ORSAY
N° D'ORDRE : 8515

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

Marie SAUVÉ

SÉLECTION DE MODÈLES EN RÉGRESSION NON GAUSSIENNE. APPLICATIONS À LA SÉLECTION DE VARIABLES ET AUX TESTS DE SURVIE ACCÉLÉRÉS.

Rapporteurs : Mme Fabienne COMTE
M. Gérard GRÉGOIRE

Soutenue le 11 décembre 2006 devant le jury composé de :

Mme	Fabienne	COMTE	Rapporteur
Mme	Elisabeth	GASSIAT	Présidente
M.	Gérard	GRÉGOIRE	Rapporteur
Mme	Sylvie	HUET	Examinatrice
M.	Pascal	MASSART	Directeur de Thèse
M.	Jean-Michel	POGGI	Examinateur

Remerciements

Un grand merci à Pascal avec qui j'ai découvert les statistiques en maîtrise, puis la sélection de modèles en DEA et en thèse. Chacune de nos discussions a été riche d'enseignements et de plaisir. Merci de m'avoir guidée et motivée, d'avoir été à la fois présent et discret.

Je remercie Fabienne Comte, Elisabeth Gassiat, Gérard Grégoire, Sylvie Huet et Jean-Michel Poggi qui me font l'honneur et le plaisir de participer au jury de cette thèse. Un merci particulier à Fabienne Comte et Gérard Grégoire pour avoir rapporté mon travail.

Je tiens à remercier sincèrement Bernard Besançon, mon professeur de math spé, qui m'a fait découvrir et aimer les mathématiques. Ses cours passionants m'ont donné les moyens et l'envie de poursuivre.

Je remercie aussi tous mes professeurs d'Orsay et en particulier Alano Ancona, Lucile Bégueri, Raphaël Cerf, Jacques Chaumat, Jean Coursol, Béatrice Laurent et Wendelin Werner pour leurs cours à la fois bien construits et ouverts à la réflexion.

Merci à toute l'équipe de Probabilités et Statistiques pour leur sympathie, leur disponibilité et leurs coups de pouce en informatique. Merci en particulier à Didier Dacunha-Castelle, Elisabeth Gassiat et Marc Lavielle pour leur soutien.

Merci aux doctorants ou ex-doctorants: Aurélien, Besma, Christine, Gilles, Ismaël B et C, Laurent, Magalie, Marc, Marion, Mina, Nicolas, Sébastien, Sophie et tous les autres pour leur présence et leur amitié. Merci surtout à Christine pour nos discussions mathématiques et autres, et pour tous ses petits conseils pratiques. Merci aussi à Marc de m'avoir initiée à la fiabilité.

Un grand merci également à Thomas Lafforgue et Luc Abergel pour leur aide, leur soutien et leurs précieux conseils.

Mille mercis à Françoise pour sa présence, son écoute, et sa bonne humeur communicative. Ce fut une chance pour moi de l'avoir eu à mes côtés durant ces années.

A mes parents, mon frère, mes grands-parents et mes beau-parents pour leur soutien et leur fierté. A Guillaume pour tellement.

Résumé

Cette thèse traite de la sélection de modèles en régression non gaussienne. Notre but est d'obtenir des informations sur une fonction $s : \mathcal{X} \to \mathbb{R}$ dont on ne connaît qu'un certain nombre de valeurs perturbées par des bruits non nécessairement gaussiens. Nous adoptons l'approche non asymptotique de la sélection de modèles par minimisation d'un critère des moindres carrés pénalisés. Nous considérons une collection de modèles $(S_m)_{m\in\mathfrak{M}_n}$. Chaque S_m est une classe de fonctions définies sur $\mathcal{X}$ et à valeurs dans $\mathbb{R}$ à laquelle nous associons l'estimateur $\hat{s}_m$ des moindres carrés sur S_m. Nous déterminons des critères pénalisés qui permettent de sélectionner un modèle $S_{\hat{m}}$ approximativement optimal au sens où le risque de l'estimateur pénalisé $\hat{s}_{\hat{m}}$ est comparable à l'infimum des risques des $(\hat{s}_m)_{m\in\mathfrak{M}_n}$. Pour chaque pénalité proposée, nous donnons une borne de risque non asymptotique. Nous utilisons la sélection de modèles pour estimer s sans faire d'hypothèses a priori et nous envisageons la détection de ruptures et la sélection de variables comme des cas particuliers de sélection de modèles.

Dans un premier temps, nous considérons des modèles de fonctions constantes par morceaux associés à une collection de partitions de $\mathcal{X}$. Nous déterminons un critère des moindres carrés pénalisés qui permet de sélectionner une partition dont l'estimateur associé (de type regressogramme) vérifie une inégalité de type oracle. La sélection d'un modèle de fonctions constantes par morceaux ne conduit pas en général à une bonne estimation de s, mais permet notamment de détecter les ruptures de s. Nous proposons aussi une méthode non linéaire de sélection de variables qui repose sur l'application de plusieurs procédures CART et sur la sélection d'un modèle de fonctions constantes par morceaux. CART permet d'associer à chaque paquet de variables une série de modèles de fonctions constantes par morceaux construits sur les variables du paquet considéré. Puis nous déterminons un critère pénalisé permettant de sélectionner un modèle et le paquet de variables sous-jacent.

Dans un deuxième temps, nous considérons des modèles de fonctions polynomiales par morceaux, dont les qualités d'approximation sont meilleures. L'objectif est d'estimer s par un polynôme par morceaux dont le degré peut varier d'un morceau à l'autre. Nous déterminons un critère pénalisé qui sélectionne une partition de $\mathcal{X} = [0,1]^p$ et une série de degrés dont l'estimateur polynomial par morceaux associé vérifie une inégalité de type oracle. Nous appliquons aussi ce résultat pour détecter les ruptures d'un signal affine par morceaux. Ce dernier travail est motivé par la détermination d'un intervalle de stress convenable pour les tests accélérés visant à obtenir dans des délais raisonnables des informations sur le temps de survie de certains composants.

