal{P}(\Lambda), \quad \forall T \preceq T_{max}^{(M)}$

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

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

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

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

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.

2.4.3 Final selection

With the same notations as in the **Subsection 2.3.3**, we validate the final selection for the two methods.
The following proposition is expressed for the $(M1)$ method.

Proposition 2.4.3
For any $\eta \in (0,1)$, we have:

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

□

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 2.4.1** and **2.4.3** involve the same expected loss, we can compare the final estimator $\tilde{\tilde{s}}$ with the entire collection of models :

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

2.5 Simulations

The aim of this section is twice. On the one hand, we illustrate by an example the theoretical procedure, described in the **Section 2.1**.
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 [20] p. 237), is composed of $p = 10$ explanatory variables $X^1, \dots, X^{10}$ such that :

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

and of the explained variable Y given by :

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

where the unobservable random variable ε is independent of $X^1, \dots, X^{10}$ 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 2.1** 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

TAB. 2.1 – 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 $X = (X^1, \dots, X^{10})$.

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 2.2**.
At the intersection of the row β and the column α appears the set of variables associated with $\tilde{s}$.

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

TAB. 2.2 – 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 2.1**).
As the expected final set $\{1, 2, 5, 3, 6, 4, 7\}$ appears in the **Table 2.2**, obviously, the final step

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

TAB. 2.3 – In this table, we see the results of the final model selection.

of the procedure, whose results are given by the **Table 2.3**, returns the "good" set.

The **Table 2.3** 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 2.3**, 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 [76] 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 :

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

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 2.2**, the set

$\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ is replaced by $\{1, 2, 5, 6, 3, 7, 4\}$.

So, we notice that the practical procedure works correctly on the simulated example. Its main advantage is that this procedure is a practical one since it seems to give good results and it is computationally efficient.
In other respects, this procedure does not have the drawback of the variable selection induced by VI. Let us introduce a new simulated example by adding another explanatory variable denoted by X^{11} and defined by $X^{11} = X^2 + \varepsilon'$, we obtain the expected Breiman's VI ranking given by the **Table 2.4**.

Variable	X^1	X^2	X^{11}	X^5	X^3	X^6	X^4	X^7	X^8	X^9	X^{10}
Rank	1	2	3	4	6	5	8	7	9	10	11

TAB. 2.4 – The Breiman's Variable Importance Ranking for the 11 explanatory variables.

If we perform the practical and theoretical procedures, the final selected set is $\{1, 2, 5, 6, 3, 7, 4\}$ which is the expected one, since the associated family $\mathcal{P}^*$ is :

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

The variable X^{11} does not appear in this family since the addition of X^{11} to the subset $\{X^1, X^2\}$ does not have significant incremental influence on the response.
If we perform variable selection by keeping the variables whose importance (VI) is greater than an arbitrary threshold, the variables X^1, X^2 and X^{11} are kept and so, by this way, we do not select the smallest subset of variables.

In conclusion, the theoretical and practical procedures give both good results in practice. However, the second one is better since it is a computationally efficient method.
Moreover, in the **Section 3.5**, we show that for real data, the proposed procedure leads to relevant results too.

2.6 Appendix

This section presents some lemmas which are useful in the proofs of the propositions of the **Sections 2.3** and **2.4**. The lemmas 1 to 3 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 2.6.1 is a concentration inequality due to Talagrand. This type of inequality allows to know how a random variable behaves around its expectation.

Lemma 2.6.1 (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 2.6.1: see Massart [15] □

The lemma 2.6.2 allows to pass from local maximal inequalities to a global one.

Lemma 2.6.2 (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:
$\forall \Phi : \mathbb{R} \rightarrow \mathbb{R}^+$ *such that :*

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

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

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

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

□

Proof of the lemma 2.6.2: see Massart and Nédélec [44], section: "Appendix: Maximal inequalities", lemma 5.5. □

Thanks to the lemma 2.6.3, we see that the Hold-Out is an adaptative selection procedure for classification.

Lemma 2.6.3 (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:

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

Then, $\forall\theta \in (0,1)$, one has:

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

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

Proof of the lemma 2.6.3: see [70], Chapter: "Statistical Learning", Section: "Advanced model selection problems". □

The lemmas 2.6.4 and 2.6.5 are concentration inequalities for a sum of squared random variables whose Laplace transform are controlled.

Lemma 2.6.4
Let $\varepsilon_1, \ldots, \varepsilon_n$ n independent and identically distributed random variables satisfying:

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

Let m_0 a partition of $\{1,\ldots,n\}$ such that, $\forall J \in m_0$, $|J| \ge N_{min}$.
We consider the collection $\mathcal{M}$ of all partitions of $\{1,\ldots,n\}$ 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 $\Omega_\delta = \left\{\forall J\in m_0;\ \left|\sum_{i\in J}\varepsilon_i\right| \le \delta\sigma^2|J|\right\}$.
Then for any $m\in\mathcal{M}$ and any $x>0$,

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

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 2.6.4:
Let $m \in \mathcal{M}$ and denote, for any $J \in m$,

$$Z_J = \frac{\left(\sum_{i\in J}\varepsilon_i\right)^2}{|J|} \wedge \left(\delta^2\sigma^4|J|\right).$$

$(Z_J)_{J\in m}$ are independent random variables.
After calculating their moments, we deduce from Bernstein inequality that, for any $x > 0$,

$$\mathbb{P}\left(\sum_{J\in m} Z_J \geq \sigma^2|m| + 4\sigma^2(1+b\delta)\sqrt{2|m|x} + 2\sigma^2(1+b\delta)\right)x \leq e^{-x}.$$

As $\sum_{J\in m} Z_J = \chi_m^2$ on the set Ω_δ, we get that for any $x > 0$,

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

Thanks to the assumption on the Laplace transform of the ε_i, we have, for any $J \in m_0$,

$$\mathbb{P}\left(\left|\sum_{i\in J}\varepsilon_i\right| \geq \delta\sigma^2|J|\right) \leq 2\exp\left(\frac{-\delta^2\sigma^2|J|}{2(1+b\delta)}\right).$$

As $|J| \geq N_{min}$, we obtain

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

□

Lemma 2.6.5
Let $\varepsilon_1, \ldots, \varepsilon_n$ n independent and identically distributed random variables satisfying:

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

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

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

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 $x > 0$,

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

and

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

□

Proof of the lemma 2.6.5: The proof is exactly the same of the preceding one. The only difference is that the set Ω_δ is smaller. □

The lemmas 2.6.6 and 2.6.7 give the expression of the weights needed in the model selection procedure.

Lemma 2.6.6
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*

$$\sum_{M\in\mathcal{P}(\Lambda)} \sum_{T\preceq T_{max}^{(M)}} e^{-x_{M,T}} \leq \Sigma(a,b) \tag{2.4}$$

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 2.6.6:
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

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

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

$$\begin{aligned}\Sigma(\mathcal{L}_1) &\leq \sum_{k=1}^{p}\binom{p}{k}\sum_{D\geq1}\frac{1}{D}e^{-(x(k,D)-(2log2)D)},\\ &\leq \sum_{k=1}^{p}\left(\frac{ep}{k}\right)^k\sum_{D\geq1}\frac{1}{D}e^{-(x(k,D)-(2log2)D)}.\end{aligned}$$

Taking $x(k,D) = aD + \alpha(k)$ with $a > 2log(2)$ an absolute constant, we have

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

Thus, 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 (2.4). □

Lemma 2.6.7
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*

$$\sum_{M\in\mathcal{P}(\Lambda)} \sum_{T\in\mathcal{M}_{n_1,M}} e^{-x_{M,T}} \leq \Sigma^{'}(a,b) \tag{2.5}$$

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 2.6.7:
We are looking for weights $x_{M,T}$ such that the sum

$$\Sigma(\{X_i;\ (X_i,Y_i)\in\mathcal{L}_1\}) = \sum_{M\in\mathcal{P}(\Lambda)} \sum_{T\in\mathcal{M}_{n_1,M}} e^{-x_{M,T}}$$

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

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

Since the Vapnik-Chervonenkis dimension of $\mathcal{S}p_M$ is $|M|+1$, it follows from lemma 2 in [44] that

$$|\{T\in\mathcal{M}_{n_1,M};\ |T| = D\}| \leq \left(\frac{n_1 e}{|M|+1}\right)^{D(|M|+1)}.$$

We get

$$\begin{aligned}\Sigma(\{X_i;\ (X_i,Y_i)\in\mathcal{L}_1\}) &\leq \sum_{k=1}^{p}\binom{p}{k}\sum_{D\geq 1} e^{D\left[(k+1)\left(1+log\left(\frac{n_1}{k+1}\right)\right)\right]-x(k,D)},\\ &\leq \sum_{k=1}^{p}\left(\frac{ep}{k}\right)^k\sum_{D\geq 1} e^{D\left[(k+1)\left(1+log\left(\frac{n_1}{k+1}\right)\right)\right]-x(k,D)}.\end{aligned}$$

Taking $x(k,D) = D\left[a+(k+1)\left(1+log\left(\frac{n_1}{k+1}\right)\right)\right]+\alpha(k)$ with $a>0$ an absolute constant, we have

$$\Sigma(\{X_i;\ (X_i,Y_i)\in\mathcal{L}_1\}) \leq \left(\sum_{k=1}^{p} e^{-\left(\alpha(k)-k\left(1+log\left(\frac{p}{k}\right)\right)\right)}\right)\left(\sum_{D\geq 1} e^{-aD}\right).$$

Thus, taking $x(k,D) = D\left[a+(k+1)\left(1+log\left(\frac{n_1}{k+1}\right)\right)\right]+bk\left(1+log\left(\frac{p}{k}\right)\right)$ with $a>0$ and $b>1$ two absolute constants, we have

$$\Sigma(\{X_i;\ (X_i,Y_i)\in\mathcal{L}_1\}) \leq \left(\sum_{k\geq 1} e^{-(b-1)k}\right)\left(\sum_{D\geq 1} e^{-aD}\right) = \Sigma'(a,b).$$

Thus the weights $x_{M,T} = |T|\left[a+(|M|+1)\left(1+log\left(\frac{n_1}{|M|+1}\right)\right)\right]+b|M|\left(1+log\left(\frac{p}{|M|}\right)\right)$, with $a>0$ and $b>1$ two absolute constants, satisfy (2.5). □

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

Lemma 2.6.8
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:

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

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,*

- $\forall \sigma \geq w(\sigma_T), \quad \sqrt{n}\mathbb{E}\left[\underset{u \in S_T,\ d(u,v) \leq \sigma}{sup} |\bar{\gamma}_n(u) - \bar{\gamma}_n(v)|\right] \leq \phi_T(\sigma),$

  with σ_T the positive solution of $\phi_T(w(x)) = \sqrt{n}x^2$.

- $\sigma_T^2 \leq \frac{K_3^2|T|}{nh}$.

□

PROOF OF LEMMA 2.6.8:
In classification, it is known that:

$$\forall u \in S_T, \quad hd^2(s,u) \leq l(s,u).$$

So:

$$\underset{u \in S_T,\ l(s,u) \leq \epsilon^2}{sup} d(s,u) \leq \frac{\epsilon}{\sqrt{h}}.$$

The existence of the function ϕ_T has been proved by Massart and Nédélec in [71]. They shown that:

$$\phi_T(x) = Kx\sqrt{\mathbb{E}\left[\mathcal{H}_{\mathcal{A}}\right] \vee 1}$$

where:

$$\begin{cases} \mathcal{A} \text{ is defined by } S_T = \{1\!\!1_A, A \in \mathcal{A}\} \\ \mathcal{H}_{\mathcal{A}} = log(\#\{A \cap \{X_1, ..., X_n\}, A \in \mathcal{A}\}) \end{cases}$$

They also proved that:

$$\sigma_T^2 = \frac{K^2(1 \vee \mathbb{E}\left[\mathcal{H}_{\mathcal{A}}\right])}{nh}.$$

If $\mathcal{A}$ is a VC-class of dimension V, with Sauer's lemma, we obtain:

$$\mathcal{H}_{\mathcal{A}} \leq V\left(1 + log\left(\frac{n}{V}\right)\right).$$

But with the structure of S_T, we can do better.

$$S_T = \{1\!\!1_A, A \in \mathcal{P}(\mathcal{F})\} \text{ with } \mathcal{F} = \{\chi_1, ..., \chi_{|T|}\}$$

where χ_j represents the partition of the observations space associated with the leaf j of T.

Then:

$$Card(\mathcal{F}) = |T| \quad \text{and} \quad \mathcal{H}_{\mathcal{A}} \leq log(2^{|T|}).$$

Thus:

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

□

Lemma 2.6.9
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 2.6.8. If $2x \geq \frac{K_3\sqrt{|T|}}{\sqrt{n}h}$, then:

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

□

Proof of lemma 2.6.9:
We consider the function $\phi = \frac{\phi_T}{\sqrt{n}}$ and the parameter σ_T defined in the lemma (2.6.8).
By application of the lemma (2.6.2) at the process $\bar{\gamma}_n$, with the function ϕ defined before and thanks to the properties of σ_T and ϕ_T, we get:
if $2x \geq \frac{K_3\sqrt{|T|}}{\sqrt{n}h}$:

$$\begin{aligned}
\mathbb{E}\left[\underset{u \in S_T}{sup} \frac{|\bar{\gamma}_n(u) - \bar{\gamma}_n(v)|}{d^2(u,v) + (2x)^2}\right] &\leq \frac{4}{(2x)^2}\frac{\phi_T(2x)}{\sqrt{n}}, \\
&\leq \frac{2}{x\sqrt{n}}\frac{\phi_T\left(\frac{\sigma_T}{\sqrt{h}}\right)}{\frac{\sigma_T}{\sqrt{h}}}, \\
&\leq \frac{2K_3\sqrt{|T|}}{x\sqrt{n}}.
\end{aligned}$$

□

2.7 Proofs

2.7.1 Regression

Proof of the proposition 2.3.1:
Let $a > 2log2$, $b > 1$, $\theta \in (0,1)$ and $K > 2 - \theta$ four constants.
Let $\delta > 0$.

Let us denote

$$\begin{aligned}
s_{M,T} &= \underset{u \in S_{M,T}}{argmin} \|s - u\|_{n2}^2 \\
\varepsilon_{M,T} &= \underset{u \in S_{M,T}}{argmin} \|\varepsilon - u\|_{n2}^2
\end{aligned}$$

Following the proof of theorem 2 in [16], we get

$$(1-\theta)\|s - \tilde{s}\|_{n2}^2 = \Delta_{\widehat{M,T}} + \underset{(M,T)}{inf} R_{M,T} \tag{2.6}$$

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

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

Thanks to lemma 2.6.4, we get that for any (M,T) and any $x > 0$

$$\begin{aligned} \mathbb{P}\Bigg(\|\varepsilon_{M,T}\|_{n_2}^2 \mathbb{1}_{\Omega_{\delta,M}} &\geq \frac{\sigma^2}{n_2}|T| + 4\frac{\sigma^2}{n_2}(1+\rho\delta)\sqrt{2|T|x} + 2\frac{\sigma^2}{n_2}(1+\rho\delta)x \ \Big| \mathcal{L}_1 \ and \ \{X_i; \ (X_i,Y_i) \in \mathcal{L}_2\} \Bigg) \\ &\leq e^{-x} \end{aligned} \tag{2.7}$$

and

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

Denoting $\Omega_\delta = \bigcap_M \Omega_{\delta,M}$, we have

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

To control $- < \varepsilon, s - s_{M,T} >_{n_2}$, we calculate its Laplace transform. Thanks to assumption (A) and $\|s\|_\infty \leq R$, we have for any (M,T) and any $\lambda \in \left(0; \frac{n_2}{2\rho R}\right)$,

$$log \ \mathbb{E}\left[e^{-\lambda < \varepsilon, s - s_{M,T} >_{n_2}} \ \Big| \mathcal{L}_1 \text{ and } \{X_i; \ (X_i,Y_i) \in \mathcal{L}_2\} \right] \ \leq \ \frac{\lambda^2\sigma^2\|s - s_{M,T}\|_{n_2}^2}{2n_2\left(1 - \lambda\frac{2\rho R}{n_2}\right)}$$

Thus, for any (M,T) and any $x > 0$

$$\begin{aligned} \mathbb{P}\Big(- < \varepsilon, s - s_{M,T} >_{n_2} &\geq \frac{\sigma}{\sqrt{n_2}}\|s - s_{M,T}\|_{n_2}\sqrt{2x} + \frac{2\rho R}{n_2}x \ \Big| \mathcal{L}_1 \text{ and } \{X_i; \ (X_i,Y_i) \in \mathcal{L}_2\} \Big) \\ &\leq e^{-x} \end{aligned} \tag{2.8}$$

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 2.6.6, and summing all inequalities (2.7) and (2.8) with respect to (M,T), we derive a set E_ξ such that

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

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

$$\begin{aligned}\Delta_{M,T} &\leq (2-\theta)\frac{\sigma^2}{n_2}|T| + 4(2-\theta)\frac{\sigma^2}{n_2}(1+\rho\delta)\sqrt{2|T|(x_{M,T}+\xi)}\\ &+2(2-\theta)\frac{\sigma^2}{n_2}(1+\rho\delta)(x_{M,T}+\xi)\\ &+2\frac{\sigma}{\sqrt{n_2}}\|s-s_{M,T}\|_{n_2}\sqrt{2(x_{M,T}+\xi)} + 4\frac{\rho R}{n_2}(x_{M,T}+\xi)\\ &-\theta\|s-s_{M,T}\|^2_{n_2} - pen(M,T)\end{aligned}$$

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

Using the inequalities $2\frac{\sigma}{\sqrt{n_2}}\|s-s_{M,T}\|_{n_2}\sqrt{2(x_{M,T}+\xi)} \leq \theta\|s-s_{M,T}\|^2_{n_2} + \frac{2}{\theta}\frac{\sigma^2}{n_2}(x_{M,T}+\xi)$ and $2\sqrt{|T|(x_{M,T}+\xi)} \leq \eta|T| + \eta^{-1}(x_{M,T}+\xi)$ with $\eta = \frac{K+\theta-2}{2-\theta}\frac{1}{2\sqrt{2}(1+\rho\delta)} > 0$, we derive that on the set $E_\xi\bigcap\Omega_\delta$, for any (M,T),

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

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

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

we get that, on the set $E_\xi\bigcap\Omega_\delta$,

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

Thus on the set E_ξ

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

Integrating with respect to ξ, we derive

$$\mathbb{E}\left[\Delta_{\widehat{M,T}}1\!\!1_{\Omega_\delta}\Big|\mathcal{L}_1\right] \leq 2\left(2(1+\rho\delta)(2-\theta)\left(1+\frac{4(1+\rho\delta)(2-\theta)}{K+\theta-2}\right)+\frac{2}{\theta}+4\frac{\rho}{\sigma^2}R\right)\frac{\sigma^2}{n_2}\Sigma(a,b) \tag{2.9}$$

We are going now to control $\mathbb{E}\left[\underset{(M,T)}{inf}\ R_{M,T}1\!\!1_{\Omega_\delta}\Big|\mathcal{L}_1\right]$.

In the same way we deduced (2.8) from assumption (A), we get that for any (M,T) and any $x>0$

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

Thus we derive a set F_ξ such that

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

- 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 $R_{M,T}$ that on the set F_ξ, for any (M,T),

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

And

$$\begin{aligned}
\mathbb{E}\left[\underset{(M,T)}{inf}\; R_{M,T}\mathbb{1}_{\Omega_\delta}\Big|\mathcal{L}_1\right] &\leq 2\underset{(M,T)}{inf}\left\{\mathbb{E}\left[\|s-s_{M,T}\|^2_{n_2}\Big|\mathcal{L}_1\right]+pen(M,T)\right\} \qquad (2.10)\\
&\quad+\left(2+4\frac{\rho}{\sigma^2}R\right)\frac{\sigma^2}{n_2}\Sigma(a,b)
\end{aligned}$$

We conclude from (2.6), (2.9) and (2.10) that

$$\begin{aligned}
(1-\theta)\mathbb{E}\left[\|s-\tilde{s}\|^2_{n_2}\mathbb{1}_{\Omega_\delta}\Big|\mathcal{L}_1\right] \leq\; & 2\underset{(M,T)}{inf}\left\{\mathbb{E}\left[\|s-s_{M,T}\|^2_{n_2}\Big|\mathcal{L}_1\right]+pen(M,T)\right\}\\
&+\left(4(1+\rho\delta)(2-\theta)\left(1+\tfrac{4(1+\rho\delta)(2-\theta)}{K+\theta-2}\right)+\tfrac{6}{\theta}+12\tfrac{\rho}{\sigma^2}R\right)\tfrac{\sigma^2}{n_2}\Sigma(a,b)
\end{aligned}$$

It remains to control $\mathbb{E}\left[\|s-\tilde{s}\|^2_{n_2}\mathbb{1}_{\Omega_\delta^c}\Big|\mathcal{L}_1\right]$.

$$\begin{aligned}
\mathbb{E}\left[\|s-\tilde{s}\|^2_{n_2}\mathbb{1}_{\Omega_\delta^c}\Big|\mathcal{L}_1\right] &= \mathbb{E}\left[\|s-s_{\widehat{M,T}}\|^2_{n_2}\mathbb{1}_{\Omega_\delta^c}\Big|\mathcal{L}_1\right]+\mathbb{E}\left[\|\varepsilon_{\widehat{M,T}}\|^2_{n_2}\mathbb{1}_{\Omega_\delta^c}\Big|\mathcal{L}_1\right]\\
&\leq \mathbb{E}\left[\|s\|^2_{n_2}\mathbb{1}_{\Omega_\delta^c}\Big|\mathcal{L}_1\right]+\sum_M \mathbb{E}\left[\|\varepsilon_{M,T^{(M)}_{max}}\|^2_{n_2}\mathbb{1}_{\Omega_\delta^c}\Big|\mathcal{L}_1\right]\\
&\leq R^2\mathbb{P}\left(\Omega_\delta^c\Big|\mathcal{L}_1\right)+\sum_M\sqrt{\mathbb{E}\left[\|\varepsilon_{M,T^{(M)}_{max}}\|^4_{n_2}\Big|\mathcal{L}_1\right]}\sqrt{\mathbb{P}\left(\Omega_\delta^c\Big|\mathcal{L}_1\right)}
\end{aligned}$$

As

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

where $C^2(\rho,\sigma)$ is a constant which overestimate $\mathbb{E}\left[\varepsilon_i^4\right]$, we get that

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

Let us recall that

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

For $p \leq log(n_2)$ and $N_{min} \geq \frac{12(1+\rho\delta)}{\delta^2\sigma^2} log(n_2)$,

- $2^p \sqrt{\mathbb{P}\left(\Omega_\delta^c \Big| \mathcal{L}_1\right)} \leq \frac{\delta\sigma}{\sqrt{1+\rho\delta}} \frac{1}{\sqrt{6}} \frac{1}{n_2\sqrt{log(n_2)}} e^{\left(\frac{3}{2}(1+log2)-3\right)log(n_2)} \leq \frac{\delta\sigma}{\sqrt{1+\rho\delta}} \frac{1}{\sqrt{6}} \frac{1}{n_2\sqrt{log(n_2)}}$
- $\mathbb{P}\left(\Omega_\delta^c \Big| \mathcal{L}_1\right) \leq \frac{\delta^2\sigma^2}{1+\rho\delta} \frac{1}{6n_2^4 log(n_2)} e^{(log2-1)log(n_2)} \leq \frac{\delta^2\sigma^2}{1+\rho\delta} \frac{1}{6n_2^4 log(n_2)}$
- $\frac{\sigma^2}{N_{min}} + \frac{C(\rho,\sigma)}{\sqrt{n_2}N_{min}} + \frac{\sqrt{3}\sigma^2}{\sqrt{n_2 N_{min}}} \leq \frac{\delta\sigma^3}{1+\rho\delta} \left(\frac{\delta\sigma}{12} + \frac{C(\rho,\sigma)\delta}{12\sigma} + \frac{\sqrt{1+\rho\delta}}{2}\right) \frac{1}{log(n_2)}$

It follows that

$$\mathbb{E}\left[\|s-\tilde{s}\|_{n_2}^2 \mathbb{1}_{\Omega_\delta^c} \Big| \mathcal{L}_1\right] \leq \frac{\delta^2\sigma^2}{1+\rho\delta}\left[\frac{R^2}{6} + \frac{\sigma^2}{\sqrt{1+\rho\delta}}\left(\frac{\delta\sigma}{12\sqrt{6}} + \frac{C(\rho,\sigma)\delta}{12\sqrt{6}\sigma} + \frac{\sqrt{1+\rho\delta}}{2\sqrt{6}}\right)\right] \frac{1}{n_2 log(n_2)}$$

Finally, we have the following result:
Denoting by $\Upsilon = \left[2(1+\rho\delta)(2-\theta)\left(1 + \frac{4(1+\rho\delta)(2-\theta)}{K+\theta-2}\right) + \frac{2}{\theta}\right]$
and taking a penalty which satisfy $\forall\ M \in \mathcal{P}(\Lambda)\ \forall\ T \preceq T_{max}^{(M)}$

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

if $p \leq log(n_2)$ and $N_{min} \geq \frac{12(1+\rho\delta)}{\delta^2\sigma^2} log(n_2)$, we have,

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

We deduce the proposition by taking $\delta = \frac{1}{\rho}$, $K = 2$, $\theta \to 1$, $a \to 2log2$ and $b \to 1$. $\square$

Proof of the proposition 2.3.2:
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:

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

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. So, we use lemma 2.6.5 instead of lemma 2.6.4.