Contents

Introduction

Notre travail se situe dans le cadre statistique d'un modèle de régression multivariée. Nous étudions le comportement d'une variable réelle Y (appelée variable réponse) en fonction d'une ou plusieurs variables explicatives $x^1, \ldots, x^p$. Nous notons $\mathbf{x}$ le p-uplet $(x^1, \ldots, x^p)$ et $\mathcal{X}$ l'ensemble des valeurs possibles pour $\mathbf{x}$ ($\mathcal{X} = \mathbb{R}^p$ en général). Nous supposons que

$$Y = s(x^1, \ldots, x^p) + \varepsilon \tag{1}$$

où $s : \mathcal{X} \longrightarrow \mathbb{R}$ est une fonction inconnue qui donne la relation entre $\mathbf{x} = (x^1, \ldots, x^p)$ et Y, et où ε est une perturbation aléatoire d'espérance nulle (conditionnellement à $\mathbf{x}$). ε correspond soit à des erreurs de mesure, soit à la dépendance de Y vis à vis de quantités autres que $(x^1, \ldots, x^p)$ qui ne sont ni contrôlées ni observées. Notre but est d'obtenir des informations sur s à partir d'un échantillon de n observations $(x_i^1, \ldots, x_i^p, Y_i)_{1 \leq i \leq n}$ obéissant à la relation (1). Nous utilisons les données $(x_i^1, \ldots, x_i^p, Y_i)_{1 \leq i \leq n}$ pour résoudre trois problèmes classiques en statistique:

- l'estimation de s, qui consiste à approcher la fonction inconnue s par une fonction mesurable des observations $(x_i^1, \ldots, x_i^p, Y_i)_{1 \leq i \leq n}$ (une telle fonction est appelée un estimateur et est notée $\hat{s}$),

- la sélection de variables, qui consiste à déterminer un petit nombre de variables parmi $x^1, \ldots, x^p$ permettant à elles seules de bien expliquer ou prédire la réponse Y,

- la détection de ruptures, qui consiste à localiser des ruptures dans le comportement de la fonction s.

Dans la suite, nous appelons modèle tout espace vectoriel de fonctions définies sur $\mathcal{X}$ et à valeurs dans $\mathbb{R}$, et nous abordons chacun des trois problèmes cités ci-dessus sous l'angle de la sélection de modèles. Nous adoptons l'approche non asymptotique de la sélection de modèles par pénalisation développée par Birgé et Massart [3, 4] dans un cadre gaussien très général. Leurs résultats s'appliquent en particulier lorsque que l'on observe n couples $(\mathbf{x}_i, Y_i)_{1 \leq i \leq n}$ vérifiant (1) avec des $\mathbf{x}_i$ déterministes et des perturbations aléatoires ε_i indépendantes, centrées et de même loi gaussienne. Dans notre étude, nous ne supposerons pas que les erreurs sont gaussiennes. Nous supposerons seulement qu'elles ont des moments exponentiels au voisinage de 0.

Dans le premier chapitre, nous considérons des modèles de fonctions constantes par morceaux. Nous déterminons un critère des moindres carrés pénalisés qui permet de sélectionner un modèle de fonctions constantes par morceaux dont l'estimateur associé (de type regressogramme)

vérifie une inégalité de type oracle. Les modèles de fonctions constantes par morceaux ont l'avantage d'être simples, mais ils sont parfois trop frustes pour permettre une bonne estimation de la fonction s. Ils permettent en revanche de détecter des ruptures dans la moyenne d'un signal. Le travail d'Emilie Lebarbier [16] sur la détection de ruptures dans la moyenne d'un signal gaussien s'appuie sur le résultat de sélection de modèles de Birgé et Massart [3], qu'elle applique à une collection de modèles de fonctions constantes par morceaux. Pour une telle collection de modèles, nous avons pu relaxer l'hypothèse gaussienne. Il est donc possible d'étendre le travail d'Emilie Lebarbier à un cadre non gaussien. Le résultat de ce chapitre permet aussi de justifier l'étape d'élagage de l'algorithme CART [7] dans un cadre de régression non gaussienne.

Dans le deuxième chapitre, nous proposons une méthode non linéaire de sélection de variables qui repose sur l'application de plusieurs procédures CART. Nous justifions notre méthode dans un cadre de régression non gaussienne par le biais d'inégalités de type oracle. Le résultat du premier chapitre permet de démontrer ce nouveau résultat. Notre méthode de sélection de variables s'applique aussi dans le cadre de la classification binaire, mais dans ce cas la démonstration repose sur une inégalité de concentration due à Talagrand.

Le troisième chapitre traite de l'estimation d'une fonction de régression s par un polynôme par morceaux dont le degré peut varier d'un morceau à l'autre. Nous déterminons un critère pénalisé qui sélectionne une partition de $\mathcal{X} = [0,1]^p$ et une série de degrés dont l'estimateur polynomial par morceaux associé vérifie une inégalité de type oracle. Ce résultat généralise celui du premier chapitre. Il permet de valider la procédure d'estimation de Comte et Rozenholc [10] dans un cadre un peu plus général (non nécessairement sous-gaussien) et de mieux comprendre l'algorithme MARS de Friedman [12]. Nous proposons aussi une extension de l'algorithme CART pour construire un estimateur polynomial par morceaux.

Dans le quatrième chapitre, nous appliquons le résultat sur la sélection d'un modèle polynomial par morceaux du chapitre 3 pour détecter les ruptures d'un signal affine par morceaux (ou approché par une fonction affine par morceaux). Nous proposons deux méthodes pour calibrer les constantes de la pénalité donnée par le théorème du chapitre 3. Ce travail est motivé par la détermination d'un intervalle de stress convenable pour les tests accélérés visant à obtenir dans des délais raisonnables des informations sur le temps de survie de certains composants.

Avant de décrire plus précisément les travaux mentionnés ci-dessus, nous présentons la méthode d'estimation par minimisation du contraste des moindres carrés et les principes de la sélection de modèles.

Prédiction et estimation

Dans le cadre de la régression définie par (1), nous appelons prédicteur toute fonction mesurable $u : \mathcal{X} \longrightarrow \mathbb{R}$. Soit $(\mathbf{x}_i, Y_i)_{1 \leq i \leq n}$ n observations vérifiant (1). s est le meilleur prédicteur au

sens où il minimise $\mathbb{E}[\gamma_n(u)]$, γ_n étant le contraste des moindres carrés:

$$\gamma_n(u) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - u(\mathbf{x}_i))^2.$$

La qualité d'un prédicteur u est alors mesurée par sa perte relative:

$$\begin{aligned} l(s,u) &= \mathbb{E}[\gamma_n(u)] - \mathbb{E}[\gamma_n(s)] \\ &= \begin{cases} \|s-u\|_n^2 & \text{si les } \mathbf{x}_i \text{ sont déterministes} \\ \|s-u\|_\mu^2 & \text{si les } \mathbf{x}_i \text{ sont des variables aléatoires indépendantes et de même loi } \mu \end{cases} \end{aligned}$$

où $\|.\|_\mu$ est la norme de $L^2(\mu)$, $\|.\|_n$ est la norme euclidienne de $\mathbb{R}^n$ divisée par $\sqrt{n}$, et où l'on note de la même façon une fonction u et le vecteur associé $(u(x_i))_{1\leq i\leq n} \in \mathbb{R}^n$.

Remarque 0.1 *Lorsque les* $\mathbf{x}_i$ *seront considérés aléatoires, nous les noterons* X_i *au lieu de* $\mathbf{x}_i$.

s étant inconnue, on veut construire un estimateur $\hat{s}$ à partir des données qui soit le plus proche possible de s au sens où son risque $\mathbb{E}[l(s,\hat{s})]$ est le plus petit possible. Une méthode classique pour estimer s consiste à minimiser le contraste γ_n sur un modèle S. L'estimateur ainsi obtenu est noté $\hat{s}_S$ et est appelé l'estimateur des moindres carrés associé à S. Lorsque les points $\mathbf{x}_i$ sont déterministes, le risque de $\hat{s}_S$ s'écrit:

$$\mathbb{E}\left(\|s-\hat{s}_S\|_n^2\right) = \inf_{u\in S}\|s-u\|_n^2 + \frac{\tau^2}{n}\dim_{\mathbb{R}^n}(S) \quad \text{où } \tau^2 = \mathbb{E}(\varepsilon_i^2). \tag{2}$$

Le premier terme, appelé le terme de biais, représente l'erreur d'approximation du modèle S. Le deuxième terme, appelé le terme de variance, représente l'erreur d'estimation dans le modèle S. Plus le modèle S est gros, plus on améliore les qualités d'approximation et donc plus le terme de biais est petit, mais plus on commet d'erreurs d'estimation et donc plus le terme de variance est grand. Inversement, plus le modèle S est petit, plus le terme de biais est grand, mais plus le terme de variance est petit. Pour obtenir un bon estimateur de s, il faut déterminer un modèle S qui fait un bon compromis entre le biais et la variance. Ce dernier point est l'objectif de la sélection de modèles.

Sélection de modèles

Nous décrivons ici l'approche non asymptotique de la sélection de modèles par pénalisation développée par Birgé et Massart [3, 4].

Soit $(S_m)_{m\in\mathcal{M}_n}$ une collection de modèles dont le nombre peut dépendre de la taille n de l'échantillon des observations. Nous considérons la collection $(\hat{s}_m)_{m\in\mathcal{M}_n}$ des estimateurs des moindres carrés associés aux modèles $(S_m)_{m\in\mathcal{M}_n}$. Le modèle idéal m^* est celui dont l'estimateur associé $\hat{s}_{m^*}$ a le plus petit risque:

$$m^* = \arg\min_{m\in\mathcal{M}_n}\mathbb{E}\left(\|s-\hat{s}_m\|_n^2\right).$$

Comme m^* dépend de s, $\hat{s}_{m^*}$ ne peut pas être utilisé comme estimateur de s. Le but de la sélection de modèles est de construire un modèle $\hat{m}$ à partir des données tel que le risque de l'estimateur associé $\hat{s}_{\hat{m}}$ soit le plus proche possible de l'oracle: $\mathbb{E}\left(\|s-\hat{s}_{m^*}\|_n^2\right) = \inf_{m\in\mathcal{M}_n} \mathbb{E}\left(\|s-\hat{s}_m\|_n^2\right)$. Vu l'expression (2) du risque, $\hat{m}$ doit pour cela faire un bon compromis entre le biais et la variance. L'idée consiste à sélectionner un modèle $\hat{m}$ en minimisant un critère des moindres carrés pénalisés:

$$\text{crit}(m) = \gamma_n(\hat{s}_m) + \text{pen}(m) \tag{3}$$

où le terme $\text{pen}(m)$ pénalise les gros modèles S_m. L'estimateur $\hat{s}_{\hat{m}}$ associé au modèle $\hat{m}$ ainsi choisi est appelé l'estimateur des moindres carrés pénalisés. Le but de l'approche non asymptotique est de déterminer une pénalité $\text{pen}(m)$ telle que l'estimateur des moindres carrés pénalisés $\hat{s}_{\hat{m}}$ vérifie:

$$\mathbb{E}\left(\|s-\hat{s}_{\hat{m}}\|_n^2\right) \quad \leq \quad C \inf_{m\in\mathcal{M}_n} \mathbb{E}\left(\|s-\hat{s}_m\|_n^2\right) \tag{4}$$

où C est une constante supérieure à 1 et le plus proche possible de 1. Une telle inégalité est appelée inégalité oracle.

Le premier critère pénalisé de type (3) est du à Mallows [17]. Il est issu de l'heuristique décrite ci-après. Notons $s_m = \arg\min_{u\in S_m} \|s-u\|_n^2$ et $D_m = \dim_{\mathbb{R}^n}(S_m)$. D'après la décomposition du risque en biais-variance (2) et d'après Pythagore, m^* minimise

$$m \longrightarrow -\|s_m\|_n^2 + \frac{\tau^2}{n} D_m \tag{5}$$

L'heuristique de Mallows consiste à dire qu'en remplaçant $\|s_m\|_n^2$ dans (5) par un estimateur sans biais, on obtient un minimiseur $\hat{m}$ dont les performances sont proches de l'oracle. Comme $\mathbb{E}\left(\|\hat{s}_m\|_n^2\right) = \|s_m\|_n^2 + \frac{\tau^2}{n}D_m$, $\|\hat{s}_m\|_n^2 - \frac{\tau^2}{n}D_m$ est un estimateur sans biais de $\|s_m\|_n^2$, et on obtient $-\|\hat{s}_m\|_n^2 + 2\frac{\tau^2}{n}D_m = -\|Y\|_n^2 + \gamma_n(\hat{s}_m) + 2\frac{\tau^2}{n}D_m$. Le critère C_p de Mallows [17] s'écrit:

$$C_p(m) = \gamma_n(\hat{s}_m) + 2\frac{\tau^2}{n} D_m.$$

Ce critère est un critère pénalisé de type (3) avec $\text{pen}(m) = 2\frac{\tau^2}{n}D_m$. Lorsque la variance $\tau^2 = \mathbb{E}(\varepsilon_i^2)$ est inconnue, on peut la remplacer par un estimateur.
Le critère C_p de Mallows ne donne de bons résultats que si le nombre de modèles de dimension donnée n'est pas trop grand. Pour que l'heuristique de Mallows fonctionne, il faudrait que $\|\hat{s}_m\|_n^2$ soit du même ordre de grandeur que son espérance pour tous les modèles $m \in \mathcal{M}_n$ simultanément. Pour obtenir une pénalité qui sélectionne un $\hat{m}$ proche de m^* en terme de risque, on ne peut pas se contenter de remplacer $\|s_m\|_n^2$ par $\|\hat{s}_m\|_n^2 - \frac{\tau^2}{n}D_m$. Il faut étudier les déviations de $\|\hat{s}_m\|_n^2 - \|s_m\|_n^2$ autour de son espérance ($\frac{\tau^2}{n}D_m$), et choisir une pénalité qui les compense. Les outils essentiels dans la détermination d'une bonne pénalité sont les inégalités de concentration.