Let us denote, for any $M \in \mathcal{P}(\Lambda)$ and any $T \in \mathcal{M}_{n_1,M}$,

$$\begin{aligned} s_{M,T} &= \underset{u \in S_{M,T}}{argmin} \ \|s - u\|^2_{n_1} \\ \varepsilon_{M,T} &= \underset{u \in S_{M,T}}{argmin} \ \|\varepsilon - u\|^2_{n_1} \end{aligned}$$

Following the proof of theorem 2 in [16], we get

$$(1-\theta)\|s - \tilde{s}\|^2_{n_1} = \Delta_{\widehat{M,T}} + \underset{(M,T)}{inf}\ R_{M,T} \tag{2.11}$$

where the infimum is taken over the (M,T), $M \in \mathcal{P}(\Lambda)$ and $T \preceq T^{(M)}_{max}$, and

$$\begin{aligned} \Delta_{M,T} &= (2-\theta)\|\varepsilon_{M,T}\|^2_{n_1} - 2 < \varepsilon, s - s_{M,T} >_{n_1} - \theta\|s - s_{M,T}\|^2_{n_1} - pen(M,T) \\ R_{M,T} &= \|s - s_{M,T}\|^2_{n_1} - \|\varepsilon_{M,T}\|^2_{n_1} + 2 < \varepsilon, s - s_{M,T} >_{n_1} + pen(M,T) \end{aligned}$$

We are going first to control $\Delta_{\widehat{M,T}}$.
Let us denote

- $\delta = 5\frac{\rho}{\sigma^2} log\left(\frac{n_1}{p}\right)$
- $\Omega = \left\{\forall 1 \leq i \leq n_1 \ \ |\varepsilon_i| \leq \delta\sigma^2\right\} = \left\{\forall 1 \leq i \leq n_1 \ \ |\varepsilon_i| \leq 5\rho log\left(\frac{n_1}{p}\right)\right\}$

Thanks to lemma 2.6.5, we get that for any $M \in \mathcal{P}(\Lambda)$, $T \in \mathcal{M}_{n_1,M}$ and any $x > 0$

$$\begin{aligned} \mathbb{P}\Bigg(\|\varepsilon_{M,T}\|^2_{n_1} \mathbb{1}_\Omega \ &\geq \ \frac{\sigma^2}{n_1}|T| + 4\frac{\sigma^2}{n_1}(1+\rho\delta)\sqrt{2|T|x} + 2\frac{\sigma^2}{n_1}(1+\rho\delta)x \ \Bigg| \{X_i;\ (X_i,Y_i) \in \mathcal{L}_1\}\Bigg) \\ &\leq e^{-x} \end{aligned} \tag{2.12}$$

and

$$\begin{aligned}\mathbb{P}\left(\Omega^c \,\middle|\{X_i;\ (X_i,Y_i)\in\mathcal{L}_1\}\right) &\leq 2n_1 exp\left(\frac{-\delta^2\sigma^2}{2(1+\rho\delta)}\right)\\ &\leq 2n_1 exp\left(\frac{-25\rho^2 log^2\left(\frac{n_1}{p}\right)}{2\left(\sigma^2+5\rho^2 log\left(\frac{n_1}{p}\right)\right)}\right)\end{aligned}$$

Thanks to assumption (A), like in the $(M1)$ case, we get that for any $M\in\mathcal{P}(\Lambda)$, $T\in\mathcal{M}_{n_1,M}$ and any $x>0$

$$\begin{aligned}\mathbb{P}\Big(-<\varepsilon, s-s_{M,T}>_{n_1} &\geq \frac{\sigma}{\sqrt{n_1}}\|s-s_{M,T}\|_{n_1}\sqrt{2x}+\frac{2\rho R}{n_1}x\,\Big|\{X_i;\ (X_i,Y_i)\in\mathcal{L}_1\}\Big)\\ &\leq e^{-x}\end{aligned} \tag{2.13}$$

Setting $x=x_{M,T}+\xi$ with $\xi>0$ and the weights

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

as defined in lemma 2.6.7, and summing all inequalities (2.12) and (2.13) with respect to $M\in\mathcal{P}(\Lambda)$ and $T\in\mathcal{M}_{n_1,M}$, we derive a set E_ξ such that

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

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

$$\begin{aligned}\Delta_{M,T} \leq\ & (2-\theta)\frac{\sigma^2}{n_1}|T|+4(2-\theta)\frac{\sigma^2}{n_1}(1+\rho\delta)\sqrt{2|T|(x_{M,T}+\xi)}\\ &+2(2-\theta)\frac{\sigma^2}{n_1}(1+\rho\delta)(x_{M,T}+\xi)\\ &+2\frac{\sigma}{\sqrt{n_1}}\|s-s_{M,T}\|_{n_1}\sqrt{2(x_{M,T}+\xi)}+4\frac{\rho R}{n_1}(x_{M,T}+\xi)\\ &-\theta\|s-s_{M,T}\|_{n_1}^2-pen(M,T)\end{aligned}$$

where $\Sigma(a,b)=\frac{e^{-a}}{1-e^{-a}}\frac{e^{-(b-1)}}{1-e^{-(b-1)}}$.

Using the inequalities $2\frac{\sigma}{\sqrt{n_1}}\|s-s_{M,T}\|_{n_1}\sqrt{2(x_{M,T}+\xi)}\leq\theta\|s-s_{M,T}\|_{n_1}^2+\frac{2}{\theta}\frac{\sigma^2}{n_1}(x_{M,T}+\xi)$ and $2\sqrt{|T|(x_{M,T}+\xi)}\leq\eta|T|+\eta^{-1}(x_{M,T}+\xi)$ with $\eta=\frac{K+\theta-2}{2-\theta}\frac{1}{2\sqrt{2}(1+\rho\delta)}>0$, we derive that on the set $E_\xi\bigcap\Omega$, for any (M,T),

$$\begin{aligned}
\Delta_{M,T} &\leq (2-\theta)\frac{\sigma^2}{n_1}|T| + 4\sqrt{2}(1+\rho\delta)(2-\theta)\frac{\sigma^2}{n_1}\sqrt{|T|(x_{M,T}+\xi)} \\
&\quad + \left(2(1+\rho\delta)(2-\theta) + \frac{2}{\theta} + 4\frac{\rho}{\sigma^2}R\right)\frac{\sigma^2}{n_1}(x_{M,T}+\xi) \\
&\quad - pen(M,T) \\
&\leq K\frac{\sigma^2}{n_1}|T| + \left(2(1+\rho\delta)(2-\theta)\left(1+\frac{4(1+\rho\delta)(2-\theta)}{K+\theta-2}\right) + \frac{2}{\theta} + 4\frac{\rho}{\sigma^2}R\right)\frac{\sigma^2}{n_1}(x_{M,T}+\xi) \\
&\quad - pen(M,T)
\end{aligned}$$

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

$$pen(M,T) \geq K\frac{\sigma^2}{n_1}|T| + \left[2(1+\rho\delta)(2-\theta)\left(1+\frac{4(1+\rho\delta)(2-\theta)}{K+\theta-2}\right) + \frac{2}{\theta} + 4\frac{\rho}{\sigma^2}R\right]\frac{\sigma^2}{n_1}x_{M,T} \quad \forall (M,T)$$

we get that, on the set $E_\xi \bigcap \Omega$,

$$\Delta_{\widehat{M,T}} \leq \left(2(1+\rho\delta)(2-\theta)\left(1+\frac{4(1+\rho\delta)(2-\theta)}{K+\theta-2}\right) + \frac{2}{\theta} + 4\frac{\rho}{\sigma^2}R\right)\frac{\sigma^2}{n_1}\xi$$

We are going now to control $\underset{(M,T)}{inf}\, R_{M,T}$.

In the same way we deduced (2.13) from assumption (A), we get that for any $M \in \mathcal{P}(\Lambda)$, $T \in \mathcal{M}_{n_1,M}$ and any $x > 0$

$$\begin{aligned}
\mathbb{P}\Big(< \varepsilon, s - s_{M,T} >_{n_1} &\geq \frac{\sigma}{\sqrt{n_1}}\|s - s_{M,T}\|_{n_1}\sqrt{2x} + \frac{2\rho R}{n_1}x \;\Big|\{X_i;\ (X_i,Y_i)\in\mathcal{L}_1\}\Big) \\
&\leq e^{-x}
\end{aligned} \tag{2.14}$$

Thus we derive a set F_ξ such that

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

- on the set F_ξ, for any (M,T),

$$< \varepsilon, s - s_{M,T} >_{n_1} \leq \frac{\sigma}{\sqrt{n_1}}\|s - s_{M,T}\|_{n_1}\sqrt{2\,(x_{M,T}+\xi)} + \frac{2\rho R}{n_1}(x_{M,T}+\xi)$$

It follows from definition of $R_{M,T}$ that on the set F_ξ, for any (M,T),

$$\begin{aligned}
R_{M,T} &\leq \|s - s_{M,T}\|_{n_1}^2 + 2\frac{\sigma}{\sqrt{n_1}}\|s - s_{M,T}\|_{n_1}\sqrt{2\,(x_{M,T}+\xi)} + \frac{4\rho R}{n_1}(x_{M,T}+\xi) + pen(M,T) \\
&\leq 2\|s - s_{M,T}\|_{n_1}^2 + \left(2 + 4\frac{\rho}{\sigma^2}R\right)\frac{\sigma^2}{n_1}(x_{M,T}+\xi) + pen(M,T) \\
&\leq 2\|s - s_{M,T}\|_{n_1}^2 + 2pen(M,T) + \left(2 + 4\frac{\rho}{\sigma^2}R\right)\frac{\sigma^2}{n_1}\xi
\end{aligned}$$

We conclude that on $E_\xi \bigcap F_\xi \bigcap \Omega$

$$(1-\theta)\|s-\tilde{s}\|_{n_1}^2 \leq 2 \underset{(M,T)}{inf} \left\{\|s-s_{M,T}\|_{n_1}^2 + pen(M,T)\right\} + \Upsilon \frac{\sigma^2}{n_1}\xi$$

And, for $p \leq log(n_1)$,

$$\mathbb{P}\left(E_\xi^c \bigcup F_\xi^c \bigcup \Omega^c\right) \leq 3e^{-\xi}\Sigma(a,b) + \frac{1}{n_1} \underbrace{2exp\left(\frac{-\frac{5\rho^2}{\sigma^2}(logn_1)^2 + \frac{50\rho^2}{\sigma^2}(logn_1)(loglogn_1) + 4logn_1}{2\left(1+\frac{5\rho^2}{\sigma^2}logn_1\right)}\right)}_{\epsilon(n_1)}$$

Finally, we have the following result:
Denoting by

$$\begin{aligned} \Upsilon &= 2(1+\rho\delta)(2-\theta)\left(1+\frac{4(1+\rho\delta)(2-\theta)}{K+\theta-2}\right)+\frac{2}{\theta} \\ &= \left[2(1+5\frac{\rho^2}{\sigma^2}log\left(\frac{n_1}{p}\right))(2-\theta)\left(1+\frac{4(1+5\frac{\rho^2}{\sigma^2}log\left(\frac{n_1}{p}\right))(2-\theta)}{K+\theta-2}\right)+\frac{2}{\theta}\right] \end{aligned}$$

and

$$\epsilon(n_1) = 2exp\left(\frac{-\frac{5\rho^2}{\sigma^2}(logn_1)^2 + \frac{50\rho^2}{\sigma^2}(logn_1)(loglogn_1) + 4logn_1}{2\left(1+\frac{5\rho^2}{\sigma^2}logn_1\right)}\right) \underset{n_1\to+\infty}{\longrightarrow} 0$$

Taking a penalty which satisfy: $\forall(M,T)$

$$\begin{aligned} pen(M,T) \geq\ & K\frac{\sigma^2}{n_1}|T| \\ &+ \left(\Upsilon + 4\frac{\rho}{\sigma^2}R\right)\frac{\sigma^2}{n_1}\left(a+(|M|+1)\left(1+log\left(\frac{n_1}{|M|+1}\right)\right)\right)|T| \\ &+ \left(\Upsilon + 4\frac{\rho}{\sigma^2}R\right)\frac{\sigma^2}{n_1}b\left(1+log\left(\frac{p}{|M|}\right)\right)|M| \end{aligned}$$

we have $\forall\xi > 0$, with probability $\geq 1 - 3e^{-\xi}\Sigma(a,b) - \frac{1}{n_1}\epsilon_{n_1}$

$$(1-\theta)\|s-\tilde{s}\|_{n_1}^2 \leq 2 \underset{(M,T)}{inf} \left\{\|s-s_{M,T}\|_{n_1}^2 + pen(M,T)\right\} + \left(\Upsilon + 2 + 8\frac{\rho}{\sigma^2}R\right)\frac{\sigma^2}{n_1}\xi$$

The proposition 2.3.2 results of

$$\begin{aligned} \Upsilon &\leq 8(2-\theta)max\left\{1;\frac{4(2-\theta)}{K+\theta-2}\right\}\left(1+25\frac{\rho^4}{\sigma^4}log^2\left(\frac{n_1}{p}\right)\right)+\frac{2}{\theta} \\ &\leq C(K,\theta)\left(1+\frac{\rho^4}{\sigma^4}log^2\left(\frac{n_1}{p}\right)\right) \end{aligned}$$

and by taking $K = 2$, $\theta \to 1$, $a \to 0$ and $b \to 1$. $\square$

PROOF OF THE PROPOSITION 2.3.3:

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

$$\left\|s - \tilde{\tilde{s}}\right\|_{n_3}^2 \leq \|s - \tilde{s}(\alpha,\beta)\|_{n_3}^2 + 2\left\langle \varepsilon, \tilde{\tilde{s}} - \tilde{s}(\alpha,\beta)\right\rangle_{n_3} \tag{2.15}$$

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

$$\begin{aligned}&log\ \mathbb{E}\left[exp\left(\lambda\left\langle \varepsilon, \tilde{s}(\alpha',\beta') - \tilde{s}(\alpha,\beta)\right\rangle_{n_3}\right)\ \middle|\mathcal{L}_1,\ \mathcal{L}_2 \text{ and } \{X_i;\ (X_i,Y_i) \in \mathcal{L}_3\}\right] \\ &\leq \frac{\sigma^2\|\tilde{s}(\alpha',\beta') - \tilde{s}(\alpha,\beta)\|_{n_3}^2\lambda^2}{2n_3\left(1 - \frac{\rho}{n_3}M_{\alpha,\beta,\alpha',\beta'}|\lambda|\right)} \quad \text{if } |\lambda| < \frac{n_3}{\rho M_{\alpha,\beta,\alpha',\beta'}}\end{aligned}$$

Thus we get that for any $\tilde{s}(\alpha,\beta)$, $\tilde{s}(\alpha',\beta') \in \mathcal{G}$ and any $x > 0$

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

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

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

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

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

Thanks to assumption (A), we get that for any $\lambda \in (-1/\rho, 1/\rho)$

$$\begin{aligned}&log\ \mathbb{E}\left[exp\left(\lambda \sum_{\substack{(X_i,Y_i)\in\mathcal{L}_2 \\ X_i \in t}} \varepsilon_i\right)\ \middle|\mathcal{L}_1 \text{ and } \{X_i;\ (X_i,Y_i) \in \mathcal{L}_2\}\right] \\ &\leq \frac{\lambda^2\sigma^2}{2(1 - \rho|\lambda|)}\,|\{i;\ (X_i,Y_i) \in \mathcal{L}_2 \text{ and } X_i \in t\}|\end{aligned}$$

It follows that for any $x>0$

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

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

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

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

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

It follows from (2.15) that, on the set $\Omega_1\bigcap E_\xi$, for any $\tilde{s}(\alpha,\beta)\in\mathcal{G}$ and any $\eta\in(0;1)$

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

and

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

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 $\geq 1-e^{-\xi}-\frac{R^2}{2(\sigma^2+\rho R)}\frac{1}{n_2^{1-log2}}$,
$\forall\eta\in(0,1)$,

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

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

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

Thanks to assumption 2.3, we get that

$$\mathbb{P}\left(\Omega_2^c\middle|\ \{X_i;\ (X_i,Y_i)\in\mathcal{L}_1\}\right)\quad\leq\quad 2n_1exp\left(-\frac{9\rho^2log^2n_1}{2(\sigma^2+3\rho^2logn_1)}\right)$$

with $\epsilon(n_1)=2n_1exp\left(-\frac{9\rho^2log^2n_1}{2(\sigma^2+3\rho^2logn_1)}\right)\underset{n_1\to+\infty}{\longrightarrow}0$

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

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

It follows from (2.15) that, on the set $\Omega_2 \bigcap E_\xi$, for any $\tilde{s}(\alpha,\beta) \in \mathcal{G}$ and any $\eta \in (0;1)$

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

and

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

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

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

$\square$

2.7.2 Classification

Proof of the proposition 2.4.1:
Let M $\in \mathcal{P}(\Lambda)$, $T \preceq T_{max}^{(M)}$ and $s_{M,T} \in S_{M,T}$. We let

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

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

Following the proof of theorem 4.2 in [15], we get

$$l(s,\tilde{s}) \leq l(s,s_{M,T}) + w_{\widehat{M,T}}(\tilde{s}) \times V_{\widehat{M,T}} + pen(M,T) - pen(\widehat{M,T}) \tag{2.16}$$

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 2.6.1, 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 2.6.6, 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 $\Omega_{\xi,(M,T)}$, $\forall(M',T')$,

$$V_{M',T'} \leq K_1\mathbb{E}\left[V_{M',T'}\right] + K_2\left(\sqrt{\frac{x_{M',T'}+\xi}{2n_2}}{y_{M',T'}}^{-1} + \frac{x_{M',T'}+\xi}{n_2}{y_{M',T'}}^{-2}\right) \tag{2.17}$$

Now we overestimate $\mathbb{E}\left[V_{M',T'}\right]$.
Let $u_{M',T'} \in S_{M',T'}$ such that $d(s,u_{M',T'}) \leq \underset{u\in S_{M',T'}}{inf} d(s,u)$.
Then

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

We prove:

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

For the second term, we have

$$\mathbb{E}\left[\underset{u\in S_{M',T'}}{sup}\left(\frac{|\bar{\gamma_{n_2}}(u) - \bar{\gamma_{n_2}}(u_{M',T'})|}{w_{M',T'}(u)}\right)\right] \leq 4\mathbb{E}\left[\underset{u\in S_{M',T'}}{sup}\left(\frac{|\bar{\gamma_{n_2}}(u) - \bar{\gamma_{n_2}}(u_{M',T'})|}{d^2(u,u_{M',T'}) + (2y_{M',T'})^2}\right)\right]$$

By application of lemma 2.6.9 for $2y_{M',T'} \geq \frac{K_3\sqrt{|T'|}}{\sqrt{n_2}h}$, we deduce

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

Thus from (2.17), 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 (2.16), 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 $\Omega_{\xi,(M,T)}$

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

Integrating with respect to ξ and by minimizing , we get

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

In brief, with a penalty function such that
$\forall M \in \mathcal{P}(\Lambda), \quad \forall T \preceq T_{max}^{(M)}$

$$\begin{aligned} pen(M,T) \quad &= \quad \alpha\frac{|T|}{n_2h} + \beta\frac{|M|}{n_2h}\left(1+log\left(\frac{p}{|M|}\right)\right) \\ &\geq \quad \frac{36(C_1+1)}{C_1-1}\left(64K_1^2K_3^2 + 2aK_2\left(\sqrt{\frac{K_2}{2}}+\sqrt{\frac{h(C_1-1)}{6(C_1+1)}}\right)^2\right)\frac{|T|}{n_2h} \\ & \quad + \frac{36(C_1+1)}{C_1-1}2K_2\left(\sqrt{\frac{K_2}{2}}+\sqrt{\frac{h(C_1-1)}{6(C_1+1)}}\right)^2 b\frac{|M|}{n_2h}\left(1+log\left(\frac{p}{|M|}\right)\right)\end{aligned}$$

we have:

$$\mathbb{E}\left[l(s,\tilde{s})|\mathcal{L}_1\right] \le C_1 \underset{M,T}{inf}\left\{l(s,S_{M,T}) + pen(M,T)\right\} + \frac{C}{n_2 h}\Sigma(a,b)$$

We notice that, the two constants α_0 and β_0, which appear in the proposition 2.4.1, are defined by

$$\begin{aligned}
\alpha_0 &= 36\left(64K_1^2K_3^2 + 4log(2)K_2\left(\sqrt{\frac{K_2}{2}} + \frac{1}{\sqrt{6}}\right)^2\right) \\
\beta_0 &= 72K_2\left(\sqrt{\frac{K_2}{2}} + \frac{1}{\sqrt{6}}\right)^2
\end{aligned}$$

□

Proof of the proposition 2.4.2:
For $M, M' \in \mathcal{P}(\Lambda)$, $T \preceq T_{max}^{(M)}$,$T' \preceq T_{max}^{(M')}$ and $s_{M,T} \in S_{M,T}$. We let

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

where $y_{M',T'}$ and$y_{M,T}$ are parameters that will be chosen later.