Résultats de Birgé et Massart dans un cadre gaussien:

Dans un cadre gaussien, grâce à une inégalité de concentration pour la somme d'un χ^2 et d'une gaussienne [4, Appendix,lemma 1], Birgé et Massart ont obtenu une inégalité de type oracle pour une pénalité de la forme:

$$\text{pen}(m) = K\frac{\tau^2}{n}\left(D_m + \mathsf{a}\sqrt{D_m x_m} + \mathsf{b}x_m\right) \tag{6}$$

où $K > 1$, $\mathsf{a} > 2$ et $\mathsf{b} > 2$ sont trois constantes, et où $(x_m)_{m\in\mathcal{M}_n}$ est une famille de poids vérifiant:

$$\sum_{m\in\mathcal{M}_n} e^{-x_m} \leq \Sigma, \qquad \Sigma \in \mathbb{R}_+^*. \tag{7}$$

La pénalité dépend de la complexité de la collection de modèles $(S_m)_{m\in\mathcal{M}_n}$ via les poids $(x_m)_{m\in\mathcal{M}_n}$.
Lorsque le nombre de modèles de dimension D donnée n'est pas trop gros, plus précisément lorsque $|\{m \in \mathcal{M}_n;\ D_m = D\}| \leq \Gamma D^r$ avec $\Gamma \in \mathbb{R}_+^*$ et $r \in \mathbb{N}$, alors les poids $x_m = LD_m$ vérifient (7) pour tout $L > 0$, et donc $\text{pen}(m) = K'\frac{\tau^2}{n}D_m$ avec $K' > 1$ convient. Le critère C_p de Mallows (qui correspond à $K' = 2$) est alors validé.

Lorsque le nombre de modèles de dimension D est beaucoup plus gros, de l'ordre de $\binom{N}{D}$ avec N grand, alors il faut prendre des poids plus gros et donc une pénalité plus forte. Dans cette situation, Birgé et Massart [4] ont montré que le critère de Mallows peut donner de très mauvais résultats.

Résultats dans un cadre non gaussien:

Dans un cadre de régression non gaussienne, la difficulté pour déterminer une bonne pénalité et obtenir un résultat du type [4, theorem 1] de Birgé et Massart est de contrôler les déviations des statistiques $\chi_m^2 = \|\varepsilon_m\|_n^2$ où $\varepsilon_m = \arg\min_{u\in S_m} \|\varepsilon - u\|_n^2$. Si $\mathbb{E}(|\varepsilon_i|^p) < +\infty$ pour un $p \geq 2$, Baraud [1, Corollary 5.1] a montré que pour tout $x > 0$

$$\mathbb{P}\left(\chi_m^2 \geq \frac{\tau^2}{n}D_m + 2\frac{\tau^2}{n}\sqrt{D_m x} + \frac{\tau^2}{n}x\right) \leq C(p)\frac{\mathbb{E}(|\varepsilon_i|^p)}{\tau^p}D_m x^{-p/2}. \tag{8}$$

Lorsque $|\{m \in \mathcal{M}_n;\ D_m = D\}| \leq \Gamma D^r$, en supposant seulement que les perturbations ε_i ont des moments d'ordre $p > 2r + 6$, Baraud [1] en a déduit qu'une pénalité de la forme $\text{pen}(m) = K'\frac{\tau^2}{n}D_m$ avec $K' > 1$ permet encore d'obtenir une inégalité de type oracle.
Nous aimerions déterminer, au delà du résultat de Baraud, une pénalité générale (telle que la pénalité (6)) qui permet d'obtenir une inégalité de type oracle quelle que soit la complexité de la collection de modèles. Pour obtenir son résultat, Baraud a du supposer que les ε_i ont des moments d'ordre $p > 2r + 6$ où r est le degré du polynôme majorant la complexité de la collection de modèles. La valeur minimale admissible pour p croît avec le degré r de la complexité. Pour traiter des collections de modèles de complexité exponentielle, nous supposerons que les ε_i ont des moments exponentiels au voisinage de 0. Cela revient à supposer qu'il existe deux constantes $b \geq 0$ et $\sigma > 0$ telles que:

$$\text{pour tout } \lambda \in (-1/b, 1/b) \qquad \log\mathbb{E}\left(e^{\lambda\varepsilon_i}\right) \leq \frac{\sigma^2\lambda^2}{2(1 - b|\lambda|)}. \tag{9}$$

σ^2 est nécessairement supérieur à $\tau^2 = \mathbb{E}(\varepsilon_i^2)$, mais il peut être choisi aussi proche de τ^2 que l'on veut à condition de prendre un plus grand b.

Nous venons de voir que la sélection de modèles permet de déterminer un estimateur dont le risque est petit relativement aux risques des estimateurs associés aux modèles d'une collection donnée. Elle permet aussi de répondre à des problèmes plus spécifiques comme la sélection de variables ou la détection de ruptures.

Sélection de variables

Etant donnée une liste de variables $x^1, \ldots, x^p$, le but de la sélection de variables est de déterminer un petit nombre de variables qui suffisent à elles seules à bien expliquer ou prédire la réponse.

Dans le cadre particulier de la régression linéaire, $s(\mathbf{x}) = \sum_{j=1}^p \alpha_j x^j$ et donc (1) devient:

$$Y = \sum_{j=1}^{p} \alpha_j x^j + \varepsilon.$$

L'estimateur des moindres carrés $\hat{s}(\mathbf{x}) = \sum_{j=1}^p \hat{\alpha}_j x^j$ avec $\hat{\alpha} = \arg\min_{\alpha} \frac{1}{n} \sum_{i=1}^n \left(Y_i - \sum_{j=1}^p \alpha_j x_i^j \right)^2$ est non biaisé mais, lorsque le nombre p de variables est grand, sa variance est grande. Le risque de $\hat{s}$ peut être amélioré en diminuant ou en éliminant certains coefficients $\hat{\alpha}_j$, de façon à réduire le terme de variance quitte à augmenter un peu le terme de biais. En éliminant certains coefficients (cad en éliminant certaines variables), on gagne aussi en interprétabilité.

Les méthodes Ridge Regression et Lasso (voir [15]) sont des versions pénalisées de la méthode des moindres carrés. Pour Ridge Regression, la pénalité est proportionnelle à la norme l^2 de $\alpha = (\alpha_1, \ldots, \alpha_p)$. Cette méthode donne des coefficients $\hat{\alpha}_j$ plus petits et un meilleur estimateur $\hat{s}$ au sens du risque. Pour Lasso, la pénalité est proportionnelle à la norme l^1 de $\alpha = (\alpha_1, \ldots, \alpha_p)$. Lasso réduit certains coefficients et en élimine certains autres. Cette méthode améliore aussi le risque et permet en plus d'obtenir un estimateur plus simple, qui fait intervenir un plus petit nombre de variables. Avec Lasso, on gagne à la fois en terme de risque et en interprétabilité. Malheureusement son coût algorithmique est très élevé.

Un autre moyen de construire un estimateur de risque petit et qui soit facile à interpréter consiste à déterminer pour chaque $K \in \{0, 1, \ldots, p\}$ le paquet de K variables qui minimise le critère des moindres carrés. L'algorithme de Furnival et Wilson [13] permet de faire cette recherche pour $p \leq 40$. Il faut ensuite choisir K en estimant par exemple les risques par validation croisée (ou grâce à un échantillon test).

Nous pouvons aussi adopter une approche de sélection de modèles, en associant à chaque paquet M de variables (i.e. M sous-ensemble de $\{x^1, \ldots, x^p\}$) le modèle S_M des fonctions linéaires en les variables de M ainsi que l'estimateur $\hat{s}_M$ des moindres carrés de s sur S_M.