Following the proof of theorem 4.2 in [15], we get

$$l(s,\tilde{s}) \le l(s,s_{M,T}) + w_{(\widehat{M,T}),(M,T)}(\tilde{s}) \times V_{(\widehat{M,T}),(M,T)} + pen(M,T) - pen(\widehat{M,T}) \qquad (2.18)$$

To control $V_{(\widehat{M,T}),(M,T)}$, we check a uniform overestimation of $V_{(M',T'),(M,T)}$. To do this, we apply the Talagrand's concentration inequality, written in lemma 2.6.1, to $V_{(M',T'),(M,T)}$, for $(M',T') \in \mathcal{P}(\Lambda) \times \mathcal{M}_{n_1,M'}$ and $(M,T) \in \mathcal{P}(\Lambda) \times \mathcal{M}_{n_1,M}$. So we obtain that for any $(M',M) \in \mathcal{P}(\Lambda)^2$, any $T' \in \mathcal{M}_{n_1,M'}$, any $(M,T) \in \mathcal{M}_{n_1,M}$ and any $x > 0$,

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

where K_1 and K_2 are universal positive constants.

Setting $x = x_{M',T'} + x_{M,T} + \xi$, with $\xi > 0$ and the weights

$$x_{M',T'} = \left(a + (|M'|+1)\left(1 + log\left(\frac{n_1}{|M'|+1}\right)\right)\right)|T'| + b|M'|\left(1 + log\left(\frac{p}{|M'|}\right)\right),$$

as defined in lemma 2.6.7, and summing all those inequalities with respect to (M',T') and (M,T), we derive a set Ω_ξ such that

- $\mathbb{P}\Big(\Omega_\xi^c|\{X_i,\ (X_i,Y_i)\in\mathcal{L}_1\}\Big)\leq e^{-\xi}(\Sigma(a,b))^2$
- on Ω_ξ, $\forall(M',T')$, (M,T),

$$\begin{aligned}V_{(M',T'),(M,T)} \;&\leq\; K_1\mathbb{E}\left[V_{(M',T'),(M,T)}\right]+K_2\left(\sqrt{\frac{x_{M',T'}+x_{M,T}+\xi}{2n_1}}(y_{M',T'}+y_{M,T})^{-1}\right.\\&\left.+\;\frac{x_{M',T'}+x_{M,T}+\xi}{n_1}(y_{M',T'}+y_{M,T})^{-2}\right)\end{aligned}\tag{2.19}$$

Now we overestimate $\mathbb{E}\left[V_{(M',T'),(M,T)}\right]$.
Let $u_{M',T'}\in S_{M',T'}$ such that $d(s,u_{M',T\prime})\leq \underset{u\in S_{M',T'}}{inf}\ d(s,u)$.
Then

$$\mathbb{E}\left[V_{M',T'}\right]\leq\mathbb{E}\left[\frac{|\bar{\gamma_{n_1}}(u_{M',T'})-\bar{\gamma_{n_1}}(s_{M,T})|}{\underset{u\in S_{M',T'}}{inf}\ (w_{M',T'}(u))}\right]+\mathbb{E}\left[\underset{u\in S_{M',T'}}{sup}\left(\frac{|\bar{\gamma_{n_1}}(u)-\bar{\gamma_{n_1}}(u_{M',T'})|}{w_{M',T'}(u)}\right)\right]$$

We prove:

$$\mathbb{E}\left[\frac{|\bar{\gamma_{n_1}}(u_{M',T'})-\bar{\gamma_{n_1}}(s_{M,T})|}{\underset{u\in S_{M',T'}}{inf}\ (w_{M',T'}(u))}\right]\leq\frac{1}{\sqrt{n_1}(y_{M',T'}+y_{M,T})}$$

For the second term, we have

$$\mathbb{E}\left[\underset{u\in S_{M',T'}}{sup}\left(\frac{|\bar{\gamma_{n_1}}(u)-\bar{\gamma_{n_1}}(u_{M',T'})|}{w_{M',T'}(u)}\right)\right]\leq 4\mathbb{E}\left[\underset{u\in S_{M',T'}}{sup}\left(\frac{|\bar{\gamma_{n_1}}(u)-\bar{\gamma_{n_1}}(u_{M',T'})|}{d^2(u,u_{M',T'})+(2(y_{M',T'}+y_{M,T}))^2}\right)\right]$$

By application of lemma 2.6.9 for $2y_{M',T'}\geq\frac{K_3\sqrt{|T'|}}{\sqrt{n_2}h}$, we deduce

$$\mathbb{E}\left[\underset{u\in S_{M',T'}}{sup}\left(\frac{|\bar{\gamma_{n_1}}(u)-\bar{\gamma_{n_1}}(u_{M',T'})|}{w_{M',T'}(u)}\right)\right]\leq\frac{8K_3\sqrt{|T'|}}{\sqrt{n_1}(y_{M',T'}+y_{M,T})}$$

Thus from (2.19), we know that on Ω_ξ and for any $(M',T'),(M,T)$

$$\begin{aligned}V_{(M',T'),(M,T)} \;&\leq\; \frac{K_1}{\sqrt{n_1}(y_{M',T'}+y_{M,T})}\left(8K_3\sqrt{|T'|}+1\right)\\&+K_2\left(\sqrt{\frac{x_{M',T'}+x_{M,T}+\xi}{2n_1}}(y_{M',T'}+y_{M,T})^{-1}\right.\\&\left.+\frac{x_{M',T'}+x_{M,T}+\xi}{n_1}(y_{M',T'}+y_{M,T})^{-2}\right)\end{aligned}$$

For $y_{M',T'} = 3K\left(\frac{K_1}{\sqrt{n_1}}\left(8K_3\sqrt{|T'|}+1\right) + K_2\sqrt{\frac{x_{M',T'}+\xi/2}{2n_1}} + \frac{1}{\sqrt{3K}}\sqrt{K_2\frac{x_{M',T'}+\xi/2}{n_1}}\right)$
with $K \geq \frac{1}{48K_1h}$, we get

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

By overestimating $w_{\widehat{M,T}}(\tilde{s})$, $y^2_{M,T}$ and replacing all of those results in (2.18), we get

$$\begin{aligned}
\left(1-\frac{2}{Kh}\right) l\left(s,\tilde{s}\right) \;\; \leq \;\; & \left(1+\frac{2}{Kh}\right) l\left(s,s_{M,T}\right) - pen(\widehat{M,T}) + pen(M,T) \\
& +36K\left(\frac{64{K_1}^2{K_3}^2}{n_1}|\hat{T}| + \frac{64{K_1}^2{K_3}^2}{n_1}|T| + 2K_2\frac{x_{\widehat{M,T}}}{n_1}\left(\sqrt{\frac{K_2}{2}}+\frac{1}{\sqrt{3K}}\right)^2\right. \\
& \left. + 2K_2\frac{x_{M,T}}{n_1}\left(\sqrt{\frac{K_2}{2}}+\frac{1}{\sqrt{3K}}\right)^2\right) \\
& +36K\left(\frac{4K_1^2}{n_1} + 2K_2\frac{\xi}{n_1}\left(\sqrt{\frac{K_2}{2}}+\frac{1}{\sqrt{3K}}\right)^2\right)
\end{aligned}$$

We let $K = \frac{2}{h}\frac{C+1}{C-1}$ with $C > 1$
Taking a penalty $pen(\widehat{M,T})$ which balances all the terms in $\left(\widehat{M,T}\right)$, i.e.

$$pen(M,T) \geq \frac{72(C+1)}{h(C-1)}\left(\frac{64K_1^2K_3^2}{n_1}|T| + 2K_2\frac{x_{M,T}}{n_1}\left(\sqrt{\frac{K_2}{2}}+\sqrt{\frac{(C-1)}{6(C+1)}}\right)^2\right)$$

We obtain that on Ω_ξ

$$l(s,\tilde{s}) \leq 2C\left\{l(s,s_{M,T}) + pen(M,T)\right\} + \frac{C_2}{n_1h} + \frac{C_3}{n_1h}\xi$$

In brief, with a penalty function such that
$\forall M \in \mathcal{P}(\Lambda), \quad \forall T \preceq T_{max}^{(M)}$

$$\begin{aligned}
pen(M,T) \;\; = \;\; & \alpha\frac{|T|}{n_1h}\left(1+(|M|+1)\left(1+log\left(\frac{n_1}{|M|+1}\right)\right)\right) + \beta\frac{|M|}{n_1h}\left(1+log\left(\frac{p}{|M|}\right)\right) \\
\geq \;\; & \frac{72(C+1)}{C-1}\left(64K_1^2K_3^2 + 2K_2\left(\sqrt{\frac{K_2}{2}}+\sqrt{\frac{C-1}{6(C+1)}}\right)^2\right)\frac{|T|}{n_1h} \\
& \times\left(a+(|M|+1)\left(1+log\left(\frac{n_1}{|M|+1}\right)\right)\right) \\
& +\frac{72(C+1)}{C-1}2K_2\left(\sqrt{\frac{K_2}{2}}+\sqrt{\frac{C-1)}{6(C+1)}}\right)^2 b\frac{|M|}{n_1h}\left(1+log\left(\frac{p}{|M|}\right)\right)
\end{aligned}$$

we have

$$l(s,\tilde{s}) \leq 2C\left\{ l(s,s_{M,T}) + pen(M,T) \right\} + \frac{C_2}{n_1 h}\left(1+\xi\right)$$

We notice that, the two constants α_0 and β_0 which appear in the proposition 2.4.2, are defined by

$$\begin{aligned} \alpha_0 &= 72\left(64K_1^2K_3^2 + 2K_2\left(\sqrt{\frac{K_2}{2}} + \frac{1}{\sqrt{6}}\right)^2\right) \\ \beta_0 &= 72.2K_2\left(\sqrt{\frac{K_2}{2}} + \frac{1}{\sqrt{6}}\right)^2 \end{aligned}$$

□

Proof of the proposition 2.4.3:
This result is obtained by a direct application of the lemma 2.6.3 which appears in the subsection 2.6

□

Chapitre 3

Objectivation de l'agrément de conduite automobile

Sommaire

Ce chapitre présente un travail effectué en collaboration avec Jean-Michel Poggi et qui est accepté par la Revue de Statistique Appliquée. Ce chapitre est composé de l'article soumis et d'une section complémentaire consacrée à la mise en œuvre de la procédure de sélection de variables décrite dans le chapitre précédent.

Dans le **Chapitre 2**, nous avons abordé le problème de la sélection de variables essentiellement d'un point de vue théorique.
Dans ce chapitre, nous traitons un problème réel de sélection de variables dans lequel les variables explicatives sont de type fonctionnel et la variable à expliquer possède cinq modalités. Dans ce travail, nous devons d'une part identifier les variables fonctionnelles pertinentes et d'autre part identifier, sur chacune d'entre elles, les plages temporelles responsables de cette pertinence.
La méthode développée pour résoudre cette double phase de sélection repose sur les ondelettes et sur une stratégie mêlant une procédure pas à pas et CART.

3.1 Introduction

Ce travail est motivé par un problème réel appelé l'objectivation. Il consiste à expliquer l'agrément de conduite traduisant un confort ressenti relativement à une prestation donnée, par exemple le comportement de la boîte de vitesses lors de la phase de mise en mouvement d'un véhicule, au moyen de critères "physiques", c'est-à-dire de variables issues de signaux (comme une vitesse, des couples ou encore la position de pédales) mesurés lors d'essais. Il s'agit d'utiliser cette quantification pour en tenir compte lors de la phase de conception du véhicule. Il s'inscrit dans la continuité de travaux menés par Renault portant sur la prestation décollage à plat pour un groupe moto-propulseur à boîte de vitesses robotisée (cf. Ansaldi [4]).

Dans cet article, nous développons une approche alternative pour le problème de la sélection des variables discriminantes en tentant de plus tirer profit du caractère fonctionnel des données. De ce point de vue, ce travail peut être rapproché de l'analyse des données fonctionnelles. Citons Deville [28], Dauxois et Pousse [27] pour les travaux pionniers dans les années 70. Plus récemment, on peut citer par exemple, Leurgans *et al.* [63], Hastie *et al.* [53] et ces dernières années, Ferraty et Vieu [39], Ferré *et al.* ([41], [40]), Rossi et Conan-Guez [81], Biau *et al.* [12] ainsi que le texte de synthèse de Besse, Cardot [11]. En outre les deux livres de Ramsay et Silverman ([78], [79]) constituent une ressource précieuse.
Dans ce travail, nous préférons utiliser la méthode CART particulièrement adapté pour la sélection de variables.

Comme cela est classique dans de nombreuses applications où les variables explicatives sont des courbes, le problème industriel qui nous occupe est mal posé, au sens où le nombre de variables explicatives est très supérieur à la taille de l'échantillon. L'un des exemples typiques de telles situations est fourni par les données d'expression du génome. On trouvera dans Dudoit *et al.* [35] la présentation de ce problème et de diverses méthodes de classification supervisée actuellement en compétition. On pourra aussi consulter Vannucci *et al.* [87] pour la situation où les variables explicatives sont des spectres, ce qui est classique en chimiométrie.

Structurellement le problème industriel qui nous intéresse présente une particularité supplémentaire : nous disposons non pas d'une seule variable explicative, qui est une courbe, mais d'un grand nombre de variables fonctionnelles parmi lesquelles il faut choisir les plus influentes. Notre approche s'intéresse donc à un double problème de sélection : celle des variables fonctionnelles et celle des descripteurs les plus discriminants, pour les signaux ainsi sélectionnés.

La démarche adoptée procède en trois étapes et utilise deux outils fondamentaux que sont d'une part la méthode des ondelettes (cf. Misiti *et al.* [73]) et d'autre part la méthode de classification non linéaire CART (cf. [20]). Les trois étapes sont constituées d'un prétraitement des signaux (incluant débruitage par ondelettes, recalage et synchronisation), d'une réduction de la dimension par compression dans une base d'ondelettes commune, puis de l'extraction et sélection des variables utiles au moyen d'une stratégie incluant des applications successives de la méthode CART.

Le plan de l'article est le suivant. Après cette introduction, le paragraphe 2 présente le contexte de l'application : le problème et les données. Dans le paragraphe 3, la démarche adoptée est détaillée. Enfin le paragraphe 4 regroupe quelques éléments de conclusion.

3.2 Le contexte applicatif

3.2.1 Le problème

La campagne d'essais réalisée par Renault (cf. Ansaldi [4]) a conduit à faire varier les facteurs suivants : le réglage de la boîte de vitesses, les conditions de roulage et les pilotes. Lors de ces essais, ont été mesurés d'une part l'agrément du pilote et d'autre part des données objectives consistant dans le relevé, à l'aide de capteurs, de plusieurs signaux temporels.

Précisons quelques éléments de terminologie utiles dans la suite. On appelle "produit" un élément de

$$\{produits\} \;=\; \{conditions\ de\ roulage\} \;\times\; \{3\ réglages\ de\ la\ boite\ de\ vitesses\}$$

où

$$\{conditions\ de\ roulage\} \;= \{2\ charges\} \;\times\; \{2\ angles\ pédale\} \times\; \{2\ vitesses\ pédale\}$$

ce qui conduit au plus à 24 produits (12 pour chacune des charges).

On appelle "essai" un élément de

$$\{essais\} \;=\; \{7\ pilotes\} \;\times\; \{24\ produits\}$$

conduisant à un maximum de 168 essais.

Les essais à 140 kg de charge ont été menés séparément des essais à 280 kg de charge. Pour chaque charge, 6 produits parmi les 12 possibles ont été testés : 4 pilotes ont comparé par paires ces 6 produits. Après analyse des résultats, 114 essais à 140 kg et 118 essais à 280 kg ont été retenus. Chacun de ces essais est représenté par un ensemble de 21 variables fonctionnelles qui correspondent aux signaux mesurés par les capteurs durant l'expérience.

L'étude menée dans [4] s'articule autour de trois phases :

- l'association d'un agrément à chacun des produits.
  Pour chaque paire d'essais, le pilote précisait son essai préféré. A partir de ces données de comparaisons par paires et à l'aide d'une méthode inspirée du "multidimensional scaling" (voir la thèse de Favre [37]) sont obtenus un classement des produits par pilote et un agrément consensuel à toute la population des pilotes, par charge. Cet agrément associe à un produit un rang de satisfaction (le rang 1 étant celui du produit le plus apprécié) ;

- l'extraction de critères puis sélection par analyse discriminante.
  A partir des signaux mesurés, de très nombreux critères sont générés puis, au moyen d'une analyse discriminante linéaire arborescente dite par moindres écarts (c'est-à-dire basée sur un critère L^1), un petit nombre d'entre eux expliquant l'agrément, sont extraits ;

- le calcul d'intervalles de tolérance.
  Pour chacun des critères pertinents, un intervalle qui maximise l'agrément sous certaines contraintes sur les produits, est construit (ce point constitue d'ailleurs la contribution majeure de la thèse d'Ansaldi [4]).

On se concentre, dans cet article, sur la deuxième étape en utilisant une approche plus fonctionnelle. Bien sûr, on ne considère que les données issues de la phase 1 qui sont seules détaillées dans le paragraphe suivant. L'agrément est le rang consensuel attribué à chacun des 6 produits testés. Ceci conduit à un problème de discrimination, au lieu d'un problème de régression avec une variable à expliquer ordinale discrète.

Dans la suite, ne seront considérés que les essais à 140 kg de charge (pour les essais à 280 kg de charge, la démarche est identique et les résultats obtenus dans l'étude [4] sont semblables).

3.2.2 Les données

Les données sont constituées des couples $((X_i^j)_{1\leq j\leq J}, Y_i)_{1\leq i\leq n}$, où $n = 114$ et $J = 21$, et :

- Y_i représente le rang attribué au produit testé lors de l'essai i ;
- X_i^j représente la $j^{ème}$ variable fonctionnelle mesurée lors de l'essai i et est le signal $\{X_i^j(t)\}_{t\in T_i}$ où T_i est la grille temporelle régulière propre à l'essai i.

Autrement dit, pour chacun des essais, on dispose de l'agrément et de 21 signaux (on parlera dans la suite, suivant le contexte, de signaux comme de variables fonctionnelles ou encore de courbes) pour la plupart d'environ 1000 points (en fait ils comportent entre 600 et 5000 points). Ces variables fonctionnelles sont principalement des positions, des vitesses, des accélérations, des couples et des régimes moteur, cependant pour des raisons de confidentialité la nature des variables ne peut pas être indiquée de façon plus précise. Notons que la fréquence d'échantillonnage de 250 Hz est la même pour tous les essais et correspond à une haute résolution temporelle.

La distribution de l'agrément Y, après regroupement en 5 modalités, est donnée par les fréquences : $33\%, 17\%, 17\%, 18\%, 15\%$. Seulement 5 modalités, et non 6, sont prises en considération, deux produits ayant obtenu le même agrément.

On trouve dans la **Figure 3.1**, les quatre variables fonctionnelles X^j correspondant à $j = 4$, $14, 17, 22$ pour les essais 7 et 19.

L'examen des graphiques permet de formuler quelques remarques préliminaires concernant ces variables fonctionnelles :

- elles sont observées sur une grille temporelle propre à l'essai, ce qui nécessitera des recalages temporels ;
- elles peuvent être d'allure générale et d'ordre de grandeur très différents, à la fois pour un même essai mais aussi au travers des différents essais, ce qui impliquera des recalages en ordonnée des courbes ;

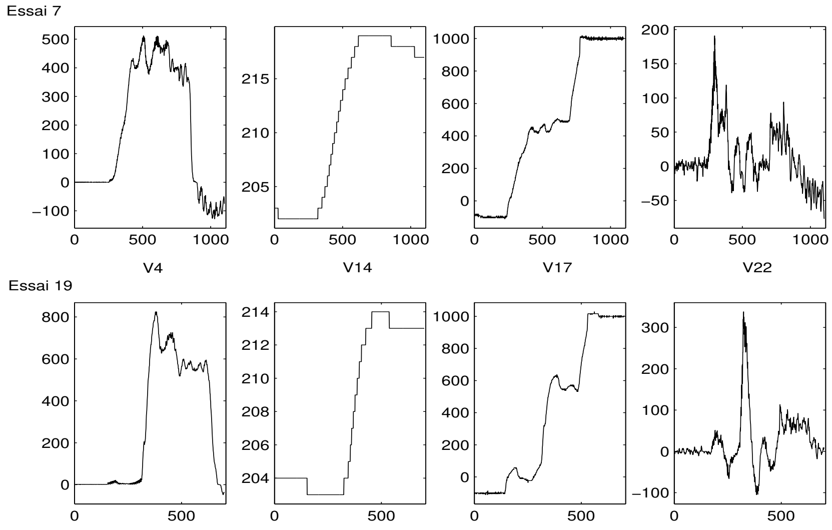

FIG. 3.1 – Pour les essais 7 et 19, les quatre variables fonctionnelles X^j correspondant à $j = 4, 14, 17, 22$, notées simplement V4, V14, V17 et V22. Elles sont observées sur une grille temporelle propre à l'essai et présentent des caractéristiques temporelles variées.

- elles présentent des caractéristiques temporelles très différentes, par exemple le rapport signal sur bruit, élevé en général, peut s'avérer modéré comme dans le cas de la variable 22 ou encore l'être localement, comme c'est le cas pour ces quatre variables sauf la variable 14 qui est une fonction constante par morceaux. Il est clair qu'un débruitage, sans être en général crucial, peut s'avérer utile ;
- la forme générale est souvent simple et peu de paramètres ou peu d'événements semblent suffisants pour la caractériser. Ceci permet d'espérer à la fois une caractérisation économe des variables fonctionnelles ainsi qu'une compression efficace.

Remarque 3.2.1
La variabilité, entre les essais, des durées d'observation et celle des amplitudes des signaux mesurés, résultent des différences de conditions de roulage et de l'exécution plus ou moins scrupuleuse des consignes par les pilotes. □

3.3 La démarche

Le cadre général dans lequel on se place est celui de la sélection de variables dans un problème de discrimination, et consiste à construire une fonction, génériquement notée F dans la suite, pour prédire Y à l'aide de :

$$\hat{Y} = F(X^1, ..., X^J).$$

Dans cette perspective, il sera utile de sélectionner parcimonieusement les variables fonctionnelles qui peuvent expliquer l'agrément, puis pour chacune d'elles, de ne retenir qu'un très faible nombre d'aspects la décrivant, pour des raisons évidentes de robustesse.

Autrement dit, on cherche à sélectionner ce que nous appelons dans ce contexte, des critères notés C^{j_k}, déduits des X^j, de façon à prédire convenablement Y par :

$$\hat{Y} = F(C^{j_1}, ..., C^{j_K})$$

avec $K << J$, typiquement de l'ordre de 5 pour l'application industrielle.

Rappelons que dans le cadre de l'objectivation, il ne s'agit pas d'expliquer au mieux l'agrément en utilisant toutes les informations disponibles, comme par exemple les conditions de roulage, qui ont un impact certain, mais de l'expliquer partiellement en se restreignant exclusivement à des variables déduites des signaux mesurés de façon à pouvoir remonter à des paramètres de conception du véhicule.

La démarche adoptée procède en trois étapes :

- un prétraitement des signaux, incluant débruitage par ondelettes, recalage et synchronisation ;
- une réduction de la taille des signaux par compression dans une base d'ondelettes commune ;
- l'extraction des variables utiles au moyen d'une stratégie pas à pas procédant par des applications successives de la méthode CART.

Une schématisation de cette procédure est illustrée par la **Figure 3.2**.

Détaillons successivement chacune de ces trois phases.

3.3.1 Prétraitement des signaux

Les données $X_i^j = \{X_i^j(t)\}_{t \in T_i}$ sont prétraitées de façon d'une part, à les débruiter individuellement c'est-à-dire pour un essai et une variable fonctionnelle donnés et, d'autre part, à les rendre plus homogènes au moyen de recalages.

Tronquer les signaux

Avant ces deux traitements, on isole une phase qui est la seule à être directement déduite de connaissances externes propres au problème. En effet, en dépit de consignes clairement définies, les durées d'enregistrement et les dates des différentes étapes de l'essai ne sont pas synchrones. Néanmoins, on peut définir deux événements à réaligner : le "vrai" début de l'essai et sa "vraie" fin qui sont lisibles au travers des variables fonctionnelles 8 et 21. Ces deux événements correspondent physiquement au démarrage réel du véhicule et à la définition de la fin de l'essai.

On trouve dans la **Figure 3.3**, trois variables fonctionnelles X^j correspondant à $j = 8, 21, 7$ pour les essais 7 et 19. Les deux premières servent de marqueur du "vrai" début de l'essai et de sa "vraie" fin, respectivement. La période utile de l'essai est visualisée sur les graphes de

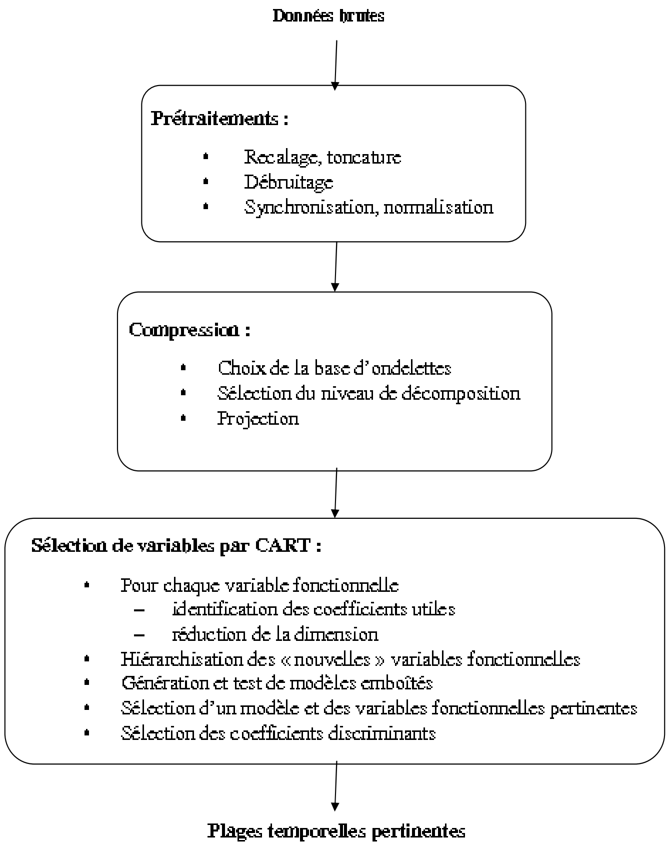

FIG. 3.2 – Représentation schématique de la démarche.

la variable fonctionnelle 7 par la portion de signal située entre les deux instants matérialisés par des lignes verticales. Bien sûr, ces instants varient en fonction de l'essai.
Pour l'essai i, on note $\widetilde{T}_i$ la grille T_i convenablement tronquée aux extrémités.

Débruiter les signaux

A i et j fixés, le signal mesuré est contaminé par un bruit de capteur. Bien sûr, il convient de l'éliminer avant tout traitement de type recalage ou interpolation des données, qui conduirait à les modifier et donc altérer la nature stochastique du bruit qui affecte le signal utile. Comme l'atteste la **Figure 3.1**, la régularité locale de celui-ci peut beaucoup varier au cours du temps, il convient donc d'utiliser des techniques de débruitage adaptatives en espace. C'est le cas de celles basées sur les méthodes d'ondelettes (cf. Donoho, Johnstone [34] pour l'un des articles fondateurs, Vidakovic [94] pour un large tour d'horizon de ces méthodes et Misiti *et al.* [73] pour une introduction aisée).

On considère le modèle suivant, usuel en traitement statistique du signal et réaliste dans cette application :

$$\forall t \in \widetilde{T}_i, \quad X_i^j(t) \ = \ f_i^j(t) \ + \ \eta_i^j(t)$$

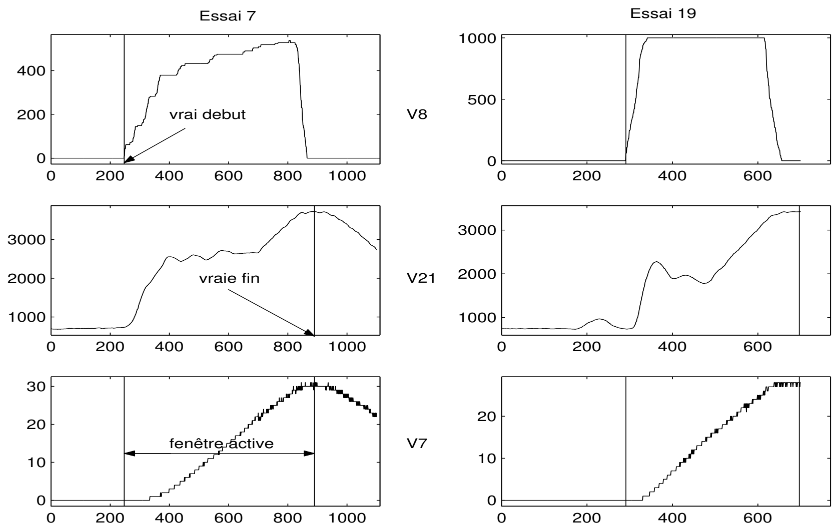

FIG. 3.3 – Pour les essais 7 et 19, les trois variables fonctionnelles X^j correspondant à $j = 8, 21, 7$ notées simplement V8, V21 et V7 sur le graphique. Les deux premières servent de marqueur au "vrai" début de l'essai et sa "vraie" fin, respectivement. La période utile de l'essai est visualisée sur les graphes de la variable fonctionnelle 7 par la portion de signal située entre les deux instants matérialisés par des lignes verticales.

où $\{\eta_i^j(t)\}_{t\in\widetilde{T}_i}$ est un bruit blanc. Dans ce cadre, le débruitage consiste à décomposer le signal dans une base d'ondelettes, à seuiller les coefficients de détail de façon à éliminer essentiellement ceux attribuables au bruit puis à reconstruire un signal débruité constitué de la somme d'une approximation lisse et de détails à diverses échelles correspondant aux fluctuations rapides du signal utile.

On obtient ainsi une estimation $\{\hat{f}_i^j(t)\}_{t\in\widetilde{T}_i}$, ou encore un signal débruité $\{\hat{X}_i^j(t)\}_{t\in\widetilde{T}_i}$.

La **Figure 3.4** présente les résultats obtenus après débruitage par ondelettes des quatre variables fonctionnelles montrées en **Figure 3.1**. La méthode utilise l'ondelette de Daubechies presque symétrique d'ordre 4 "sym4", un niveau de décomposition entre 3 et 5 (suivant les signaux) et le seuillage dit "universel" (cf. Donoho et Johnstone [34]).

Comme on peut le remarquer, le débruitage par ondelettes permet de supprimer de façon satisfaisante le bruit tout en préservant les composantes à haute fréquence du signal utile.

Synchroniser et normaliser les signaux

L'objectif de cette étape est d'éliminer la dépendance en i de la grille temporelle. On procède pour chaque signal, tout d'abord à un recalage linéaire en temps en ramenant la grille $\widetilde{T}_i$ sur l'intervalle $[0, 1]$. Puis, on effectue une interpolation linéaire du signal, suivie d'un échantillonnage pour se ramener à la grille régulière à m points de $[0, 1]$ (ici on fixe $m = 512$, valeur

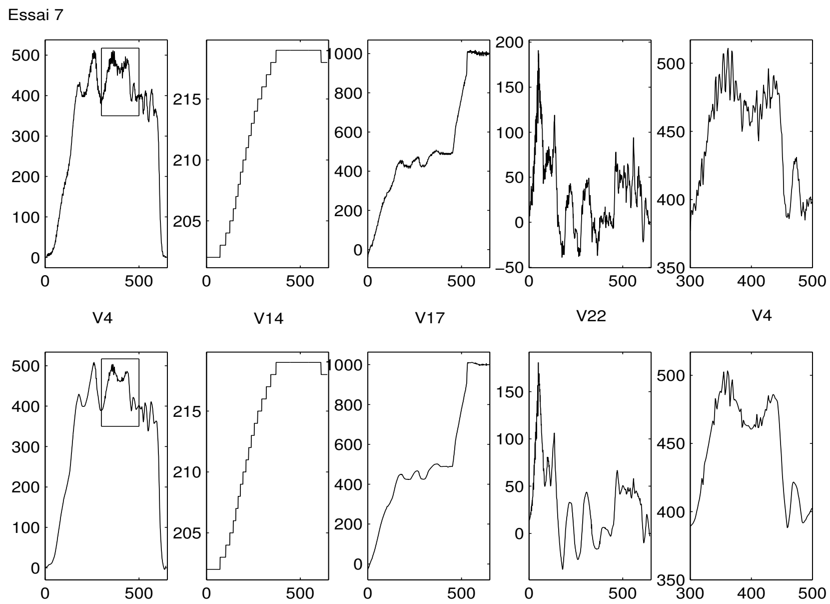

FIG. 3.4 – Pour l'essai 7, en haut de la figure les quatre signaux $X_7^j(t)$ $(j = 4, 14, 17, 22)$ et, en bas, leurs versions débruitées. Dans les deux derniers graphiques à droite, un zoom sur une portion du premier signal permet d'apprécier la qualité du débruitage par ondelettes, à la fois efficace pour débruiter les parties lisses tout en préservant les composantes à haute fréquence du signal utile.

largement suffisante pour des durées de fenêtres actives comprises entre 300 et 700 observations). Un instant dans cette nouvelle "unité" de temps s'interprète comme la proportion de la durée de l'essai écoulée.

On dispose donc de $\{\widetilde{X}_i^j(t)\}_{t\in T}$, sur la grille fixe $T = \{\frac{1}{m}, ..., \frac{m-1}{m}, 1\}$.

Enfin, pour éliminer certains effets d'échelle, en partie liés aux conditions de roulage, les signaux sont normalisés en ordonnée.

Remarque 3.3.1
Un autre prétraitement consiste à effectuer un recalage non linéaire en alignant pour tout j, les n signaux à l'aide de marqueurs convenablement choisis (cf. Bigot [13]). Ceci amènerait à considérer le problème plus sous un aspect de classification de formes. Cependant, cela serait extrêmement lourd et engendrerait une difficulté quant à la remontée dans le temps d'origine en particularisant de nouveau les variables fonctionnelles, et limiterait l'interprétation.

En revanche, cela permettrait de poursuivre un objectif plus ambitieux consistant à rechercher à renforcer l'homogénéité à Y fixé, en mettant au point le recalage pour chaque modalité de la réponse.

Mentionnons que des méthodes de recalage temporel intermédiaires entre la solution adoptée et celle-ci sont envisageables, comme par exemple le type de méthode de recalage décrit dans [78] qui cherche à rapprocher des fonctions de leur moyenne. □

Remarque 3.3.2
De manière implicite, dans la suite du travail (mais aussi dans les travaux antérieurs menés dans ce contexte par Renault), les essais sont considérés comme des réplications indépendantes. Des classifications non supervisées et des ACP fonctionnelles (cf. Ramsay, Silverman [78]) permettent de corroborer raisonnablement l'idée que les effets dus au pilote et aux conditions de roulage sont négligeables devant les autres facteurs de variabilité. □

3.3.2 Compression des signaux

A l'issue de la phase de prétraitement, on dispose donc pour chaque essai, de $J = 21$ signaux débruités, de $m = 512$ points. Chacun de ces signaux peut donc être représenté dans une base d'ondelettes ou de paquets d'ondelettes par très peu de coefficients (cf. Mallat [66] et Coifman, Wickerhauser [24]). Il suffit, par exemple, pour un signal donné, de sélectionner les coefficients les plus grands en valeur absolue, exploitant ainsi la capacité des ondelettes à concentrer l'énergie d'un signal (pour des classes très larges de signaux), en un très petit nombre de ses grands coefficients d'ondelettes.

Le problème est ici de choisir, variable fonctionnelle par variable fonctionnelle, une base commune à tous les essais pour les représenter de façon compacte. Pour déterminer une base commune de décomposition, on peut se restreindre à un petit nombre de bases différentes comme les espaces d'approximation en ondelettes de résolution de plus en plus grossière. Comme $512 = 2^9$, seule une demie douzaine de bases, l'ondelette étant choisie (ici on utilise l'ondelette de Daubechies presque symétrique d'ordre 4 "sym4"), sont à mettre en compétition. Le choix peut être :

- effectué indépendamment de la variable Y et guidé par la définition d'un critère de qualité comme par exemple la moyenne de l'erreur d'approximation du signal par sa projection convenablement pénalisé.

  Afin de déterminer le niveau de décomposition de chacun des signaux j, on considère le critère $EQ_j(p)$ lié à l'énergie et défini comme suit :

  - pour une variable fonctionnelle j et pour un individu i, soit $X_i^j(t)$ le signal d'origine et $A_{i,p}^j(t)$ le signal reconstruit à partir des coefficients d'approximation du niveau p ;
  - on définit l'erreur de la variable fonctionnelle j par

$$EQ_j(p) = \sum_{i=1}^{114} \|X_i^j(t) - A_{i,p}^j(t)\|^2.$$

Remarque 3.3.3
Notons que, lorsque le niveau de décomposition p augmente, le nombre de coefficients et la qualité d'approximation diminuent. Le choix du niveau de décomposition résulte d'un compromis entre le nombre de coefficients retenus et la qualité d'approximation. □

Le choix du niveau de décomposition de la variable j consiste alors à déterminer la plus petite valeur de p pour laquelle on détecte un changement de pente "suffisant" dans le graphe de $(p, EQ_j(p))_{1\leq p\leq 9}$ et à ôter 1 à titre conservatoire.

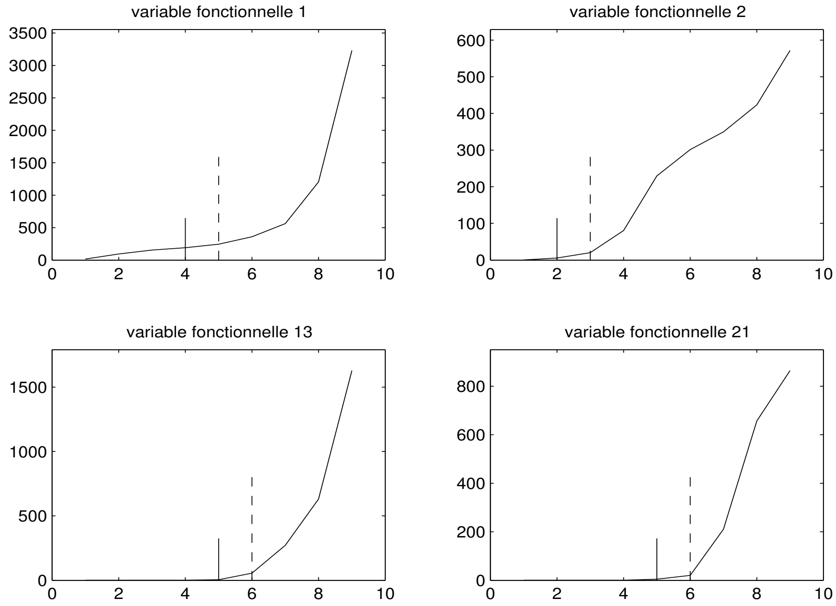

FIG. 3.5 – Pour les variables fonctionnelles 1, 2, 13 et 21, on représente $(p, EQ(p))_{1\leq p\leq 9}$ en trait plein, la plus petite valeur de m pour laquelle on détecte un changement de pente "suffisant" en traits pointillés et cette valeur ôtée de 1 en traits pleins.

La **Figure 3.5** esquisse la façon dont le niveau de décomposition lors de la compression par ondelettes est déterminé pour chacun des signaux.

- basé sur un critère dépendant de la variable Y, comme par exemple l'erreur de classification d'un arbre CART (voir paragraphe suivant).

L'emploi d'une procédure inspirée du premier choix ci-dessus avec recherche d'une cassure dans la répartition moyenne de l'énergie, conduit à retenir majoritairement 16 coefficients et donc à réduire $\mathbb{R}^{J\,\times\,m}$ à $\mathbb{R}^{\sum m_j}$ avec $\sum m_j \approx 300$ ou 400 suivant la stratégie adoptée pour comprimer une variable fonctionnelle (d'ailleurs non discriminante) dont les fluctuations à haute fréquence sont significatives.

La **Figure 3.6** présente pour deux variables fonctionnelles, les résultats obtenus après compression par ondelettes : le signal après compression superposé au signal prétraité est représenté dans le premier graphique, le second (en dessous) contient les coefficients d'approximation associés. Ceux-ci peuvent, bien sûr, être de taille différente puisque le niveau de décomposition retenu dépend de la variable considérée.

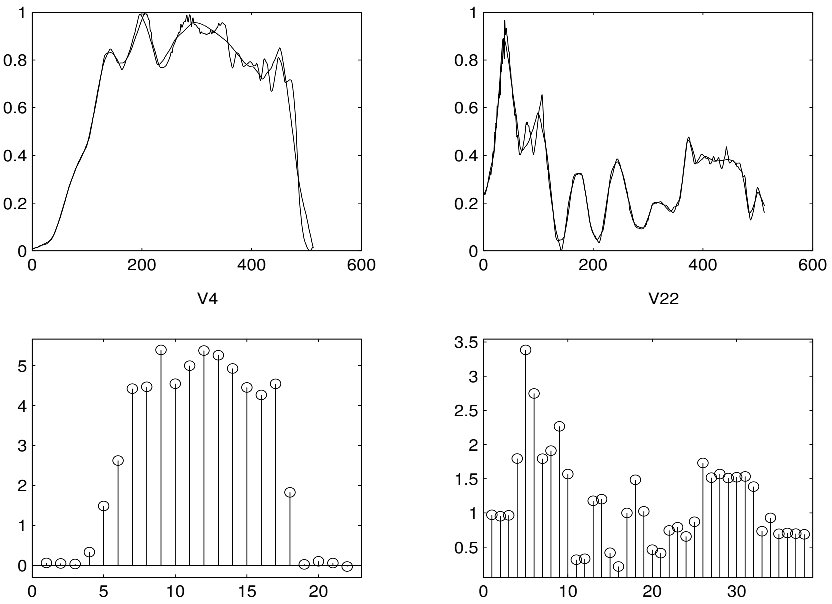

FIG. 3.6 – Pour l'essai 7 et pour les deux variables correspondant à $j = 4, 22$: en haut, le signal après compression superposé au signal original (prétraité), en bas les coefficients d'approximation associés.

Les deux graphiques du haut de la **Figure 3.6** contiennent, pour l'essai 7 et pour deux variables fonctionnelles différentes, le signal après compression superposé au signal d'origine. Ils sont très proches bien que représentés par peu de coefficients. En effet, les deux graphiques du bas de la figure contiennent les coefficients d'approximation associés aux représentations comprimées. Ainsi, la forme des graphiques du haut et du bas de la figure se ressemblent sauf aux extrémités de l'axe des abscisses à cause d'extra-coefficients, engendrés par les prolongements appliqués aux signaux dans les calculs des coefficients par la transformée en ondelettes discrète (voir [73]).

Remarque 3.3.4
Signalons que la connexion entre les développements sur des bases orthogonales d'ondelettes de processus stochastiques et les décompositions issues de la transformée discrète en ondelettes est donnée, par exemple, dans Amato *et al.* [3]. □

Remarque 3.3.5
Une autre approche associant plus étroitement les phases de compression et de sélection des variables discriminantes est proposée par Coifman, Saito [23]. Il s'agit de choisir une base optimale, parmi les bases associées à une décomposition en paquets d'ondelettes, en maximisant la séparation entre classes. □

Elle n'est pas retenue ici, une voie médiane est empruntée : des gains massifs en compression sont obtenus même au prix d'une politique de sélection un peu conservative de façon à ne pas trop obérer la phase suivante qui fera le choix des variables les plus discriminantes. On note $C^j = (C^{j,1}, ..., C^{j,K_j})$ le paquet des K_j coefficients associés à la variable fonctionnelle X^j.

3.3.3 Sélection de variables par CART

A la fin de l'étape précédente, il y a une réduction de la dimension de l'espace des variables, mais elle demeure insuffisante puisque l'on dispose de 114 individus à comparer à 300 ou 400 variables.

Les nouvelles données ainsi construites sont donc : $(((C_i^{j,k})_{1\leq k\leq K_j})_{1\leq j\leq J}, Y_i)_{1\leq i\leq n}$.

On propose une procédure pas à pas basée sur la méthode CART. Celle-ci permet d'ajuster aux données, un modèle additif du type $Y = F((C^{j,k})_{j,k})$ où F est additive et plus précisément constante sur des polyèdres dont les côtés sont parallèles aux axes, sous la forme d'un arbre binaire de décision. On peut se reporter au livre de Breiman *et al.* [20] les fondateurs de la méthode ou Hastie *et al.* [56] pour un rapide aperçu. Dans la suite, on considère l'erreur de classification définie comme usuellement mais en pénalisant les fausses classifications par le truchement de la matrice de coût définie par $\Gamma(k, k') = |k - k'|$, définition qui découle naturellement du fait que Y est une variable ordinale discrète.

La procédure est présentée ci-dessous en cinq phases :

1. Pour chaque j, on construit l'arbre CART A^j expliquant Y par le paquet de coefficients C^j et on sélectionne, au moyen de l'importance des variables au sens de Breiman *et al.* [20] (voir aussi [47] et [46]), le paquet des coefficients utiles, noté $\tilde{C}^j$, en seuillant l'importance comme illustré dans la **Figure 3.7**.

   On peut noter que les pics dans les graphes de l'importance des variables correspondent non pas seulement, à des marqueurs significatifs de la forme du signal mais bien à des événements significatifs discriminants.

2. On en déduit un ordre sur les “nouvelles” variables fonctionnelles (c'est-à-dire sur les paquets $(\tilde{C}^j)_j$) au moyen de l'erreur de classification, évaluée par validation croisée, commise par l'arbre A^j (voir **Figure 3.8**).

3. On construit une suite ascendante $(M^j)_j$ d'au plus $J = 21$ modèles CART emboîtés, en invoquant et en testant les paquets de variables $\tilde{C}^j$, pas à pas, suivant l'ordre précédemment obtenu. Autrement dit, M^j explique Y par l'ensemble de paquets de coefficients $(\tilde{C}^l)_{l\leq j}$ privés des paquets qui se sont révélés, après test, comme insuffisamment informatifs.