Nous sélectionnons un paquet $\hat{M}$ en minimisant un critère pénalisé:

$$\text{crit}(M) = \gamma_n(\hat{s}_M) + \text{pen}(M)$$

avec une pénalité pen(M) qui pénalise les gros paquets de variables, et qui doit être choisie de façon à ce que l'estimateur $\hat{s}_{\hat{M}}$ vérifie une inégalité de type oracle. En général, la pénalité pen(M) ne dépend du paquet M que via le nombre de variables qu'il contient: $K_M = |M|$. Nous choisissons donc pour chaque $K \in \{0, 1, \ldots, p\}$ le paquet de K variables $\hat{M}_K$ qui minimise le critère des moindres carrés. La pénalité permet de choisir le nombre $\hat{K}$ tel que le modèle associé fait un bon compromis entre le biais et la variance.

En associant à chaque paquet de variables M, le modèle S_M des fonctions linéaires en les variables de M, nous nous somme ramenés au problème de sélection d'un modèle, et nous pouvons donc appliquer les résultats précédents. Si l'on considère uniquement les paquets de variables de la forme $M = \{x^1, \ldots, x^j\}$, avec $1 \leq j \leq p$, alors une pénalité de la forme $\text{pen}(M) = K' \frac{\tau^2}{n} |M|$ convient comme le démontre le résultat de Birgé et Massart [3, 4] dans un cadre gaussien et comme le confirme le résultat de Baraud [1] dans un cadre non gaussien. Si l'on considère tous les paquets possibles, alors $|\{M;\ |M| = K\}| = \binom{p}{K}$ et Birgé et Massart [4] montre que la pénalité précédente n'est plus suffisante. Dans ce cas, nous disposons d'une pénalité validée uniquement dans le cadre gaussien.

Les méthodes de sélection de variables discutées ci-dessus sont des méthodes linéaires de sélection de variables dans le sens où elles considèrent des interactions linéaires entre les variables explicatives et la réponse. Dans le chapitre 2, nous proposons une méthode non linéaire de sélection de variables. Nous adoptons une approche sélection de modèles, mais au lieu d'associer à chaque paquet M de variables le modèle S_M des fonctions linéaires en les variables de M, nous associons à M des modèles S_M de fonctions constantes par morceaux construits sur les variables de M.

Les modèles de fonctions constantes par morceaux sont des modèles simples notamment utilisés en détection de ruptures.

détection de ruptures

La détection de ruptures peut elle aussi être abordée par une approche de sélection de modèles (voir le chapitre 2 de la thèse de Lebarbier [16]).

Supposons que s est constante par morceaux ou qu'elle peut être convenablement approchée par une fonction constante par morceaux:

$$s = \sum_{J \in M} \alpha_J \mathbb{1}_J$$

avec $M = (J_1, \ldots, J_D)$ une partition de $\mathcal{X}$. L'objectif est d'estimer cette partition M. Pour cela, nous commençons par définir une collection $\mathcal{M}_n$ de partitions de l'espace $\mathcal{X}$. Lorsque $\mathcal{X} = [0, 1]$ et $x_i = \frac{i}{n}$ pour tout $1 \leq i \leq n$, la collection $\mathcal{M}_n$ naturelle est la collection de

toutes les partitions dont les noeuds appartiennent à la grille $(\frac{i}{n})_{1\leq i\leq n}$. Puis, nous associons à chaque partition $M \in \mathcal{M}_n$, le modèle S_M des fonctions constantes par morceaux construites sur la partition M, et $\hat{s}_M$ l'estimateur des moindres carrés de s sur S_M. Nous sélectionnons une partition $\hat{M}$ en minimisant un critère pénalisé:

$$\text{crit}(M) = \gamma_n(\hat{s}_M) + \text{pen}(M)$$

avec une pénalité $\text{pen}(M)$ qui pénalise les partitions ayant un grand nombre de morceaux, et qui doit être choisie de façon à ce que l'estimateur $\hat{s}_{\hat{M}}$ vérifie une inégalité de type oracle.

Cette méthode vise à déterminer la partition dont l'estimateur des moindres carrés associé approche le mieux possible le vrai signal. Elle ne déterminera pas forcément le vrai nombre de ruptures, car, lorsque certaines ruptures sont peu marquées, l'oracle ne correspond pas à la partition tenant compte de toutes les ruptures.

Dans le cas où $\mathcal{X} = [0,1]$ et où $\mathcal{M}_n$ est la collection de toutes les partitions dont les noeuds appartiennent à la grille $(x_i = \frac{i}{n})_{1\leq i\leq n}$, en appliquant le résultat gaussien de sélection de modèles de Birgé et Massart, Lebarbier a obtenu la pénalité suivante:

$$\text{pen}(M) = \tau^2 \frac{|M|}{n}\left(\alpha \log \frac{n}{|M|} + \beta\right)$$

où $\tau^2 = \mathbb{E}(\varepsilon_i^2)$, α et β sont deux constantes absolues, et $|M|$ est le nombre de morceaux de la partition M. Nous verrons dans le chapitre 1 qu'une pénalité de forme similaire peut être utilisée pour détecter les ruptures dans la moyenne d'un signal non nécessairement gaussien.

Lorsqu'il s'agit de détecter les sauts ou les changements de pente d'un signal affine par morceaux (ou approché par une fonction affine par morceaux), nous pouvons procéder de la même manière en associant à chaque partition $M \in \mathcal{M}_n$, le modèle S_M des fonctions affines par morceaux construites sur la partition M. Cette méthode est appliquée ici dans le chapitre 4.

Dans les 4 paragraphes qui suivent, je présente mes travaux.

Sélection d'un modèle de type histogramme en régression non gaussienne

Dans le premier chapitre, nous considérons des observations $(\mathbf{x}_i, Y_i)_{1\leq i\leq n}$ vérifiant (1) avec $\mathbf{x}_i \in \mathcal{X}$ déterministes et ε_i des perturbations aléatoires centrées, indépendantes et de même loi ayant des moments exponentiels au voisinage de 0. Etant donnée une collection $\mathcal{M}_n$ de partitions de $\mathcal{X}$, nous associons à chaque partition $M \in \mathcal{M}_n$ le modèle S_M des fonctions constantes par morceaux construites sur la partition M et l'estimateur $\hat{s}_M$ obtenu par minimisation du contraste des moindres carrés sur S_M. Les estimateurs $(\hat{s}_M)_{M\in\mathcal{M}_n}$ sont des fonctions constantes par morceaux appelées regressogrammes. Nous déterminons un critère des moindres carrés pénalisés qui permet de sélectionner une partition $\hat{M}$ dont le regressogramme associé $\hat{s}_{\hat{M}}$ vérifie une inégalité de type oracle. Lorsque $\mathcal{X} = [0,1]$ et que les x_i prennent N valeurs distinctes $0 = v_0 < v_1 < \ldots < v_{N-1} < 1$, la collection $\mathcal{M}_n$ naturelle est la collection de toutes les partitions dont les noeuds appartiennent à la grille de points $(v_1, \ldots, v_{N-1})$. Le

nombre de modèles de dimension D donnée est alors $\binom{N-1}{D-1}$, et le résultat de Baraud [1] ne s'applique donc pas. Notre résultat donne une forme générale de pénalité permettant de traiter à la fois des collections de complexité polynomiale et des collections plus complexes telles que celle citée ci-dessus. Pour l'obtenir, nous avons construit une nouvelle inégalité de concentration pour les statistiques $\chi^2_M = \|\varepsilon_M\|^2_n$ où $\varepsilon_M = \arg\min_{u\in S_M} \|\varepsilon - u\|^2_n$. Pour des modèles S_M de fonctions constantes par morceaux, nous avons pu construire à la main des inégalités de concentration en utilisant seulement l'inégalité de Bernstein [19, section 2.2.3]. En nous plaçant sur un événement Ω_δ de grande probabilité, nous avons pu montrer que les déviations de χ^2_M autour de son espérance sont du même ordre que dans le cas gaussien. Nous avons obtenu (voir section 1.4, lemma 1.1): pour tout $x > 0$,

$$\mathbb{P}\left(\chi^2_M \mathbb{1}\!\mathrm{I}_{\Omega_\delta} \geq \frac{\sigma^2}{n}|M| + 4\frac{\sigma^2}{n}(1+b\delta)\sqrt{2|M|x} + 2\frac{\sigma^2}{n}(1+b\delta)x\right) \leq e^{-x} \tag{10}$$

où b et σ^2 sont définis par (9), et où $|M|$ est le nombre de morceaux de la partition M. Grâce à cette inégalité de concentration, nous montrons (voir section 1.3, theorem 1.1) qu'en prenant une pénalité de la forme

$$\mathrm{pen}(M) = K\frac{\sigma^2}{n}|M| + \kappa_1\frac{\sigma^2}{n}\sqrt{|M|x_M} + \left(\kappa_2\frac{\sigma^2}{n} + \frac{4Rb}{n}\right)x_M$$

avec des poids $(x_M)_{M\in\mathcal{M}_n}$ vérifiant (7) on obtient une inégalité de type oracle. La pénalité obtenue ici est de la même forme que la pénalité (6) obtenue dans le cas gaussien, sauf qu'elle ne concerne que des modèles S_M de fonctions constantes par morceaux.

Le principal attrait des modèles S_M de fonctions constantes par morceaux est leur simplicité. Les estimateurs $\hat{s}_M$ associés donnent des informations faciles à interprêter. Le revers de la médaille est qu'ils sont fortement discontinus. Même les meilleurs d'entre eux ne donnent souvent pas une bonne estimation de s, surtout si s est très régulière. La sélection d'un modèle de fonctions constantes par morceaux ne permet pas de construire un bon estimateur. Elle permet en revanche de détecter les ruptures de $s : \mathcal{X} \longrightarrow \mathbb{R}$. Pour $\mathcal{X} = [0,1]$, considérons à nouveau la collection $\mathcal{M}_n$ de toutes les partitions dont les noeuds appartiennent à la grille de valeurs $(v_1, \ldots, v_{N-1})$. Les poids $x_M = |M|\left(a + \log\frac{N}{|M|}\right)$ avec $a > 1$ satisfont l'inégalité (7) et l'on obtient une pénalité de la forme:

$$\mathrm{pen}(M) = (\sigma^2 + Rb)\frac{|M|}{n}\left(\alpha\log\frac{N}{|M|} + \beta\right)$$

Lebarbier [16] a travaillé sur la détection de ruptures d'un signal gaussien. Grâce au résultat de Birgé et Massart, elle obtient la même forme de pénalité avec $b = 0$. Elle donne ensuite une méthode de calibration des constantes de la pénalité à partir des données, et obtient une procédure qui permet de sélectionner une partition $\hat{M}$ (et donc de déterminer des ruptures) de manière automatique à partir des données. Le résultat de ce chapitre permet de proposer une procédure de détection de ruptures similaire sans supposer les observations gaussiennes.

Si la détection de ruptures est l'application la plus immédiate de la sélection de modèles constants par morceaux, ce n'est pas la seule. Nous utilisons dans le chapitre 2 des modèles

de fonctions constantes par morceaux pour déterminer les variables influentes.

Des prédicteurs de type regressogrammes sont construits par le célèbre algorithme CART. La première étape de l'algorithme CART consiste à construire de manière récursive dyadique une partition fine de l'espace $\mathcal{X}$. La partition initiale est celle constituée d'une seule région: l'espace $\mathcal{X}$ tout entier. A chaque étape, on découpe en deux les régions de la partition existante. Cette construction est naturellement représentée par un arbre de profondeur maximale noté T_{max} et dont les feuilles forment une partition fine de $\mathcal{X}$ notée M_0. A chaque sous-arbre élagué T de T_{max} correspond une partition M_T de $\mathcal{X}$ construite à partir de M_0. On dispose alors de la collection de partitions $(M_T)_{T \preceq T_{max}}$, où $T \preceq T_{max}$ signifie T sous-arbre élagué de T_{max}. La deuxième étape de l'algorithme CART consiste à élaguer l'arbre T_{max} en minimisant le critère $\gamma_n(\hat{s}_{M_T}) + \alpha' \frac{|M_T|}{n}$. Or, comme $|\{T \preceq T_{max}; |M_T| = D\}| \leq \frac{1}{D} \binom{2(D-1)}{D-1} \leq \frac{2^{2D}}{D}$, on peut prendre $x_M = a|M|$ avec $a > 2 \log 2$, et d'après notre résultat de sélection de modèles, une pénalité de la forme $\text{pen}(M) = \alpha(\sigma^2 + Rb)\frac{|M|}{n}$ permet d'obtenir une inégalité de type oracle. Notre résultat valide donc la procédure d'élagage de CART dans le cadre d'une régression non gaussienne avec des points $\mathbf{x}_i$ déterministes.

Sélection de variables au travers de CART

Nous disposons d'un échantillon d'observations $\mathcal{L} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ constitué de n copies indépendantes d'un couple (X, Y) où $X = (X^1, \ldots, X^p)$ est à valeurs dans $\mathbb{R}^p$ et a pour loi μ et où Y est à valeurs soit dans $\mathbb{R}$ soit dans $\{0, 1\}$. Nous considérons le cadre de la régression dans lequel Y est à valeurs dans $\mathbb{R}$ et le cadre de la classification binaire dans lequel Y est à valeurs dans $\{0, 1\}$. Les variables $X^1, \ldots, X^p$ sont les variables explicatives et Y est la variable réponse. A l'aide des données, notre but est de déterminer un petit nombre de variables parmis $\{X^1, \ldots, X^p\}$ qui suffisent à elles seules à bien expliquer ou prédire la réponse Y. Nous proposons ici une méthode de sélection de variables qui utilise l'algorithme CART [7] et nous adoptons une approche sélection de modèles par pénalisation.