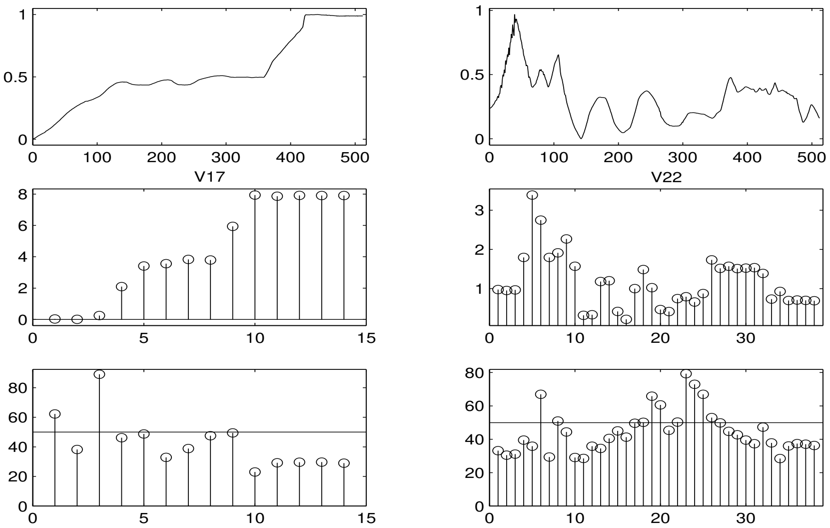

FIG. 3.7 – Pour les essais 7 et 19, et pour les variables correspondant à $j = 17, 22$, en haut les signaux prétraités, au milieu le paquet C^j des coefficients d'approximation de niveau retenu et en bas l'importance de chacun de ces coefficients. Les coefficients utiles constituant $\tilde{C}^j$ sont ceux dont l'importance dépasse le seuil.

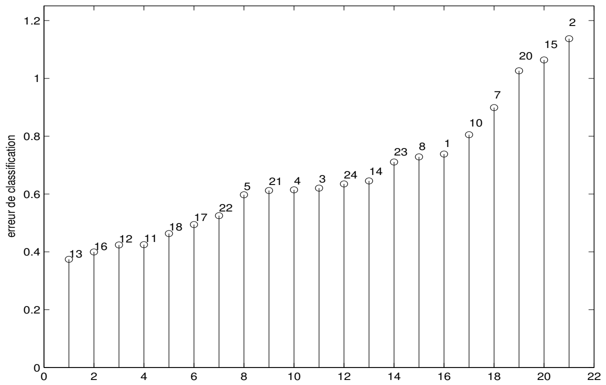

FIG. 3.8 – L'erreur de classification évaluée par validation croisée des arbres A^j, de la meilleure à la pire. Elle fluctue dans un rapport de 1 à 3. Cet ordre sur les "nouvelles" variables fonctionnelles est celui qui sera utilisé pour les invoquer pas à pas.

4. On sélectionne ensuite les variables fonctionnelles pertinentes en choisissant celles définissant le modèle M^{j_0} minimisant l'erreur de classification. L'allure de celle-ci (cf. **Figure 3.9**) est attendue : elle décroît d'abord fortement avant de lentement croître lorsque les variables introduites n'apportent plus rien à la discrimination.

5. Enfin, en calculant l'importance des variables explicatives du modèle M^{j_0} : les coefficients $\{\tilde{C}^j, j \in M^{j_0}\}$ et en retenant la tête de ce classement, on sélectionne les critères

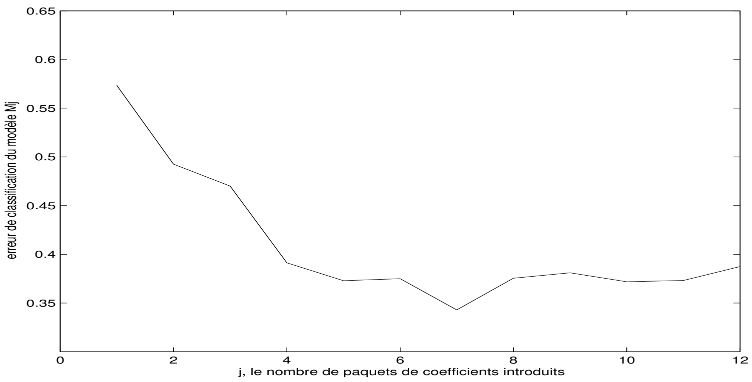

FIG. 3.9 – L'erreur de classification du modèle M^j évaluée par validation croisée, en fonction de j, le nombre de paquets de coefficients introduits.

pertinents (voir **Figure 3.10**).

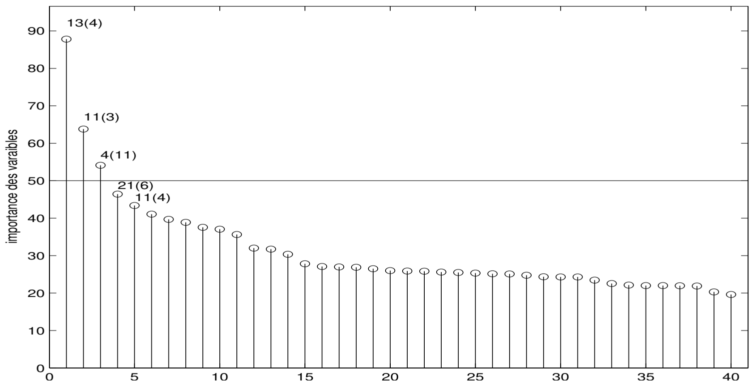

FIG. 3.10 – Importance des variables calculée sur le modèle M^{j_0} sélectionné précédemment et sélection finale des trois critères dont les importances ressortent nettement en tête.

Une première façon de procéder, très dépendante du problème, consiste à ne retenir que les 5 premières variables, 5 étant le nombre souhaité de critères. On obtient alors un arbre dont l'erreur de validation croisée est de 24 sur 114 pour 12 erreurs apparentes (c'est-à-dire l'erreur de resubstitution).
Une alternative consiste à considérer l'erreur de validation croisée sur la suite de modèles emboîtés induite par l'ordre issu du calcul de l'importance des variables. On sélectionne alors le modèle dont l'erreur est la plus faible.

La **Table 3.1** donne, pour les modèles de cette suite dont le nombre de variables est inférieur à

nombre de variables sélectionnées	2	3	4	5	6	7	8	9	10	11	12	13	14	15
erreur apparente	35	15	41	12	11	11	13	9	13	9	8	8	12	7
validation croisée	47	40	30	24	29	27	21	19	25	21	17	19	21	21

TAB. 3.1 – Nombre d'erreurs commises sur l'échantillon d'apprentissage en fonction du nombre de variables retenues.

15, l'erreur apparente et l'erreur de validation croisée. Le meilleur modèle est celui comportant 12 variables. L'erreur commise est de 17 sur 114 (15%) et l'erreur apparente de 8 sur 114 (7%), ce qui est très satisfaisant.
Enfin, si l'on examine l'arbre CART construit en se restreignant à ces 12 variables (cf. **Figure 3.11**), il est intéressant de noter que 5 variables seulement étiquettent les nœuds de l'arbre et 4 d'entre elles sont en tête du classement fourni par la **Figure 3.10**.

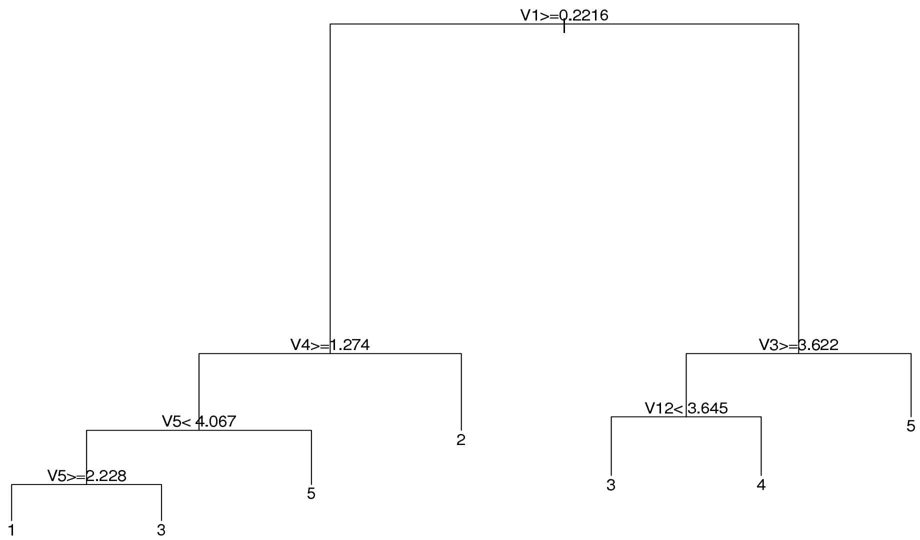

FIG. 3.11 – Arbre dont l'erreur de classification, évaluée par validation croisée, est la plus faible.

Remarque 3.3.6
Terminons par une remarque générale dont la portée méthodologique est cruciale.
Un inconvénient classique de l'usage des arbres de classification est leur instabilité, c'est-à-dire que le classifieur construit peut fluctuer "beaucoup" pour des "petites" variations de l'échantillon d'apprentissage (cf. Hastie *et al.* [56]). Un remède désormais classique à cette propriété indésirable est d'utiliser le bagging qui permet de stabiliser la prédiction en utilisant non pas un classifieur mais l'agrégé d'un ensemble de classifieurs construits par rééchantillon-

nage bootstrap de l'échantillon d'apprentissage (voir Ghattas [46]).

Suivant cette idée (voir Ghattas [47]), l'importance des variables et l'erreur de classification sont évaluées par rééchantillonnage. Plus précisément, pour la phase 1, on considère la moyenne des importances des variables calculées sur des arbres obtenus par rééchantillonnage n pour n, des 114 observations. Pour l'estimation de l'erreur de classification, elle est évaluée par validation croisée grâce à un schéma de découpage en 10 de l'échantillon puis stabilisée en randomisant cette phase de découpage. □

3.4 Conclusion

Du point de vue de l'application, les critères qui ressortent comme les plus discriminants sont associés à quatre variables fonctionnelles. Parmi eux, deux sont très proches des critères obtenus par la méthode basée sur la méthode discriminante linéaire et deux sont nouveaux et considérés par les experts comme intéressants. Il faut noter que dans notre cas, ces critères ont été obtenus sans intégrer de connaissances *a priori*, sauf dans la phase de troncature de la grille temporelle des observations. Signalons cependant que les conditions d'arrêt dépendent de seuils fixés pour le moment en fonction de l'application.

Complémentairement à ce travail, des avancées concernent l'étude théorique de pénalités adéquates pour faire de la sélection de variables dans des contextes voisins (cf. **Chapitre 2**). Typiquement il s'agit d'utiliser une approche par sélection de modèle "à la Birgé-Massart" (cf. Barron, Birgé, Massart [7]) pour sélectionner des variables dans un modèle de régression non linéaire, au moyen d'applications répétées de la méthode CART. Des résultats de type inégalités oracles permettent de préciser la forme des pénalités convenables et peuvent suggérer des alternatives au choix ad-hoc effectués ici.

Remerciements
Les auteurs remercient la Direction de la Recherche de Renault d'avoir mis à leur disposition les données relatives aux essais qui motivent ce travail et, en particulier, Nadine Ansaldi pour les discussions associées.
Cette collaboration se poursuit actuellement dans le cadre d'un contrat de recherche entre le Laboratoire de Mathématiques d'Orsay et la Direction de la Recherche de Renault.

3.5 Complément aux chapitres 2 et 3

Lors de l'étape numéro 5 de la phase de sélection de variables par CART, on commence par calculer l'importance des variables explicatives du modèle M^{j_0}, chacune de ces variables étant un coefficient d'ondelette, autrement dit, une variable scalaire.
Par conséquent, une alternative à la sélection finale par seuillage de l'importance des variables consiste à mettre en application la procédure de sélection de variables décrite dans le **Chapitre 2**.

Nous considérons alors les $p = 15$ variables $Z^1, \ldots, Z^p$ qui sont les 15 plus importants coefficients d'ondelettes du modèle M^{j_0} au sens de l'importance définie par Breiman *et al.*.
Ensuite, en raison de la contrainte imposée par les industriels, à savoir le faible nombre de critères, nous considérons la famille $\mathcal{P}^*$ composée de tous les paquets de taille inférieure ou égale à 10 qu'il est possible de constituer à partir des 15 variables $Z^1, \ldots, Z^p$.

Les résultats obtenus par la procédure de sélection de variables proposée par Sauvé et Tuleau (voir **Chapitre 2**) sont résumés dans la **Table 3.2**.

signaux retenus	13	11
coefficients associés	4	4

TAB. 3.2 – Variables sélectionnées à l'issue de la procédure automatique de sélection de variables décrite dans le **Chapitre 2**.

Autrement dit, la procédure de sélection de variables retient uniquement deux coefficients d'ondelettes associés respectivement aux variables fonctionnelles 13 et 11, ces deux coefficients s'avérant être ceux de rang 1 et 5 dans l'importance définie par Breiman.

Ce résultat n'est pas en désaccord avec les résultats obtenus à l'issue de notre démarche puisque ces deux coefficients étaient retenus dans les méthodes envisagées précédemment. Simplement, ils étaient alors accompagnés d'autres coefficients. L'étape numéro 3 de la phase de sélection par CART peut apporter une explication à cette différence.
En effet, lors de cette phase, on construit une suite de modèles M^j en invoquant et testant les paquets de variables $\tilde{C}^j$. Or, lors de ces tests, on a fixé un seuil arbitraire qui n'était peut-être pas suffisamment élevé pour éliminer tous les effets de "dépendance" entre les coefficients. Ceci pourrait par conséquent expliquer pourquoi une procédure quasi exhaustive ne prend pas en compte les coefficients intermédiaires dans le classement induit par l'importance des variables.

Par ailleurs, si l'on compare les performances de l'arbre associé à ce paquet de 2 variables, on constate des résultats assez proches puisque cet arbre commet 25 erreurs sur 114 évaluées par validation croisée et 10 erreurs apparentes sur 114.

Ainsi, la mise en application de la procédure exhaustive de sélection de variables, proposée au **Chapitre 2**, aura permis d'une part de valider le fait que seulement quelques grandeurs physiques sont représentatives de l'agrément de conduite. D'autre part, cette procédure ne tient pas réellement compte de l'ordre établi par l'importance des variables lors de l'étape 5 de la phase de sélection de variables par CART de ce chapitre. Pour autant, les résultats sont très proches, et notamment le coefficient le plus important au sens de Breiman est sélectionné dans chacune des méthodes proposées.

D'autre part, puisque les résultats obtenus semblent cohérents avec l'application considérée, cela nous permet de conforter la validité, sur des données réelles, de la procédure de sélection de variables proposées au **Chapitre 2**, procédure testée, jusqu'à présent, uniquement sur des données simulées.

Chapitre 4

k-Nearest Neighbor for functional data

Sommaire

Ce chapitre présente un travail en collaboration avec Magalie Fromont qui fera prochainement l'objet d'un article.

Ce chapitre aborde le thème de la classification de données fonctionnelles à l'aide des k-plus proches voisins et poursuit l'étude menée par Biau *et al.* [12].
Comme dans ce travail, nous commençons par projeter nos données dans un espace de dimension d inconnue, puis nous appliquons une procédure de type k-plus proches voisins aux projections. Il faut noter que l'on doit simultanément sélectionner la dimension d et le nombre de voisins k.
Théoriquement, nous montrons qu'envisager une procédure des k-plus proches légèrement pénalisée a des performances comparables à la procédure classique, autrement dit non pénalisée. Par ailleurs, un travail sur des données réelles et simulées montre qu l'ajout d'un léger terme de pénalité permet d'obtenir une meilleure stabilité de la dimension d sélectionnée.

4.1 Introduction

Let (X, Y) be a random pair of variables such that X takes its values in a measurable space $\mathcal{X}$ and Y in $\{0, 1\}$, with unknown distribution denoted by P. Given n independent copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of (X, Y), the purpose of classification is to construct a function, called a *classifier*, $\phi_n : \mathcal{X} \to \{0, 1\}$ based on $(X_1, Y_1), \ldots, (X_n, Y_n)$, which allows to predict the value of Y from the observation of X. When we do not assume that the value of Y is fully determined by X and $(X_1, Y_1), \ldots, (X_n, Y_n)$, the prediction suffers from the classification error defined by $L(\phi_n) = \mathbb{P}[\phi_n(X) \neq Y | (X_i, Y_i), i = 1 \ldots n]$. Introducing the regression function $\eta : x \mapsto \mathbb{P}[Y = 1 | X = x]$, the function ϕ^* that minimizes the classification error $L(\phi) = \mathbb{P}[\phi(X) \neq Y]$ over all the measurable functions $\phi : \mathcal{X} \to \{0, 1\}$ is defined by $\phi^*(x) = \mathbb{1}_{\eta(x) > 1/2}$. In statistical terms, classification deals with the estimation of this function ϕ^* which is called the *Bayes classifier* from the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, and the theoretical performance of any estimator ϕ_n can be evaluated by comparing $\mathbb{E}[L(\phi_n)]$ with $L(\phi^*)$. In particular, ϕ_n is said to be universally consistent if $\mathbb{E}[L(\phi_n)]$ tends to $L(\phi^*)$ as n tends to ∞ whatever the distribution P.

The case where $\mathcal{X} = \mathbb{R}^d$ has been widely studied and many references are devoted to it (see Devroye *et al.* [29] for a review). In many real-world applications, coming from speech recognition or food industry issues for example, data that have to be classified are better represented by discretized functions than by standard vectors. In this paper, we focus on such applications, and hence we assume that $\mathcal{X}$ is a functional space, for example $\mathbb{L}_2(I)$, where I is an interval.

The most simple and popular classifiers are probably the kernel and k-Nearest Neighbor rules. From a theoretical point of view, these rules are known to be universally consistent when $\mathcal{X} = \mathbb{R}^d$ since Stone's [84] and Devroye and Krzyżak's [30] striking results. But the issue is not so clear when $\mathcal{X}$ is an infinite dimensional space, and authors investigate it only since a few years. Dabo-Niang and Rhomari [26], and Abraham *et al.* [1] deal with kernel rules. In particular, Abraham *et al.* [1] show that the moving window classifier can not be universally consistent without restrictive conditions on the functional space $\mathcal{X}$ and the distribution of X. As for k-Nearest Neighbor rules, Cover and Hart [25] first consider Banach space valued data, but they do not establish consistency. Kulkarni and Posner [61] give convergence rates of the k-Nearest Neighbor regression estimator under some regularity conditions on η and the support of the distribution of X. With slight changes in the proofs by Abraham *et al.* [1], one can obtain a universal consistency result for the k-Nearest Neighbor rule when η is assumed to be continuous. Cerou and Guyader [22] finally establish the consistency property in a separable space under a less restrictive regularity condition termed the Besicovich condition.

However, such direct approaches suffer from the phenomenon commonly referred to as the *curse of dimensionality*. Thus, they are not expected to achieve good rates of convergence. The reader will find further details about the curse of dimensionality in the book by Hastie *et al.* [56] page 22 for instance.

To overcome this difficulty, most of the traditional effective methods for $\mathbb{R}^d$-valued data analysis have been adapted to handle functional data under the general name of Functional Data Analysis. A key reference for this growing research field is the series of books by Ramsay and

Silverman [78] and [79]. Using ideas related to the functional canonical correlation analysis approach, Hastie *et al.* [53] develop functional versions of optimal scoring and linear discriminant analysis for classification. Nonlinear classification schemes have then been exploited. Hall *et al.* [51] method combines functional Principal Components Analysis with a usual kernel density estimator. Ferraty and Vieu [39] extend this approach by introducing more general kernel-type estimators. Hastie and Tibshirani [55] propose to use the k-Nearest Neighbor rule or any other neighborhood-based classifier, with some modified neighborhoods which are computed via a local linear discriminant analysis.
In a recent paper, Biau *et al.* [12] propose to filter the functional data X_i in the Fourier basis and to apply the k-Nearest Neighbor rule to the first d coefficients of the expansion. The choice of both the appropriate dimension d and number of neighbors k is made automatically by a minimization of a penalized empirical classification error performed after some data-splitting device. The resulting classifier is proved to satisfy an oracle type inequality, and hence to be universally consistent. As pointed out by the authors, similar results could be obtained for other universally consistent classification procedures in finite dimension. In this spirit, the Support Vector Machines procedures are investigated by Rossi and Villa [82].

The approach of Biau *et al.* [12] is central to our paper. After a careful study of this work, two main issues remain unsolved. The authors underline their preferring to implement the procedure based on the minimization of the empirical classification error without any penalization. However, the theoretical properties presented do not explain why this choice is expected to give better results. Furthermore, as pointed out in the beginning of the simulation study and previously by Hengartner *et al.* [58] for the question of bandwidth selection in local linear regression smoothers, the data-splitting device can be unstable. To overcome this problem, Biau *et al.* [12] suggest to consider many different random splits of the data, and then to combine the corresponding classifiers. This combination technique is commonly used in practice, nevertheless it does not have theoretical background in the present classification context. In this paper, we address to both issues.

In **Section 4.2**, from a recent result due to Bousquet, Boucheron, and Massart (see [70]), we propose a theoretical justification for the improvements observed when choosing to minimize the nonpenalized empirical classification error rather than the penalized one. The result actually shows that the penalty of order $n^{-1/2}$ considered by Biau *et al.* [12] is too large when some margin-type assumptions are satisfied. This suggests to take a penalty equal to zero, or possibly of order smaller than $n^{-1/2}$.
In **Section 4.3**, we illustrate this theoretical advance with an experimental study. We deal with realistic data coming from speech recognition or food industry contexts first and then with simple simulated data.
Section 4.4 is devoted to the problem of the instability of the data-splitting device. Our intention is not to give a justification for the combination technique used by Biau, Bunea, and Wegkamp's [12], since we are not able to. In fact, we propose another stabilization process, which is based on a small order penalization, such as the one allowed by the above theoretical result. The experiments are still performed on both simple simulated and realistic data.

4.2 Functional classification via (non)penalized criteria

In this section, we present Biau, Bunea, and Wegkamp's [12] classification scheme that we consider all along the paper. After briefly describing the procedure and the theoretical properties obtained by Biau, Bunea, and Wegkamp in a general context, we investigate them under some margin-type assumptions such as the ones introduced by Mammen and Tsybakov [68], Tsybakov [86] or Massart and Nédélec [71].

4.2.1 Functional classification in a general context

By using the same framework and notation as in the introduction, we assume that $\mathcal{X}$ is an infinite dimensional separable space. We consider a complete system of $\mathcal{X}$ that we denote by $\{\psi_j,\ j \in \mathbb{N}^*\}$. For every i in $\{1, \ldots, n\}$, X_i can thus be expressed as a series expansion $X_i = \sum_{j=1}^{\infty} X_{i,j}\psi_j$ and for d in $\mathbb{N}^*$, we set $\mathbf{X}_i^d = (X_{i,1}, \ldots, X_{i,d})$. In the same way, $\mathbf{x}^d$ denotes the first d coefficients in the expansion of any new element x in $\mathcal{X}$. The procedure developed by Biau *et al.* [12] is described as follows.

- The data are split into a training set $\mathbb{D}_{\mathcal{T}_l} = \{(X_i, Y_i), i \in \mathcal{T}_l\}$ of length l and a validation set $\mathbb{D}_{\mathcal{V}_m} = \{(X_i, Y_i), i \in \mathcal{V}_m\}$ of length m such that $n = l + m$ with $1 \leq l \leq n - 1$.
  A usual choice for l will be $n/2$ when n is even, $(n+1)/2$ when n is odd.

- For each k in $\{1, \ldots, l\}$, d in a subset $\mathcal{D}$ of $\mathbb{N}^*$, let $\hat{p}_{l,k,d}$ be the k-Nearest Neighbor rule on $\mathbb{R}^d$ constructed from the set $\{(\mathbf{X}_i^d, Y_i), i \in \mathcal{T}_l\}$. Let $\mathbf{x}$ be an element of $\mathbb{R}^d$. The set $\{(\mathbf{X}_i^d, Y_i), i \in \mathcal{T}_l\}$ is reordered according to increasing Euclidean distances $\|\mathbf{X}_i^d - \mathbf{x}\|$, and the reordered variables are denoted by $(\mathbf{X}_{(1)}^d(\mathbf{x}), Y_{(1)}(\mathbf{x})), \ldots, (\mathbf{X}_{(l)}^d(\mathbf{x}), Y_{(l)}(\mathbf{x}))$. Thus $\mathbf{X}_{(k)}^d(\mathbf{x})$ is the k-th nearest neighbor of $\mathbf{x}$ amongst $\{\mathbf{X}_i^d, i \in \mathcal{T}_l\}$. When $\|\mathbf{X}_{i_1}^d - \mathbf{x}\| = \|\mathbf{X}_{i_2}^d - \mathbf{x}\|$, $\mathbf{X}_{i_1}^d$ is declared closer to $\mathbf{x}$ if $i_1 < i_2$. Then $\hat{p}_{l,k,d}(\mathbf{x})$ is defined by

$$\hat{p}_{l,k,d}(\mathbf{x}) = \begin{cases} 0 & \text{if } \sum_{i=1}^{k} \mathbb{1}_{Y_{(i)}(\mathbf{x})=0} \geq \sum_{i=1}^{k} \mathbb{1}_{Y_{(i)}(\mathbf{x})=1} \\ 1 & \text{otherwise.} \end{cases}$$

  We introduce the corresponding functional classifier defined by

$$\hat{\phi}_{l,k,d}(x) = \hat{p}_{l,k,d}(\mathbf{x}^d) \text{ for all } x \text{ in } \mathcal{X}.$$

- The appropriate k and d are simultaneously selected from the validation set by minimizing a penalized empirical classification error :

$$(\hat{k}, \hat{d}) = \underset{k \in \{1, \ldots, l\},\ d \in \mathcal{D}}{\operatorname{argmin}} \left(\frac{1}{m} \sum_{i \in \mathcal{V}_m} \mathbb{1}_{\{\hat{\phi}_{l,k,d}(X_i) \neq Y_i\}} + \operatorname{pen}(d) \right), \tag{4.1}$$

  where $\operatorname{pen}(d)$ is a positive penalty term that can be equal to zero.

- The final classifier is defined by

$$\hat{\phi}_n(x) = \hat{\phi}_{l,\hat{k},\hat{d}}(x) \text{ for all } x \text{ in } \mathcal{X}. \tag{4.2}$$

Let us now recall the central result of the paper by Biau *et al.* [12].

Proposition 4.2.1 (Biau, Bunea, Wegkamp)
Introduce $(\lambda_d, d \in \mathbb{N}^*)$ such that $\Delta = \sum_{d \in \mathbb{N}^*} e^{-2\lambda_d^2} < +\infty$. Let $l > 1/\Delta$, m with $l + m = n$, and $\hat{\phi}_n$ be the classification rule defined by (4.2) with $\mathcal{D} = \mathbb{N}^*$ and $pen(d) = \lambda_d/\sqrt{m}$ (penalized case). Then there exists a constant $c(\Delta) > 0$ such that

$$\mathbb{E}[L(\hat{\phi}_n)] - L(\phi^*) \leq$$
$$\inf_{d \in \mathbb{N}^*} \left\{ L_d^* - L(\phi^*) + \inf_{1 \leq k \leq l} \left\{ \mathbb{E}[L(\hat{\phi}_{l,k,d})] - L_d^* \right\} + pen(d) \right\} + c(\Delta)\sqrt{\frac{\log l}{m}}, \quad (4.3)$$

where L_d^* is the minimal classification error when the feature space is $\mathbb{R}^d$.

The same result holds when $\hat{\phi}_n$ is the classification rule defined by (4.2) with $\mathcal{D} = \{1, \ldots, d_n\}$ and $pen(d) = 0$ (nonpenalized case), but at the price that the last term $c(\Delta)\sqrt{\log l/m}$ is replaced by $c(\Delta)\sqrt{\log l d_n/m}$. □

The quantity $L_d^* - L(\phi^*)$ can be viewed as an approximation term. By using some classical martingale arguments, one proves that it tends to 0 as d tends to ∞. Moreover, from Stone's [84] consistency result in $\mathbb{R}^d$, one deduces that for d in $\mathbb{N}^*$, $\mathbb{E}[L(\hat{\phi}_{l,k,d})]$ tends to L_d^* as $l \to \infty$, $k \to \infty$, $k/l \to 0$ whatever the distribution P. The classifier $\hat{\phi}_n$ is thus universally consistent. Precisely, if $\lim_{n\to\infty} l = \infty$, $\lim_{n\to\infty} m = \infty$ and $\lim_{n\to\infty} \log l/m = 0$ in the penalized case or $\lim_{n\to\infty} \log(l d_n)/m = 0$ in the nonpenalized case, then

$$\lim_{n\to\infty} \mathbb{E}[L(\hat{\phi}_n)] = L(\phi^*) \text{ for any distribution } P.$$

A universal strong consistency result can also be obtained under a mild condition on the distribution of X.

Of course, such results are fully satisfactory from an asymptotic point of view. In the following section, we consider the proposed classifier from a nonasymptotic point of view. In particular, in order to evaluate the accuracy of the rates achieved in (4.3), a brief overview of the present knowledge on the minimax bounds in the general classification framework is given.

4.2.2 Minimax bounds

Classical bounds for Vapnik-Chervonenkis classes

We consider here a class of classifiers $\mathcal{S} = \{\mathbb{1}_C, C \in \mathcal{C}\}$, where $\mathcal{C}$ is a class of subsets of $\mathcal{X}$, with Vapnik-Chervonenkis dimension $V(\mathcal{C}) < \infty$ (see Vapnik [91] or Devroye *et al.* [29] for further details on Vapnik-Chervonenkis classes).

A first risk bound was given by Vapnik and Chervonenkis [92] for the Empirical Risk Minimizer (ERM) defined by $\hat{\varphi}_n = \operatorname{argmin}_{\phi \in \mathcal{S}} \sum_{i=1}^n \mathbb{1}_{Y_i \neq \phi(X_i)}$. For any distribution P such that $\phi^* \in \mathcal{S}$, one has $\mathbb{E}[L(\hat{\varphi}_n)] - L(\phi^*) \leq \kappa_1 \sqrt{V(\mathcal{C}) \log n/n}$. Moreover, as pointed out by Lugosi

[64], the factor $\log n$ in the upper bound can be removed by using some classical chaining arguments, and the bound becomes optimal in a minimax sense. Vapnik and Chervonenkis [89] actually proved that for any classifier φ_n,

$$\sup_{P,\, \phi^* \in \mathcal{S}} \mathbb{E}[L(\varphi_n)] - L(\phi^*) \geq \kappa_2 \sqrt{\frac{V(\mathcal{C})}{n}} \quad \text{if } n \geq \kappa_3 V(\mathcal{C}).$$

In the following, we call this case the *global case*, since the risk bounds given here are obtained without any restriction on the class $\mathcal{S}$ except it is based on a VC class.

The global case is overpessimistic in the sense that the two preceding inequalities are obtained considering all the distribution P such ϕ^* belongs to $\mathcal{S}$. With some assumption on these distributions P the results can be improved.
For example, considering the over-optimistic situation called the *zero error case* which consists of assuming that $Y = \eta(X)$, Devroye and Wagner [33], Vapnik [88], and Blumer *et al.* [17] obtained various forms of the following result : the ERM $\hat{\varphi}_n$ has a mean classification error not larger than $\kappa_4 V(\mathcal{C}) \log n/n$. Up to a logarithmic factor, this upper bound is known to be optimal in a minimax sense since Vapnik and Chervonenkis [89] and Haussler *et al.* [57] established that for any classifier φ_n,

$$\sup_{P,\, \phi^* \in \mathcal{S},\, L(\phi^*)=0} \mathbb{E}[L(\varphi_n)] \geq \kappa_5 \frac{V(\mathcal{C})}{n} \quad \text{if } n \geq \kappa_6 V(\mathcal{C}).$$

The main point here is that the minimax risk in the zero-error case is of smaller order of magnitude than in the global case, and that the difference is really significant ($V(\mathcal{C}) \log n/n$ instead of $\sqrt{V(\mathcal{C})/n}$). This leads to the intuition that if the Bayes classification error is not exactly equal to zero but very small, the bounds in the global case can be refined.

Refined bounds for Vapnik-Chervonenkis classes

The following result is due to Lugosi [64]. For any distribution P such that $\phi^* \in \mathcal{S}$, the ERM $\hat{\varphi}_n$ satisfies

$$\mathbb{E}[L(\hat{\varphi}_n)] - L(\phi^*) \leq \kappa_7 \left(\sqrt{\frac{L(\phi^*) V(\mathcal{C}) \log n}{n}} + \frac{V(\mathcal{C}) \log n}{n} \right).$$

The corresponding minimax lower bound was obtained by Devroye and Lugosi [32]. Given $L_0 \in]0, 1/2[$, for any classifier φ_n,

$$\sup_{P,\, \phi^* \in \mathcal{S},\, L(\phi^*)=L_0} \mathbb{E}[L(\varphi_n)] - L(\phi^*) \geq \kappa_8 \sqrt{\frac{L_0 V(\mathcal{C})}{n}} \quad \text{if } n \geq \kappa_9 \frac{V(\mathcal{C})}{L_0(1-2L_0)^2}.$$

These bounds can be viewed as a kind of interpolation between the global case and the zero-error one. In particular, we can notice that (up to a possible logarithmic factor) the minimax risk can be of order $V(\mathcal{C})/n$ when $L(\phi^*)$ is of order $V(\mathcal{C})/n$.

By carefully studying the proofs of the above results, the behavior of the regression function η around 1/2 turns out to be crucial. Mammen and Tsybakov [68] first analyzed the influence of this behaviour by introducing some *margin* assumptions.

Risk bounds under margin assumptions

Let us now consider such margin assumptions. Let P_X be the marginal distribution of X, and $\mathbb{E}_X$ be the expectation with respect to P_X. For $\theta \geq 1$, we denote by GMA(θ) the general margin assumption introduced by Mammen and Tsybakov [68] and Tsybakov [86].

$$\text{GMA}(\theta): \quad \exists h > 0,\ L(\phi) - L(\phi^*) \geq h\mathbb{E}_X[|\phi(X) - \phi^*(X)|]^\theta,\ \forall \phi : \mathcal{X} \to \{0,1\}.$$

We also introduce some versions of this general margin assumption that can be more easily interpreted with

$$\text{MA}(\alpha): \quad \mathbb{P}[|\eta(X) - 1/2| \leq u] \leq C_1 u^\alpha,\ \forall u > 0.$$

The limit case

$$\text{MA}(\infty): \quad \exists h > 0,\ |2\eta(x) - 1| \geq h,\ \forall x \in \mathcal{X},$$

was studied in details by Massart and Nédélec [71].

From the well-kown equality

$$L(\phi) - L(\phi^*) = \mathbb{E}_X[|2\eta(X) - 1||\phi^*(X) - \phi(X)|],$$

which is satisfied by any classifier $\phi : \mathcal{X} \to \{0,1\}$ (see Devroye *et al.* [29] page 16 for a proof), Tsybakov [86] proves that MA(α) implies GMA($(1+\alpha)/\alpha$) (See Proposition 1 from Tsybakov [86]), and one can easily see that MA(∞) implies GMA(1).

The risk bounds are given under two kinds of complexity assumptions on the set ϕ^* belongs to.

- CA1 : $\phi^* \in \mathcal{S} = \{1\!\!I_C, C \in \mathcal{C}\}$, $\mathcal{C}$ being a class of subsets of $\mathcal{X}$, with VC dimension $V(\mathcal{C}) < \infty$, and there exists some countable subset S of $\mathcal{S}$ such that for every ϕ in $\mathcal{S}$, there exists a sequence $(\phi_k)_{k\geq 1}$ of elements of S such that for all (x, y) in $\mathcal{X} \times \{0,1\}$, $1\!\!I_{y\neq\phi_k(x)}$ tends to $1\!\!I_{y\neq\phi(x)}$ as k tends to ∞.
- CA2: $\phi^* \in \mathcal{S}$, $\mathcal{S}$ satisfying the condition $H(\varepsilon, \mathcal{S}, \mathbb{L}_1(P_X)) \leq C_2\varepsilon^{-\rho}$ for any $\varepsilon > 0$, where $H(\varepsilon, \mathcal{S}, \mathbb{L}_1(P_X))$ denotes the ε-entropy with bracketing of the set $\mathcal{S}$ with respect to the

$\mathbb{L}_1(P_X)$-norm.

Massart and Nédélec [71] establish that if the assumption CA1 is satisfied, and GMA(θ) holds with $\theta \geq 1$ and $h \geq (V(\mathcal{C})/n)^{1/(2\theta)}$, then the ERM $\hat{\varphi}_n$ satisfies :

$$\mathbb{E}[L(\hat{\varphi}_n)] - L(\phi^*) \leq \kappa_{10} \left(\frac{V(\mathcal{C})(1 + \log(nh^{2\theta}/V(\mathcal{C})))}{nh} \right)^{\frac{\theta}{2\theta-1}}. \tag{4.4}$$

They also discuss the optimality of this upper bound in a minimax sense for the special case MA(∞). They prove in particular that for any classifier φ_n,

$$\sup_{P,\, \mathrm{MA}(\infty),\, \mathrm{CA1 \ with}\ V(\mathcal{C}) \geq 2} \mathbb{E}[L(\varphi_n)] - L(\phi^*) \geq \kappa_{11} \left\{ \left(\frac{V(\mathcal{C})}{nh} \right) \wedge \sqrt{\frac{V(\mathcal{C})}{n}} \right\} \text{ if } n \geq V(\mathcal{C}).$$

As for the assumption CA2, Tsybakov [86] obtains the following upper bound for the ERM $\hat{\varphi}_n$ computed on an $\varepsilon-$net on $\mathcal{S}$ with respect to the $\mathbb{L}_1(P_X)$ norm. For $\varepsilon = cn^{-1/(1+\rho)}$, if GMA($\theta$) and CA2 are satisfied, then

$$\mathbb{E}[L(\hat{\varphi}_n)] - L(\phi^*) \leq \kappa_{12} \left\{ n^{-\frac{\theta}{2\theta+\rho-1}} \wedge n^{-1/2} \right\}.$$

Massart and Nédélec [71] refine this result by exhibiting the dependency of the risk bound with respect to the margin parameter h in GMA(θ). This upper bound is then proved to be optimal in a minimax sense by Tsybakov [86] when $\mathcal{X} = [0,1]^d$, and by Massart and Nédélec [71] in the special case MA(∞).

Recently, Tsybakov and Audibert [6] proved that when $\mathcal{X} = \mathbb{R}^d$, under another type of margin assumption (expressed in terms of smoothness for the regression function η), such fast rates of convergence are not only achieved by ERM type classifiers but also by some plug-in classifiers. This result is essential for our study since the k-Nearest Neighbor estimator is not an ERM classifier but a plug-in one.

In view of the above risk bounds, the result in Proposition 4.2.1 gives some rates of convergence that perfectly fit the minimax risk bounds in the global case. However, when we assume that some margin assumption is satisfied for instance, it will not be satisfactory any more. Indeed, on the one hand, one can see that the order of magnitude of the penalty term ($\lambda_d/\sqrt{m}$) in the penalized case is too large as compared to the rates faster than $1/\sqrt{m}$ that are expected under such a margin assumption. On the other hand, in the nonpenalized case, the right hand side of the inequality (4.3) makes a term of order $\sqrt{\log l d_n/m}$ appear. This term can not be seen as a residual term when considering any margin assumption any more, and hence the oracle type inequality (4.3) has to be refined.

4.2.3 Functional classification with margin conditions

The key point in the proof of Proposition 4.2.1 is a general inequality which is at the root of many model selection results. Setting

$$L_m(\phi) = \frac{1}{m} \sum_{i \in \mathcal{V}_m} \mathbb{1}_{\{\phi(X_i) \neq Y_i\}} \text{ for all } \phi : \mathcal{X} \to \{0,1\},$$

one has

$$\begin{aligned} L(\hat{\phi}_n) - L(\phi^*) &\leq L(\hat{\phi}_{l,k,d}) - L(\phi^*) + L(\hat{\phi}_n) - L(\hat{\phi}_{l,k,d}) \\ &\leq L(\hat{\phi}_{l,k,d}) - L(\phi^*) + L(\hat{\phi}_n) - L_m(\hat{\phi}_n) - L(\hat{\phi}_{l,k,d}) + L_m(\hat{\phi}_{l,k,d}) \\ &\quad + L_m(\hat{\phi}_n) - L_m(\hat{\phi}_{l,k,d}). \end{aligned}$$

From the definition (4.2) of $\hat{\phi}_n$, $L_m(\hat{\phi}_n) - L_m(\hat{\phi}_{l,k,d}) \leq \text{pen}(d) - \text{pen}(\hat{d})$, and therefore

$$L(\hat{\phi}_n) - L(\phi^*) \leq L(\hat{\phi}_{l,k,d}) - L(\phi^*) + \text{pen}(d) + L(\hat{\phi}_n) - L_m(\hat{\phi}_n) - L(\hat{\phi}_{l,k,d}) + L_m(\hat{\phi}_{l,k,d}) - \text{pen}(\hat{d}),$$

that is

$$\begin{aligned} L(\hat{\phi}_n) - L(\phi^*) \leq L(\hat{\phi}_{l,k,d}) &- L(\phi^*) + \text{pen}(d) \\ &+ L(\hat{\phi}_{l,\hat{k},\hat{d}}) - L_m(\hat{\phi}_{l,\hat{k},\hat{d}}) - L(\hat{\phi}_{l,k,d}) + L_m(\hat{\phi}_{l,k,d}) - \text{pen}(\hat{d}). \end{aligned} \tag{4.5}$$

Since $L(\hat{\phi}_{l,k,d}) - L_m(\hat{\phi}_{l,k,d})$ is centered, the oracle type inequality (4.3) is obtained by choosing a penalty such that $\text{pen}(\hat{d})$ is large enough to compensate for the quantity $L(\hat{\phi}_{l,\hat{k},\hat{d}}) - L_m(\hat{\phi}_{l,\hat{k},\hat{d}})$, but such that $\text{pen}(d)$ is small enough (of order at most $1/\sqrt{m}$) to fit the minimax risk bounds in the global case. The main issue is then to evaluate the fluctuations of $L(\hat{\phi}_{l,\hat{k},\hat{d}}) - L_m(\hat{\phi}_{l,\hat{k},\hat{d}})$. Biau, Bunea, and Wegkamp use for this Hoeffding's concentration inequality. Their inequality is thus essentially based on the fact that the functions involved are bounded.

The following result is obtained via a Bernstein type inequality which allows to control the fluctuations of the whole quantity $L(\hat{\phi}_{l,\hat{k},\hat{d}}) - L_m(\hat{\phi}_{l,\hat{k},\hat{d}}) - L(\hat{\phi}_{l,k,d}) + L_m(\hat{\phi}_{l,k,d})$ by taking its variance into account.

Proposition 4.2.2
Assume that $n \geq 2$ and let $\hat{\phi}_n$ be the classifier defined by (4.2) with a finite subset $\mathcal{D}$ of $\mathbb{N}^*$ and with the penalty terms $pen(d)$ that can be equal to zero. For any $\beta > 0$, if $GMA(\theta)$ holds with $\theta \geq 1$ and $h \leq 1$, then

$$\begin{aligned} (1-\beta)\mathbb{E}[L(\hat{\phi}_n) - L(\phi^*) | \mathbb{D}_{\mathcal{T}_l}] \leq (1+\beta) &\inf_{k \in \{1,\ldots,l\},\ d \in \mathcal{D}} \left\{ \left(L(\hat{\phi}_{l,k,d}) - L(\phi^*) \right) + pen(d) \right\} \\ &+ \frac{\beta^{-1}(1 + \log(l|\mathcal{D}|)) + 4\beta}{2(mh)^{\frac{\theta}{2\theta-1}}} + \frac{1 + \log(l|\mathcal{D}|)}{3m}. \end{aligned} \tag{4.6}$$

□

Remark 4.2.1

- The oracle type inequality (4.6) can also be expressed in the following form :

$$(1-\beta)\mathbb{E}[L(\hat{\phi}_n)-L(\phi^*)] \leq (1+\beta) \inf_{d\in\mathbb{N}^*} \left\{ L^*_d - L(\phi^*) + \inf_{1\leq k\leq l} \left\{ \mathbb{E}[L(\hat{\phi}_{l,k,d})] - L^*_d \right\} \right.$$
$$\left. + \operatorname{pen}(d) \right\} + \frac{\beta^{-1}\left(1+\log(l|\mathcal{D}|)\right)+4\beta}{2(mh)^{\frac{\theta}{2\theta-1}}} + \frac{1+\log(l|\mathcal{D}|)}{3m}.$$

By using the same arguments as Biau *et al.* [12], this inequality directly leads to a universal consistency result.

From a nonasymptotic point of view, the terms $((2\beta)^{-1}\left(1+\log(l|\mathcal{D}|)\right)+2\beta)(mh)^{-\theta/(2\theta-1)}$ and $(1+\log(l|\mathcal{D}|))/3m$ are at most of the same order as the minimax risk given by (4.4). This guarantees that they can actually be viewed as residual terms in the final risk bound.

As the penalty term $pen(d)$ can be equal to zero, the obtained result proves the efficiency of the procedure when we use a nonpenalized version of the k-Nearest Neighbor.

Furthermore, as soon as the penalty is small enough, that is with a smaller order of magnitude than the residual terms or than $\inf_{1\leq k\leq l}\{\mathbb{E}[L(\hat{\phi}_{l,k,d})]-L^*_d\}$, it will not alterate the rate of convergence. In a recent work, Györfi [50] proved that under some local Lipschitz condition on the regression function η, assuming that the margin condition MA(α) is satisfied,

$$\inf_{1\leq k\leq l}\left\{\mathbb{E}[L(\hat{\phi}_{l,k,d})]-L^*_d\right\} \leq \kappa(\log m)^{\frac{1+\alpha}{2}} m^{-\frac{1+\alpha}{2+d}}.$$

This allows us to consider in practice penalties such that $\operatorname{pen}(d)=0$, $\operatorname{pen}(d)=\log d/m$, $\operatorname{pen}(d)=d^{1/\gamma}/m$ with $\gamma\in\mathbb{N}^*$ or $\operatorname{pen}(d)=d/m^2$ for instance. Some experimental results are presented in **Section 4.3** in order to compare the performance of the classification procedure with these various penalties.