Dans le cadre de la régression, Y est à valeurs dans $\mathbb{R}$ et nous appelons prédicteur toute fonction mesurable $u : \mathbb{R}^p \longrightarrow \mathbb{R}$. La fonction de régression s est définie par:

$$s(x) = \mathbb{E}\left(Y | X = x\right).$$

s est le meilleur prédicteur au sens des moindres carrés, i.e. $s = \arg \min_{u:\mathbb{R}^p \to \mathbb{R}} \mathbb{E}\left(\gamma(u, X, Y)\right)$, où $\gamma(u, x, y) = (y - u(x))^2$ est le contraste des moindres carrés.
Dans le cadre de la classification binaire, Y est à valeurs dans $\{0, 1\}$ et nous appelons classifieur (ou prédicteur) toute fonction mesurable $u : \mathbb{R}^p \longrightarrow \{0, 1\}$. Nous préférons alors noter η la fonction de régression et garder la notation s pour le classifieur de Bayes défini par:

$$s(x) = \mathbb{1}_{\eta(x) \geq 1/2} \quad \text{avec } \eta(x) = \mathbb{E}\left(Y | X = x\right).$$

s est le meilleur classifieur au sens des moindres carrés, i.e. $s = \arg \min_{u:\mathbb{R}^p \to \{0,1\}} \mathbb{E}\left(\gamma(u, X, Y)\right)$.
Remarquons que, comme Y et les classifieurs u sont à valeurs dans $\{0, 1\}$, $\gamma(u, x, y) = \mathbb{1}_{y \neq u(x)}$.

Dans les deux cadres, nous voulons déterminer un petit paquet M de variables (i.e. M sous-ensemble de $\{X^1, \ldots, X^p\}$) tel que l'on puisse construire à partir de M un prédicteur $\tilde{s}_M$ qui ne dépend que des variables de M et dont le risque quadratique $\mathbb{E}[l(s, \tilde{s}_M)]$ est petit. l est ici la perte quadratique définie par $l(s,u) = \mathbb{E}\left(\gamma(u,X,Y)\right) - \mathbb{E}\left(\gamma(s,X,Y)\right)$.

Comme nous allons utiliser CART, rappelons que la première étape de CART consiste à construire récursivement une partition fine de l'espace des covariables (ici $\mathbb{R}^p$). La partition initiale est constituée d'une seule région: l'espace $\mathbb{R}^p$ tout entier. A chaque étape, les régions de la partition existante sont découpées en deux selon un découpage du type "$X^j \leq c$" où $1 \leq j \leq p$ et $c \in \mathbb{R}$. Cette construction est naturellement représentée par un arbre de profondeur maximale noté T_{max} et dont les feuilles forment une partition fine de $\mathbb{R}^p$.

Notre procédure est la suivante. Nous commençons par appliquer la première étape de CART à chaque paquet M de variables. Grâce à CART, pour chaque paquet M de variables, nous construisons à l'aide des données et en autorisant uniquement les découpages faisant intervenir les variables du paquet M un arbre maximal $T_{max}^{(M)}$. Nous considérons alors pour tout sous-arbre élagué T de $T_{max}^{(M)}$ (noté $T \preceq T_{max}^{(M)}$) le modèle $S_{M,T}$ constitué des fonctions constantes par morceaux définies sur la partition associée aux feuilles de l'arbre T. Nous notons $\hat{s}_{M,T}$ l'estimateur des moindres carrés de s sur $S_{M,T}$. A chaque paquet M de variables est donc associé une liste de modèles $(S_{M,T})_{T \preceq T_{max}^{(M)}}$ et une liste d'estimateurs $(\hat{s}_{M,T})_{T \preceq T_{max}^{(M)}}$. Le paquet idéal M^* est le plus petit paquet tel que:

$$\exists\, T^* \preceq T_{max}^{(M^*)} :\ (M^*, T^*) = \arg \min_{(M,T)} \mathbb{E}\left(l(s, \hat{s}_{M,T})\right)$$

où le minimum est pris parmis tous les couples (M,T) avec M sous-ensemble de $\{X^1, \ldots, X^p\}$ et $T \preceq T_{max}^{(M)}$. s étant inconnu, M^* est aussi inconnu. Notre but est de déterminer à l'aide des données un paquet $\hat{M}$ "proche" de M^*. Nous adoptons l'approche de la sélection de modèles par pénalisation. Nous estimons (M^*, T^*) par

$$\widehat{(M,T)} = \arg \min_{(M,T)} \gamma_n(\hat{s}_{M,T}) + \mathrm{pen}(M,T)$$

Le résultat principal de ce chapitre recommande une pénalité de la forme:

$$\mathrm{pen}(M,T) = \alpha' \frac{|T|}{n} + \beta' \frac{|M|}{n} \left(1 + \log\left(\frac{p}{|M|}\right)\right).$$

où $|T|$ est le nombre de feuilles de l'arbre T et $|M|$ est le nombre de variables dans le paquet M. Les propositions 2.3.1 et 2.3.2 dans le cadre de la régression et les propositions 2.4.1 et 2.4.2 dans le cadre de la classification justifient cette pénalité par le biais d'inégalités de type oracle. Le premier terme de la pénalité $\left(\alpha' \frac{|T|}{n}\right)$ correspond à la pénalité utilisée dans le critère d'élagage de CART. A M fixé, nos inegalités de type oracle permettent donc de justifier l'étape d'élagage de CART dans le cadre de la régression non gaussienne et de la classification binaire. Le deuxième terme de la pénalité pénalise les gros paquets de variables. Nous notons pour tout M

$$\hat{T}_M = \arg \min_{T \preceq T_{max}^{(M)}} \left\{ \gamma_n(\hat{s}_{M,T}) + \alpha' \frac{|T|}{n} \right\}.$$

L'arbre $\hat{T}_M$ s'obtient par la deuxième étape de CART: l'étape d'élagage. Les modèles $(M, \hat{T}_M)$ servent de modèles de référence. Le paquet $\hat{M}$ est sélectionné en minimisant $\gamma_n(\hat{s}_{M,\hat{T}_M}) + \text{pen}(M, \hat{T}_M)$, qui est un critère permettant de faire un compromis entre l'adéquation aux données du modèle $(M, \hat{T}_M)$ et sa complexité (mesurée par le nombre de feuilles de $\hat{T}_M$ et le nombre de variables dans M).

Estimation d'une fonction de régression par un polynôme par morceaux

Dans le troisième chapitre, comme dans le premier, nous avons des observations $(\mathbf{x}_i, Y_i)_{1 \leq i \leq n}$ vérifiant (1) avec $\mathbf{x}_i \in \mathcal{X} = [0,1]^p$ déterministes et $Y_i \in \mathbb{R}$ aléatoires. Chaque Y_i correspond à la valeur bruitée d'une fonction inconnue s au point $\mathbf{x}_i$. Les bruits, notés ε_i, sont des variables aléatoires supposées centrées, indépendantes et de même loi ayant des moments exponentiels au voisinage de 0. Notre but est de construire un estimateur polynomial par morceaux de la fonction s.