- The proof of **Proposition 4.2.2** follows the same lines as the proof of Corollary 3.2. due to Bousquet, Boucheron, and Massart (see [70]) except that it is applied conditionally given $\mathbb{D}_{\mathcal{T}_l}$ with $\mathbb{D}_{\mathcal{V}_m}$ instead of $\xi_1,\ldots,\xi_n$, $\mathcal{M}=\{(k,d), k\in\{1,\ldots,l\}, d\in\mathcal{D}\}$, $f_{(k,d)}(x,y)=\mathbb{1}_{\hat{\phi}_{l,k,d}(x)\neq y}-\mathbb{1}_{\phi^*(x)\neq y}$ and $w(\varepsilon)=\varepsilon^{1/\theta}h^{-1/2}$. The proof is rather simple in the present case. □

Proof of the proposition 4.2.2:
Starting from the general inequality (4.5), we need to evaluate the fluctuations of $L(\hat{\phi}_{l,\hat{k},\hat{d}}) - L_m(\hat{\phi}_{l,\hat{k},\hat{d}}) - L(\hat{\phi}_{l,k,d}) + L_m(\hat{\phi}_{l,k,d})$. The pair $(\hat{k},\hat{d})$ being randomly selected, it is therefore a matter of controlling $L(\hat{\phi}_{l,k',d'}) - L_m(\hat{\phi}_{l,k',d'}) - L(\hat{\phi}_{l,k,d}) + L_m(\hat{\phi}_{l,k,d})$ uniformly for (k',d') in

$\{1, \ldots, l\} \times \mathcal{D}$.

We use here a special version of Bernstein's inequality given by Birgé and Massart recalled in the lemma 4.2.1 at the end of this proof.
Since

$$L(\hat{\phi}_{l,k',d'}) - L_m(\hat{\phi}_{l,k',d'}) - L(\hat{\phi}_{l,k,d}) + L_m(\hat{\phi}_{l,k,d}) = \frac{1}{m} \sum_{i \in \mathcal{V}_m} \Big(1\!\!\mathrm{I}_{\{\hat{\phi}_{l,k,d}(X_i) \neq Y_i\}} - 1\!\!\mathrm{I}_{\{\hat{\phi}_{l,k',d'}(X_i) \neq Y_i\}}$$
$$- \mathbb{E}\Big[1\!\!\mathrm{I}_{\{\hat{\phi}_{l,k,d}(X_i) \neq Y_i\}} - 1\!\!\mathrm{I}_{\{\hat{\phi}_{l,k',d'}(X_i) \neq Y_i\}} \big| \mathbb{D}_{\mathcal{T}_l} \Big] \Big),$$

we have to find an appropriate upper bound for $\mathbb{E}\Big[\Big| 1\!\!\mathrm{I}_{\{\hat{\phi}_{l,k,d}(X_i) \neq Y_i\}} - 1\!\!\mathrm{I}_{\{\hat{\phi}_{l,k',d'}(X_i) \neq Y_i\}} \Big|^q \big| \mathbb{D}_{\mathcal{T}_l} \Big]$ for every i in $\mathcal{V}_m$, $q \geq 2$.

For any integer $q \geq 2$, we have

$$\mathbb{E}\left[\left| 1\!\!\mathrm{I}_{\{\hat{\phi}_{l,k,d}(X_i) \neq Y_i\}} - 1\!\!\mathrm{I}_{\{\hat{\phi}_{l,k',d'}(X_i) \neq Y_i\}} \right|^q \big| \mathbb{D}_{\mathcal{T}_l} \right] = \mathbb{E}_X \left[\left| \hat{\phi}_{l,k,d}(X) - \hat{\phi}_{l,k',d'}(X) \right| \right],$$

so

$$\mathbb{E}\left[\left| 1\!\!\mathrm{I}_{\{\hat{\phi}_{l,k,d}(X_i) \neq Y_i\}} - 1\!\!\mathrm{I}_{\{\hat{\phi}_{l,k',d'}(X_i) \neq Y_i\}} \right|^q \big| \mathbb{D}_{\mathcal{T}_l} \right]$$
$$\leq \mathbb{E}_X \left[\left| \hat{\phi}_{l,k,d}(X) - \phi^*(X) \right| \right] + \mathbb{E}_X \left[\left| \hat{\phi}_{l,k',d'}(X) - \phi^*(X) \right| \right].$$

Under the assumption GMA(θ), we then have

$$\begin{aligned} &\mathbb{E}\big[\big| 1\!\!\mathrm{I}_{\{\hat{\phi}_{l,k,d}(X_i) \neq Y_i\}} - 1\!\!\mathrm{I}_{\{\hat{\phi}_{l,k',d'}(X_i) \neq Y_i\}} \big|^q \big| \mathbb{D}_{\mathcal{T}_l} \big] \\ &\qquad \leq \left(\frac{L(\hat{\phi}_{l,k,d}) - L(\phi^*)}{h} \right)^{1/\theta} + \left(\frac{L(\hat{\phi}_{l,k',d'}) - L(\phi^*)}{h} \right)^{1/\theta} \\ &\qquad \leq \frac{q!}{2} \left(\frac{1}{3} \right)^{q-2} \left(\left(\frac{L(\hat{\phi}_{l,k,d}) - L(\phi^*)}{h} \right)^{1/\theta} + \left(\frac{L(\hat{\phi}_{l,k',d'}) - L(\phi^*)}{h} \right)^{1/\theta} \right). \end{aligned}$$

Introduce now a collection of positive numbers $\{x_{(k',d')}, k' \in \{1, \ldots, l\}, d' \in \mathcal{D}\}$ such that $\Sigma = \sum_{k' \in \{1,\ldots,l\}, d' \in \mathcal{D}} e^{-x_{(k',d')}} < \infty$. From Lemma 4.2.1, we deduce that for any $x > 0$, $k' \in \{1, \ldots, l\}$, $d' \in \mathcal{D}$, conditionally given $\mathbb{D}_{\mathcal{T}_l}$, with probability at least $\left(1 - e^{-\left(x + x_{(k',d')}\right)}\right)$,

$$\begin{aligned} &L(\hat{\phi}_{l,k',d'}) - L_m(\hat{\phi}_{l,k',d'}) - L(\hat{\phi}_{l,k,d}) + L_m(\hat{\phi}_{l,k,d}) \\ &\leq \sqrt{\frac{2(x + x_{(k',d')})}{m}} \sqrt{\left(\frac{L(\hat{\phi}_{l,k,d}) - L(\phi^*)}{h} \right)^{1/\theta} + \left(\frac{L(\hat{\phi}_{l,k',d'})) - L(\phi^*)}{h} \right)^{1/\theta}} + \frac{1}{3} \frac{x + x_{(k',d')}}{m}. \end{aligned}$$

From the general inequality (4.5), we obtain that conditionally given $\mathbb{D}_{\mathcal{I}_l}$, with probability at least $(1-\Sigma e^{-x})$,

$$L(\hat{\phi}_n)-L(\phi^*)\leq L(\hat{\phi}_{l,k,d})-L(\phi^*)+\mathrm{pen}(d)-\mathrm{pen}(\hat{d})$$
$$+\sqrt{\frac{2\left(x+x_{(\hat{k},\hat{d})}\right)}{m}}\sqrt{\left(\frac{L(\hat{\phi}_{l,k,d})-L(\phi^*)}{h}\right)^{1/\theta}+\left(\frac{L(\hat{\phi}_n)-L(\phi^*)}{h}\right)^{1/\theta}}+\frac{1}{3}\frac{x+x_{(\hat{k},\hat{d})}}{m}. \tag{4.7}$$

Since h is assumed to be smaller than 1, and $\theta\geq 1$, we have

$$\sqrt{\frac{2\left(x+x_{(\hat{k},\hat{d})}\right)}{m}}\sqrt{\left(\frac{L(\hat{\phi}_{l,k,d})-L(\phi^*)}{h}\right)^{1/\theta}+\left(\frac{L(\hat{\phi}_n)-L(\phi^*)}{h}\right)^{1/\theta}}$$
$$\leq\sqrt{\frac{2\left(x+x_{(\hat{k},\hat{d})}\right)}{(mh)^{\frac{\theta}{2\theta-1}}}}\sqrt{\left(\frac{L(\hat{\phi}_{l,k,d})-L(\phi^*)}{h}\right)^{1/\theta}+\left(\frac{L(\hat{\phi}_n)-L(\phi^*)}{h}\right)^{1/\theta}}\sqrt{\frac{(mh)^{\frac{\theta}{2\theta-1}}}{m}}$$
$$\leq\sqrt{\frac{2\left(x+x_{(\hat{k},\hat{d})}\right)}{(mh)^{\frac{\theta}{2\theta-1}}}}\sqrt{\left(\left(L(\hat{\phi}_{l,k,d})-L(\phi^*)\right)^{1/\theta}+\left(L(\hat{\phi}_n)-L(\phi^*)\right)^{1/\theta}\right)(mh)^{\frac{\theta}{2\theta-1}-1}}$$
$$\leq\sqrt{\frac{2\left(x+x_{(\hat{k},\hat{d})}\right)}{(mh)^{\frac{\theta}{2\theta-1}}}}\sqrt{\left(\left(L(\hat{\phi}_{l,k,d})-L(\phi^*)\right)^{1/\theta}+\left(L(\hat{\phi}_n)-L(\phi^*)\right)^{1/\theta}\right)\left(\frac{1}{(mh)^{\frac{\theta}{2\theta-1}}}\right)^{1-\frac{1}{\theta}}}.$$

By successively using the elementary inequalities $a^{\frac{1}{\theta}}b^{1-\frac{1}{\theta}}\leq a+b$ and $\sqrt{ab}\leq\beta^{-1}a/4+\beta b$ for $a\geq 0$, $b\geq 0$, $\beta>0$, we establish that

$$\sqrt{\frac{2\left(x+x_{(\hat{k},\hat{d})}\right)}{(mh)^{\frac{\theta}{2\theta-1}}}}\sqrt{\left(\left(L(\hat{\phi}_{l,k,d})-L(\phi^*)\right)^{1/\theta}+\left(L(\hat{\phi}_n)-L(\phi^*)\right)^{1/\theta}\right)\left(\frac{1}{(mh)^{\frac{\theta}{2\theta-1}}}\right)^{1-\frac{1}{\theta}}}$$
$$\leq\sqrt{\frac{2\left(x+x_{(\hat{k},\hat{d})}\right)}{(mh)^{\frac{\theta}{2\theta-1}}}}\sqrt{L(\hat{\phi}_{l,k,d})-L(\phi^*)+L(\hat{\phi}_n)-L(\phi^*)+\frac{2}{(mh)^{\frac{\theta}{2\theta-1}}}}$$
$$\leq\frac{\beta^{-1}\left(x+x_{(\hat{k},\hat{d})}\right)}{2(mh)^{\frac{\theta}{2\theta-1}}}+\beta\left(L(\hat{\phi}_{l,k,d})-L(\phi^*)+L(\hat{\phi}_n)-L(\phi^*)+\frac{2}{(mh)^{\frac{\theta}{2\theta-1}}}\right).$$

It then follows from (4.7) that conditionally given $\mathbb{D}_{\mathcal{T}_l}$, with probability at least $(1-\Sigma e^{-x})$,

$$\begin{aligned} L(\hat{\phi}_n) - L(\phi^*) \leq & L(\hat{\phi}_{l,k,d}) - L(\phi^*) + \mathrm{pen}(d) \\ & + \beta \left(L(\hat{\phi}_{l,k,d}) - L(\phi^*) + L(\hat{\phi}_n) - L(\phi^*) + \frac{2}{(mh)^{\frac{\theta}{2\theta-1}}} \right) \\ & + \frac{\beta^{-1}\left(x + x_{(\hat{k},\hat{d})}\right)}{2(mh)^{\frac{\theta}{2\theta-1}}} + \frac{1}{3}\frac{x + x_{(\hat{k},\hat{d})}}{m} - \mathrm{pen}(\hat{d}). \end{aligned}$$

By taking $x_{(k',d')} = \log(l|\mathcal{D}|)$ for every $k' \in \{1, \dots, l\}$, $d' \in \mathcal{D}$, we have that conditionally given $\mathbb{D}_{\mathcal{T}_l}$, with probability at least $(1-e^{-x})$,

$$\begin{aligned} (1-\beta)(L(\hat{\phi}_n) - L(\phi^*)) \leq & (1+\beta)(L(\hat{\phi}_{l,k,d}) - L(\phi^*)) + \mathrm{pen}(d) + \frac{2\beta}{(mh)^{\frac{\theta}{2\theta-1}}} \\ & + \frac{\beta^{-1}\left(x + \log(l|\mathcal{D}|)\right)}{2(mh)^{\frac{\theta}{2\theta-1}}} + \frac{1}{3}\frac{x + \log(l|\mathcal{D}|)}{m}. \end{aligned}$$

Integrating the inequality

$$\begin{aligned} \mathbb{P}\Bigg(\Bigg[& (1-\beta)(L(\hat{\phi}_n) - L(\phi^*)) - (1+\beta)(L(\hat{\phi}_{l,k,d}) - L(\phi^*)) - \mathrm{pen}(d) \\ & - \frac{2\beta}{(mh)^{\frac{\theta}{2\theta-1}}} - \frac{\beta^{-1}\log(l|\mathcal{D}|)}{2(mh)^{\frac{\theta}{2\theta-1}}} - \frac{1}{3}\frac{\log(l|\mathcal{D}|)}{m} \Bigg]_+ \geq \left(\frac{\beta^{-1}}{2(mh)^{\frac{\theta}{2\theta-1}}} + \frac{1}{3m} \right) x \Bigg| \mathbb{D}_{\mathcal{T}_l} \Bigg) \leq e^{-x}, \end{aligned}$$

with respect to x finally leads to

$$\begin{aligned} (1-\beta)\mathbb{E}[L(\hat{\phi}_n) - L(\phi^*)|\mathbb{D}_{\mathcal{T}_l}] \leq & (1+\beta)\left(L(\hat{\phi}_{l,k,d}) - L(\phi^*)\right) + \mathrm{pen}(d) \\ & + \frac{\beta^{-1}\left(1 + \log(l|\mathcal{D}|)\right) + 4\beta}{2(mh)^{\frac{\theta}{2\theta-1}}} + \frac{1 + \log(l|\mathcal{D}|)}{3m}. \end{aligned}$$

Since (k, d) can be arbitrarily chosen, this concludes the proof of Proposition 4.2.2. □

Lemma 4.2.1 (Birgé, Massart [14])
Let $\xi_1, \dots, \xi_n$ be independent random variables satisfying the moments condition

$$\frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} |\xi_i|^q\right] \leq \frac{q!}{2} v c^{q-2} \ \forall q \geq 2,$$

for some positive constants v and c. Then, for any positive x,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(\xi_i - \mathbb{E}[\xi_i]) \geq \sqrt{\frac{2vx}{n}} + \frac{cx}{n}\right) \leq e^{-x}.$$

□

4.3 Experimental study

In this section, we study from a practical point of view the performance of the classifier $\hat{\phi}_n$ defined by (4.2) with various penalty functions, that we list in **Table 4.1**.

penalty names	pen_0	pen_B	pen_1	pen_2	pen_3	pen_4
formula	0	$\frac{log(d)}{\sqrt{n}}$	$\frac{log(d)}{n}$	$\frac{\sqrt{d}}{n}$	$\frac{log(d)}{n^2}$	$\frac{\sqrt{d}}{n^2}$

TAB. 4.1 – List of the considered penalties.

In particular, we aim at showing that the penalty $pen_B(d) = \log(d)/m$ proposed by Biau *et al.* [12] is too heavy and that lower penalization schemes can have a significant impact on the performance of the considered procedure.

To quantify this impact, we first investigate the test error as defined below. The available data are randomly split into two parts of size $n = 3p/4$ and $t = p/4$ respectively. The first part is then split into a training set $\mathbb{D}_{\mathcal{T}_l}$ of length $l = p/4$ and a validation set $\mathbb{D}_{\mathcal{V}_m}$ of length $m = p/2$, and the classifier $\hat{\phi}_n$ is constructed as described in **Section 4.2.1**. Precisely, the training set is used to build the collection of estimators $\{\hat{\phi}_{l,k,d},\ k \in \{1,\ldots,l\},\ d \in \mathcal{D}$, whereas the validation set is used to select the parameters $\hat{k}$ and $\hat{d}$, and thus to select the estimator $\hat{\phi}_n = \hat{\phi}_{l,\hat{k},\hat{d}}$ in the collection. The responses of the second part of the data can hence be predicted thanks to this estimator, and compared to the "true" labels. Finally, the test error that we consider is the mean of the differences between the predictions and the "true" responses.

In addition to the above test error, we also evaluate the performance of the procedure in terms of dimension selection. More precisely, we study the order and stability of the selected dimension $\hat{d}$.

For our experiments, we use both realistic data coming from speech recognition and food industry problems and simple simulated data.

We summarize the results in tables containing for each penalty function, on the first row the mean of the test error over N iterations of the splitting device and computations, and on the second one the order of the selected dimension $\hat{d}$.

4.3.1 Application to realistic data

A speech recognition problem

In this section, we study data coming from the speech recognition problem considered by Biau *et al.* [12]. These data, created by Biau, consist of three sets, each containing $p = 100$ recordings of two words.

The first set deals with the words *"boat"* and *"goat"* whose labels are respectively 1 and 0. The second set corresponds to the phonemes *"sh"* (as in *she*) and *"ao"* (as in *water*) with labels 1 and 0 respectively, and the third to the words *"yes"* (label 1) and *"no"* (label 0).
These data are available at http ://www.math.univ-montp2fr/∼biau/bbwdata.tgz.

Each recording is the discretization of the corresponding signal and it contains 8192 points. **Figure 4.1** displays, for each set, two speech signals : one with label 0 and one with label 1.

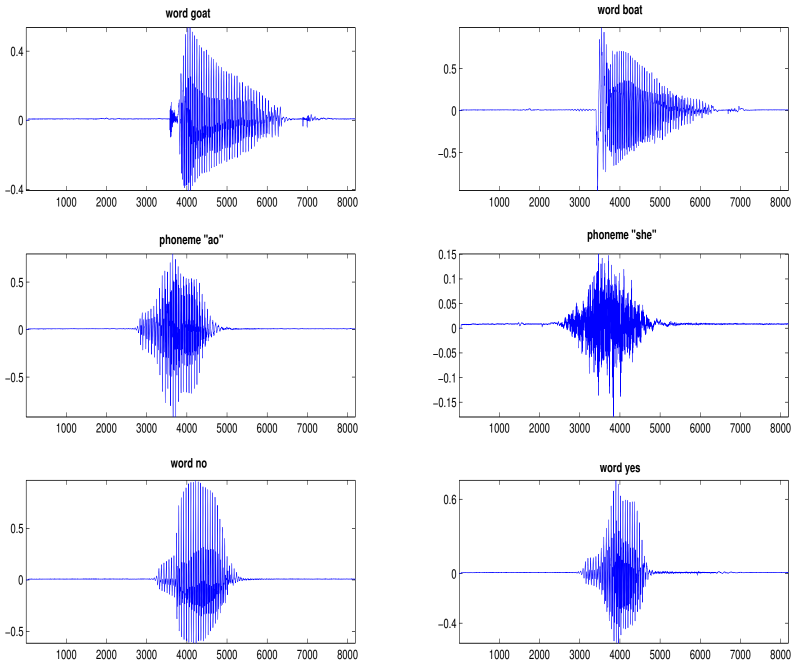

FIG. 4.1 – On this figure the speech signals associate with the words or phonemes *goat*, *boat*, *ao*, *sh*, *yes* and *no*.

Assuming that the original speech signals belong to $\mathbb{L}_2(\mathbb{R})$, we choose, for the complete system $\{\psi_j,\ j \in \mathbb{N}^*\}$ involved in the procedure, the Fourier basis called trigonometric basis too :

$$\psi_1(t) = 1, \quad \psi_{2q}(t) = \sqrt{2}cos(2\pi qt), \quad \psi_{2q+1}(t) = \sqrt{2}sin(2\pi qt), \quad q = 1, 2, \ldots.$$

The coefficients of the Fourier series expansion of each data are evaluated using a Fast Fourier Transform.

We compute the mean of the test error for the classifier $\hat{\phi}_n$ by using $N = 100$ splitting rules of the data (into two parts of size $n = l + m$, with $l = 25$ and $m = 50$, and $t = 25$ respectively).

Table 4.2 summarizes the results obtained for each of the three data sets.

words *goat boat*	pen_0	pen_B	pen_1	pen_2	pen_3, pen_4
mean test error	0.162	0.3128	0.2428	0.2526	0.2192
selected d	16(13.3)	1(0.3)	7(9.2)	5(5.3)	9(11.6)
phonemes *ao sh*	pen_0	pen_B	pen_1	pen_2	pen_3, pen_4
mean test error	0.1528	0.3948	0.1744	0.176	0.1688
selected d	5(6.1)	1(1)	5(4.5)	5(3.1)	5(5.9)
words *yes no*	pen_0	pen_B	pen_1	pen_2	pen_3, pen_4
mean test error	0.0976	0.4384	0.1372	0.134	0.1356
selected d	10(6.3)	1(0.7)	9(4.2)	9(2.3)	9(5.7)

TAB. 4.2 – For each penalty, on the first row, the mean of test error is given. On the second row, the first number is the median of the selected dimension d and the number in parentheses is the standard deviation.

In the **Table 4.2**, great values appear for the standard deviation since, according to the penalty function, great values of the dimension could be selected.

We can notice that the penalty pen_B first proposed by Biau, Bunea and Wegkamp from a theoretical point of view is not relevant since the corresponding mean of test error is the largest one observed. For example, for the third set, the mean of test error with pen_B is more than twice the one with pen_0 one. This phenomenon has already been pointed out by Biau, Bunea and Wegkamp, as they chose to study in practice the nonpenalized procedure. Furthermore, the fact that the selected dimension $\hat{d}$ with pen_B is always very small ($\hat{d} = 1$ for more than 90% of the experiments) corroborates the idea that this penalty is too heavy, and that the nonpenalized procedure will be more appropriate.

However, a refined study of the other penalization schemes shows that it can be interesting to consider procedures with some penalties of small order. For the penalties pen_1 and pen_2, we obtain larger mean test errors than the ones obtained with the nonpenalized procedure. Nevertheless, the difference between these mean test errors and the best ones is not so great, whereas one can notice that the penalized procedures with pen_1 and especially pen_2 select lower dimensions $\hat{d}$. Infact, the mean of the selected dimensions is lower as the standard deviations. This stabilization of the dimension selection process does not occur with the penalties pen_3 and pen_4, which are too small in fact to have a real impact. In view of these results, it is clear that the use of a penalized procedure, but with an appropriate penalty, can be of particular relevance.

We should here notice that the trigonometric basis may not be the most appropriate one for the decomposition of the data in the present case. Indeed, a recent work by Rouvière [10] actually proves that the above mean test errors can be improved by considering wavelet expansions. However, it appears that the instability in the selection of the dimension $\hat{d}$ by the nonpenalized procedure still holds, and that slightly penalizing the criteria could overcome this difficulty. The comparison of the different penalized procedures with such wavelet expansions, may be the subject of a future work.

We have to take into account the fact that the number of observations is small ($p = 100$), and that this could entail some difficulties in the interpretation of the results. We will hence remain reserved about our conclusions at this stage.

A food industry problem

We focus here on a classification problem which comes from the food industry. The data available on statlib (`http ://lib.stat.cmu.edu/datasets/tecator`) are recorded on a Tecator Infratec Food and Feed Analyzer working in the wavelet range 850-1050 nm by the Near Infrared Transmission principle. They contain $p = 215$ observations of the neared infrared absorbance spectrum of finely chopped pure meat, with different fat contents.

Each observation is composed of a 100 channel spectrum of absorbances and the corresponding fat content. The absorbance is $-\log_{10}$ of the transmittance measured by the spectrometer, and the fat content, measured in percent, is determined by analytic chemistry. The classification problem consists in separating meat samples with a high fat content (more than 20%) from samples with a low fat content (less than 20%). **Figure 4.2** displays some spectra for both classes.

We use here the same procedure as in the previous section, with a number of iterations N equal to 100. The splitting device is made with $n = l + m$, $l = 53$, $m = 108$, and $t = 54$. The obtained results are summarized in the **Table 4.3**.

	pen_0	pen_B	pen_1	pen_2	$pen_3,\ pen_4$
mean test error	0.2698	0.3509	0.2988	0.3016	0.2928
selected d	9(9.6)	1(0.1)	5(4)	5(3.6)	6(6.8)

TAB. 4.3 – For each penalty, on the first row, the mean of test error is given. On the second row, the first number is the median of the selected dimension d and the number in parentheses is the standard deviation.

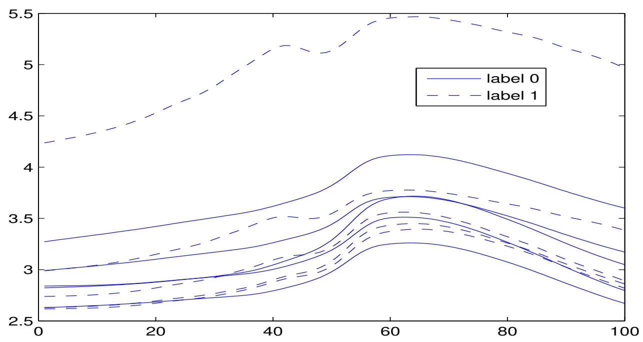

FIG. 4.2 – The dotted lines represent 5 spectra of absorbance for observations labeled 1. The other ones represent spectra of absorbance for data with label 0.

The same behavior as in the previous application to the speech recognition problem can be observed here. Hence, our conclusions remain valid, all the more since the number of observations is more important here.

The aim of the following section is to confirm these conclusions, when the number of observations is as large as we want, that is for a simple simulated problem.

4.3.2 Application to simple simulated data

In this section, we consider simulated data $(X_1, Y_1), \ldots, (X_p, Y_p)$ which are p independent copies of a pair of random variables (X, Y) with distribution defined as follows.
Let x be a regular grid of 100 points on the interval $[0; 1]$, and h be some (positive) margin parameter. The random pair (X, Y) is drawn according to the model

$$\begin{aligned} X &= cos(Ux)\mathbb{1}_{U<(b+a)/2} + sin(Ux)\mathbb{1}_{U>(b+a)/2} \\ Y &= Z_1\mathbb{1}_{U\leq\pi} + Z_2\mathbb{1}_{U>\pi}, \end{aligned}$$

where U is a random variable uniformly distributed on $[a, b]$, Z_1 and Z_2 two random variables with a Bernoulli distribution $\mathcal{B}(0.5+h)$ and $\mathcal{B}(0.5-h)$ respectively.

Figure 4.3 gives the representation of such data set.

For this simulation study, we create $p = 500$ data, and the splitting device is made with $n = l+m$, $l = 125$, $m = 250$, and $t = 125$. We apply the same procedures as in the previous sections, with a number of iterations N equal to 100. The performance of the different penalized procedures is still evaluated in terms of both mean test error, and dimension selection. The results are summarized in **Table 4.4**.

The results given in **Table 4.4** tend to confirm our previous conclusions, in the sense that they show that using a penalized procedure with a well-chosen penalty term (of order $\log(d)/n$

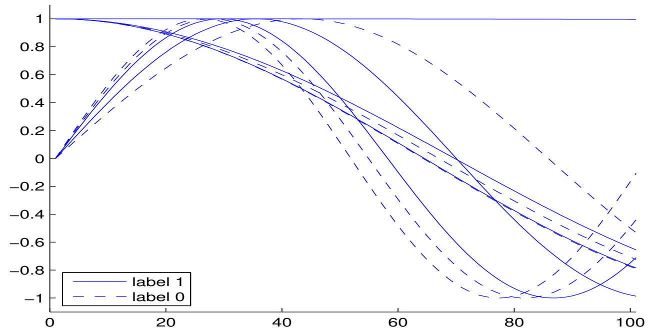

FIG. 4.3 – The dotted lines represent 5 signals X_i labeled 0. The others are functional data X_i associated with the response 1.

	pen_0	pen_B	pen_1	pen_2	pen_3, pen_4
mean test error	0.2956	0.33728	0.30296	0.30296	0.30296
selected d	3(9.5)	2(0.5)	3(1.9)	3(1.9)	3(3)

TAB. 4.4 – For each penalty, on the first row, the mean of test error is given. On the second row, the first number is the median of the selected dimension d and the number in parentheses is the standard deviation.

or $\sqrt{d}/n$) may improve the stability of the dimension selection process, whereas it does not alterate the mean test error too much. Such procedures may hence be a good compromise between the requirement of efficiency in terms of mean test error and stability of the dimension selection process.

The question of the stabilization of the dimension selection process however requires more attention. In the above experimental study, we have to take into account that the observed instability of the nonpenalized procedure may come not only from the procedure used to construct the estimator $\hat{\phi}_n$ itself, but also from the split of the original data set into two parts, the second one being used for the computation of the test error.

In the following section, we address the question of the stabilization of the procedure used to construct the estimator $\hat{\phi}_n$ only.

4.4 Stabilizing the data-splitting device

As explained in **Section 4.2**, the classification scheme proposed by Biau, Bunea, and Wegkamp requires a data-splitting device, which can lead to some instability. We aim here at confirming the intuition, emerged from the previous section, that penalizing the procedure with a well-chosen penalty may overcome this difficulty, without altering too much the performance of the procedure in terms of test error.

Since we exclusively focus on the data-splitting device involved in the procedure itself, we first split the original data into two *fixed* parts of size $n = \frac{2p}{3}$ and $t = \frac{p}{3}$ respectively. Then, on the first part, we perform, N times, the procedure described in Section 4.2.1 involving for each repetition a *random* splitting. The test error is finally computed in the same way through the second part of the data. The results obtained for Biau's data are given in **Table 4.5**.

words *goat boat*	pen_0	pen_B	pen_1	pen_2	pen_3, pen_4
mean test error	0.1684	0.272	0.18	0.168	0.1932
selected d	15(15.5)	1(0.2)	6(10.3)	5(3.5)	14(14.7)
phonemes *ao sh*	pen_0	pen_B	pen_1	pen_2	pen_3, pen_4
mean test error	0.1684	0.4088	0.2064	0.2112	0.208
selected d	7(5.4)	1(0.8)	5(3.3)	5(2.6)	6(5.3)
words *yes no*	pen_0	pen_B	pen_1	pen_2	pen_3, pen_4
mean test error	0.0604	0.378	0.1016	0.1012	0.1036
selected d	10(5.9)	1(1.5)	9(4.3)	9(2.3)	9(5.7)

TAB. 4.5 – For each penalty, on the first row, the mean of test error is given. On the second row, the first number is the median of the selected dimension d and the number in parentheses is the standard deviation.

The analysis of the results of **Table 4.5** shows that a real stabilization occurs for well-adapted penalties (pen_1 and pen_2) since the possible values for the parameter $\hat{d}$ do not belong to a large range in these cases.

The results for the other data sets (food industry and simple simulated data) are similar. For sake of shortness, we do not include them.
At present, we get an idea of the good order of the appropriate penalization, which is $\log(d)/m$ or $\sqrt{d}/m$. A refined analysis should now allow to evaluate the influence of a multiplicative factor in the penalties, and to determine an automatic procedure for the exact calibration of the penalty function to use.

Acknowledgments:

The authors thank Gérard Biau and Laurent Rouvière for valuable discussions and for making available to us data used to illustrate our method and some programs to allow fair comparisons with variants of their method.

Bibliographie

[1] C. Abraham, G. Biau, and B. Cadre. On the kernel rule for function classification. *Technical report*, (2003). University Montpellier, http ://www.math.univ-montp2.fr/∼biau/publications.html.

[2] H. Akaike. Information theory and an extension of the maximum likehood principle. In *2nd International Symposium on Information Theory*, pages 267–281. P.N. Petrov and F. Csaki, (1973).

[3] U. Amato, A. Antoniadis, and I. De Feis. Dimension reduction in functional regression with applications. *to appear in Comp. Stat. and Data. Anal.*, (2005).

[4] N. Ansaldi. *Contributions des méthodes statistiques à la quantification de l'agrément de conduite*. PhD thesis, Marne-la-Vallée, (2002).

[5] A. Antoniadis, J. Bigot, and T. Sapatinas. Wavelet estimators in nonparametric regression : a comparative simulation study. *Journal of Statistical Software*, 6(6) :1–83, (2001).

[6] J-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in estimators under margin condition. *Preprint*, (2005). Laboratoire de Probabilités et Modèles Aléatoires.

[7] A.R. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113 :301–413, (1999).

[8] A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Trans. Inf. Theory*, 37 :1034–1054, (1991).

[9] P. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Mach. Learn.*, 48((1-3)) :85–113, (2002).

[10] A. Berlinet, G. Biau, and L. Rouvière. Functional learning with wavelets. soumis à IEEE Trans. Inf. Theory, (2005).

[11] P. Besse and H. Cardot. Modélisation statistique de données fonctionnelles. In G. Govaert, editor, *Analyse de données*. Hermes, (2003).

[12] G. Biau, F. Bunea, and M. Wegkamp. Functional classification in Hilbert spaces. *IEEE Trans. Inf. Theory*, 51(6) :2163–2172, 2005.

[13] J. Bigot. *Recalage des signaux et analyse de la variance fonctionnelle par ondelettes ; application au domaine biomédical*. PhD thesis, Grenoble, (2003).

[14] L. Birgé and P. Massart. Minimum contrast estimators on sieves : exponential bounds and rates of convergence. *Bernoulli*, 4(3) :329–375, (1998).

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

[16] L. Birgé and P. Massart. Minimal penalties for gaussian model selection. à paraître dans Probability Theory and Related Fields, (2005).

[17] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. Assoc. Comput. Mach.*, 36(4) :929–965, 1989.

[18] L. Breiman. Bagging predictors. *Machine Learning*, 24(2) :123–140, (1996).

[19] L. Breiman. Random forests. *Machine Learning*, 45(1) :5–32, (2001).

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

[21] G. Celeux and J-P. Nakache. *Analyse discriminante sur variables qualitatives.* Polytechnica, (1994).

[22] F. Cerou and A. Guyader. Nearest neighbor classification in infinite dimension. Technical report, IRISA, Rennes, France, (2005).

[23] R. Coifman and N. Saito. Constructions of local orthonormal bases for classification and regression. *C. R. Acad. Sci. Paris*, 319(Ser. 1) :191–196, (1994).

[24] R. Coifman and M. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Trans. Inform. Theory*, 38(2) :713–719, (1992).

[25] T.M. Cover and P.E. Hart. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory*, 13 :21–27, (1967).

[26] S. Dabo-Niang and N. Rhomari. Nonparametric regression estimation when the regressor takes its values in a metric space. Technical report, University Paris VI, (2001). http ://www.ccr.jussieu.fr/lsta.

[27] J. Dauxois and A. Pousse. *Les analyses factorielles en calcul des probabilités et en statistique : essai d'étude synthétique.* PhD thesis, Université Toulouse III, (1976).

[28] J.C. Deville. Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'Insee*, 15 :7–97, (1974).

[29] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition.* Applications of Mathematics. Springer, New York, (1996).

[30] L. Devroye and A Krzyzak. An equivalence theorem for L^1 convergence of the kernel regression estimate. *Journal of the Statistical Planning and Inference*, 23 :71–82, (1989).

[31] L. Devroye and A Krzyzak. On the hilbert kernel density estimate. *Statistics and Probability theory letters*, 44 :299–308, (1999).

[32] L. Devroye and G. Lugosi. Lower bounds in pattern recognition and learning. *Pattern Recognition*, 28 :1011–1018, (1995).

[33] L. Devroye and T. Wagner. Nonparametric discrimination and density estimation. Technical Report 183, Electronics Research Center, University of Texas, 1976.

[34] D. Donoho and I. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3) :425–455, (1994).

[35] S. Dudoit, J. Fridlyand, and T. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457) :77–87, (2002).

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

[37] C. Favre. *Analyse en normes L^1 et L^0 des distances et des préférences. Planification en analyse sensorielle. Application au confort d'accueil de sièges automobiles.* PhD thesis, Université de Rennes II, (1999).

[38] F. Ferraty, A. Goia, and P. Vieu. Régression non-paramétrique pour des variables aléatoires fonctionnelles mélangeantes. *C. R. Acad. Sci. Paris*, 334(Ser. I) :217–220, (2002).

[39] F. Ferraty and P. Vieu. Curves discrimination : a nonparametric functional approach. *Comp. Stat. and Data Anal.*, 44(1-2) :161–173, (2003).

[40] L. Ferré and N. Villa. Discrimination de courbes par régression inverse fonctionnelle. *Revue de Statistique Appliquée*, LIII(1) :39–57, (2005).

[41] L. Ferré and A.F. Yao. Functional sliced inverse regression analysis. *Statistics*, 37(6) :475–488, (2003).

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

[43] S. Gey. *Bornes de risque, détection de ruptures, boosting : trois thèmes statistiques autour de CART en régression.* PhD thesis, Université Paris XI Orsay, (2002).

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

[45] S. Gey and J.M. Poggi. Boosting and instability for regression trees. *Comp. Stat. and Data Anal.*, 50(2) :533–550, (2006).

[46] B. Ghattas. Agrégation d'arbres de classification. *Revue de Statistique Appliquée*, XLVIII(2) :85–98, (1999).

[47] B. Ghattas. Importance des variables dans les méthodes CART. *Revue de Modulad*, 24 :29–39, (1999).

[48] B. Ghattas. *Agrégation d'arbres de décision binaires ; Application à la prévision de l'ozone dans les Bouches du Rhône.* PhD thesis, Université de la Méditerrannée, (2000).

[49] A. Gueguen and J.P. Nakache. Méthode de discrimination basée sur la construction d'un arbre de la décision binaire. *Revue de Statistique Appliquée*, XXXVI((1)) :19–38, (1998).

[50] L. Györfi. On the rate of convergence of k-nearest-neighbor classification rule. Preprint, (2005).

[51] P. Hall, D.S. Poskitt, and B. Presnell. A functional data-analytic approach to signal discrimination. *Technometrics*, 43 :1–9, (2001).

[52] W. Hardle and L. Simar. *Applied Multivariate Statistical Analysis.* Springer, (2003).

[53] T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *Annals of Statistics*, 23 :73–102, (1995).

[54] T. Hastie, S. Rosset, R. Tibshirani, and J Zhu. The entire regularization path for support vector machine. *Journal of Machine Learning Research*, 5 :1391–1415, (2004).

[55] T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(6) :607–616, 1996.

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

[57] D. Haussler, N. Littlestone, and M.K. Warmuth. Predicting $\{0, 1\}$-functions on randomly drawn points. *Inf. Comput.*, 115(2) :248–292, (1994).

[58] N.W. Hengartner, E. Matzner-Lober, and M.H. Wegkamp. Bandwidth selection for local linear regression. *Journal of the Royal Statistical Society, Series B*, 64 :1–14, (2002).

[59] G. James. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society Series B*, 64(3) :411–432, (2002).

[60] G. James and B. Silverman. Functional adaptive model estimation. *Journal of the American Statistical Association*, 100 :565–576, (2005).

[61] S.R. Kulkarni and S.E. Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Trans. Inf. Theory*, 41 :1028–1039, (1995).

[62] E. Lebarbier. *Quelques approches pour la détection de ruptures à horizon fini*. PhD thesis, Université Paris XI Orsay, (2002).

[63] S. Leurgans, R. Moyeed, and B. Silverman. Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society Series B*, 55 :725–740, (1993).

[64] G. Lugosi. Pattern classification and learning theory. In L. Györfi, editor, *Principles of Nonparametric Learning*, pages 1–56. Springer, Wien, New York, (2002).

[65] G. Lugosi and K. Zeger. Concept learning using complexity regularization. *IEEE Trans. Inf. Theory*, 42(1) :48–54, (1996).

[66] S. Mallat. *A wavelet tour of signal processing*. Academic Press, (1998).

[67] C. Mallows. Some comments on Cp. *Technometrics*, 15 :661–675, (1973).

[68] E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *Annals of Statistics*, 27(6) :1808–1829, (1999).

[69] P. Massart. Some applications of concentration inequalities to Statistics. *Annales de la faculté des Sciences de Toulouse*, 2 :245–303, (2000).

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

[71] P. Massart and E. Nédélec. Risk bounds for statistical learning. accepted to the Annals of Statistics, (2005).

[72] M. Misiti, Y. Misiti, G. Oppenheim, and J-M. Poggi. *Wavelet Toolbox For Use with Matlab*. The Mathworks Inc., (1997).

[73] M. Misiti, Y. Misiti, G. Oppenheim, and J-M. Poggi. *Les ondelettes et leurs applications*. Hermes, (2003).

[74] J-P. Nakache and J. Confais. *Statistique explicative appliquée*. Technip, (2003).

[75] J. Pagès and B. Escofier. *Analyses factorielles simples et multiples : Objectifs, méthodes et interprétation*. Dunod, (1998).

[76] 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).

[77] B.T. Polyak and A.B. Tsybakov. Asymptotic optimality of the C_p-criteria in regression projective estimation. *Theory Probab. Appl.*, 35 :293–306, (1990).

[78] J. Ramsay and B. Silverman. *Functional Data Analysis*. Springer, (1997).

[79] J. Ramsay and B. Silverman. *Applied Functional Data Analysis*. Springer, (2002).

[80] J. Rissanen. Modeling by shortest data description. *Automatica*, 14 :465–471, (1978).

[81] F. Rossi and B. Conan-Guez. Functional multi-layer perceptron : a non-linear tool for functional data analysis. *Neural networks*, 18(1) :45–60, (2005).

[82] F. Rossi and N. Villa. Classification in Hilbert spaces with support vector machines. In *Proceedings of ASMDA 2005*, pages 635–642, Brest, France, (2005).

[83] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6 :461–464, (1978).

[84] C.J. Stone. Consistent nonparametric regression. *Annals of Statistics*, 5 :595–645, (1977).

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

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

[87] M. Vannucci, P.J. Brown, and T. Fearn. A decision theoretical approach to wavelet regression on curves with a high number of regressors. *Journal of Statistical Planning and Inference*, 112 :195–212, (2003).

[88] V. Vapnik. *Estimation of dependences based on empirical data*. Springer-Verlag, New York, 1982.

[89] V. N. Vapnik and A. Ya. Chervonenkis. *Teoriya raspoznavaniya obrazov. Statisticheskie problemy obucheniya (Theory of pattern recognition. Statistical problems of learning)*. Nauka, Moscow, (1974).

[90] V.N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, (1982).

[91] V.N. Vapnik. *Statistical learning theory*. Wiley, New York, (1998).

[92] V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16 :264–280, (1971).

[93] V.N. Vapnik and A.Ya. Chervonenkis. Ordered risk minimization. *Autom. Remote Control*, 35 :1226–1235,1403–1412, (1974).

[94] B. Vidakovic. *Statistical modeling by wavelets*. Wiley, (1999).

Sélection de variables pour la discrimination en grande dimension et classification de données fonctionnelles

Résumé : Cette thèse s'inscrit dans le cadre de la statistique non paramétrique et porte sur la classification et la discrimination en grande dimension, et plus particulièrement sur la sélection de variables. Une première partie traite de la sélection de variables à travers CART, dans un cadre de régression et de classification binaire. La procédure exhaustive développée s'appuie sur le principe de la sélection de modèle qui permet d'obtenir des inégalités "oracle" et de réaliser la sélection de variables par contraste pénalisé. Une seconde partie est motivée par un problème industriel. Il s'agit de déterminer parmi les signaux temporels, mesurés au cours d'essais, ceux capables d'expliquer le ressenti de confort du conducteur, puis d'identifier les pages temporelles responsables de cette pertinence. La démarche adoptée s'articule autour du prétraitement des signaux, de la réduction de la dimension par projection dans une base d'ondelettes commune, et de la sélection de variables en mêlant CART et une stratégie pas à pas. Une dernière partie aborde le thème de la classification de données fonctionnelles au moyen des k-plus proches voisins. La procédure consiste à appliquer les k-plus proches voisins sur les coordonnées de la projection des données dans un espace fini dimensionnel. Cette procédure implique de déterminer simultanément la dimension de l'espace de projection et le nombre de voisins. La version usuelle des k-plus proches voisins et une version légèrement pénalisée sont considérées théoriquement. Un travail sur données réelles et simulées semble montrer qu'un faible terme de pénalité stabilise la sélection en conservant de bonnes performances.

Mots Clés : sélection de variables, sélection de modèle, pénalisation, ondelettes, CART, données fonctionnelles, classification, k-plus proches voisins.

Variable selection for discrimination in high dimension and functional data classification

Abstract : This thesis deals with non parametric statistics and is related to classification and discrimination in high dimension, and more particularly on variable selection. A first part is devoted to variable selection through CART, both on the regression and binary classification frameworks. The proposed exhaustive procedure is based on model selection which leads to "oracle" inequalities and allows to perform variable selection by penalized empirical contrast. A second part is motivated by an industrial problem. Il consists of determining among the temporal signals, measured during experiments, those able to explain the subjective drivability, and then to define the ranges responsible for this relevance. The adopted methodology is articulated around the preprocessing of the signals, dimensionality reduction by compression using a common wavelet basis and selection of useful variables involving CART and a strategy step by step. A last part deals with functional data classification with the k-nearest neighbors. The procedure consists of applying k-nearest neighbors on the coordinates of the projection of the data on a suitable chosen finite dimensional space. The procedure involves selecting simultaneously the space dimension and the number of neighbors. The traditional version of k-nearest neighbors and a slightly penalized version are theoretically considered. A study on real and simulated data shows that the introduction of a small penalty term stabilizes the selection while preserving good performance.

Key Words : variable selection, model selection, penalization, wavelets, CART, functional data, classification, k-nearest neighbors.

Classification AMS : 62P30, 62H30, $62-07$, 62G05, 62G08