Pour cela, nous considérons une collection $\mathfrak{M}_n$ de couples $m = (M, \underline{\mathsf{d}})$ avec M une partition de $[0,1]^p$ et $\underline{\mathsf{d}} = (d_J)_{J \in M} \in \mathbb{N}^M$. Nous associons à chaque couple $m = (M, \underline{\mathsf{d}}) \in \mathfrak{M}_n$ le modèle S_m des fonctions polynomiales par morceaux définies sur la partition M et de degré variable inférieur ou égal à d_J sur la région J de M, ainsi que l'estimateur des moindres carrés de s sur S_m noté $\hat{s}_m$. Nous disposons alors d'une collection $(S_m)_{m \in \mathfrak{M}_n}$ de modèles de polynomes par morceaux et d'une collection $(\hat{s}_m)_{m \in \mathfrak{M}_n}$ d'estimateurs polynomiaux par morceaux. Puis nous déterminons un critère des moindres carrés pénalisés qui permet de sélectionner un modèle $\hat{m}$ dont l'estimateur polynomial par morceaux associé $\hat{s}_{\hat{m}}$ vérifie une inégalité de type oracle. Ce résultat généralise celui du chapitre 1 pour des modèles de polynomes par morceaux au lieu des modèles de fonctions constantes par morceaux.

Pour obtenir ce nouveau résultat de sélection de modèles, nous construisons encore une inégalité de concentration pour les statistiques $\chi^2_m = \|\varepsilon_m\|^2_n$ où $\varepsilon_m = \arg\min_{u \in S_m} \|\varepsilon - u\|^2_n$. Par rapport à l'inégalité de concentration obtenue dans le cas de modèles de fonctions constantes par morceaux, la difficulté supplémentaire est de contrôler la norme infinie de polynômes en fonction de leur norme l^2 discrète relative à la suite de points $(\mathbf{x}_i)_{1 \leq i \leq n}$ (ou à une sous-suite). Pour comparer ces deux normes (voir section 3.6, lemma 3.5), nous supposons que les points $(\mathbf{x}_i)_{1 \leq i \leq n}$ sont "bien répartis" dans $[0,1]^p$. Nous obtenons alors une inégalité de concentration qui généralise (10) (voir section 3.5, lemma 3.3). Puis nous en déduisons (voir section 3.3, theorem 3.1) qu'une pénalité de la forme

$$\text{pen}(m) = K\frac{\sigma^2}{n}D_m + \kappa_1(d,p)\frac{\sigma^2}{n}\sqrt{D_m x_m} + \kappa_2(d,p)\frac{\sigma^2 + Rb}{n}x_m \tag{11}$$

avec p le nombre de variables, d le degré maximal des polynômes, et des poids $(x_m)_{m \in \mathcal{M}_n}$ vérifiant (7), permet d'obtenir une inégalité de type oracle. La pénalité (11) a la même forme que la pénalité (6) obtenue dans le cas gaussien, sauf que κ_1 et κ_2 dépendent ici du nombre p de variables et du degré maximal d des polynômes.

Après avoir choisi une collection $\mathfrak{M}_n$, il faut en pratique déterminer des poids x_m vérifiant (7), préciser le développement de pen(m) et calibrer les constantes inconnues apparaissant dans pen(m) à partir des données. Pour $p = 1$, ce travail est réalisé par Comte et Rozenholc [10]. Ils obtiennent un algorithme qui détermine automatiquement une partition, une série de degrés, et construit un estimateur polynomial par morceaux. La forme de la pénalité utilisée dans leur algorithme est validée théoriquement par [2] quand les bruits sont sous-gaussiens. Notre résultat permet de la valider dans un cas plus général.

Dans une deuxième partie de ce chapitre, pour p quelconque, nous proposons une procédure d'estimation polynomiale par morceaux basée sur l'algorithme CART. Cette procédure est simple et rapide, mais impose un degré uniforme sur chaque morceau de la partition.

Le célèbre algorithme MARS de Friedman [12] peut lui aussi être interprété comme une extension de CART à l'estimation polynomiale par morceaux. La différence majeure avec notre travail est qu'il construit des estimateurs continus même aux noeuds des partitions. Pour cela, il considère des modèles engendrés par des splines, et il sélectionne un modèle par un critère pénalisé. Malheureusement, notre résultat ne s'applique pas pour des modèles engendrés par des splines. La pénalité (11) que nous obtenons pour des modèles de polynômes par morceaux a la même forme que la pénalité (6) obtenue par Birgé et Massart [4] dans un cadre gaussien pour des collections de modèles quelconques. Nous pouvons donc supposer que la pénalité (11) donne aussi de bon résultats pour les modèles utilisés par MARS. En étudiant la complexité de la collection de modèles obtenue à l'issue de la première étape de MARS (i.e. en comptant le nombre de modèles de dimension donnée), nous déterminons des poids vérifiant (7), et en les substituant dans (11) nous obtenons la même forme de pénalité (à un facteur log près) que celle proposée par Friedman.

Application en fiabilité

Dans le dernier chapitre, nous présentons un travail motivé par un problème de fiabilité. Nous nous intéressons aux tests accélérés dont l'objectif est d'obtenir dans des délais raisonnables des informations sur le temps de survie de composants ou de systèmes. Ces tests accélérés consistent à soumettre des composants tests à des niveaux de stress plus élevés que la normale. Puis, grâce à des modèles statistiques physiquement raisonnables, les résultats sont extrapolés pour obtenir des informations sur le temps de survie du même composant dans les conditions standards.

L'un des modèles les plus utilisés est le modèle Weibull log-linéaire qui suppose que le temps de survie suit une loi de Weibull de paramètres η et β:

$$f(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} e^{-\left(\frac{t}{\eta}\right)^{\beta}} \mathbb{1}_{t>0},$$

avec β une constante strictement positive et η une fonction log-linéaire des variables de stress:

$$\log \eta = c_0 + \sum_{j=1}^{p} a_j x^j.$$

Les variables x^j sont des variables de stress ou plus généralement des fonctions connues d'une ou plusieurs variables de stress. Le modèle d'Arrhenius, par exemple, suppose que η est une fonction log-linéaire de la variable $1/T$ où T est la température.

Nous notons Y le logarithme du temps de survie. Y suit alors une loi des valeurs extrêmes de moyenne $\mu = \log \eta - \frac{\gamma}{\beta}$ où $\gamma = 0.5772\ldots$ est la constante d'Euler, et de variance $\tau^2 = \frac{1}{\beta^2}\frac{\pi^2}{6}$. Nous avons donc

$$\begin{aligned} Y &= \mu + Z, \\ \mu &= a_0 + \sum_{j=1}^{p} a_j x^j, \end{aligned} \tag{12}$$

avec Z une variable qui suit une loi des valeurs extrêmes de moyenne 0 et de variance τ^2, et $(a_0, a_1, \ldots, a_p, \tau)$ des paramètres inconnus.

A l'issu du test, nous obtenons des données $(\mathbf{x}_i, Y_i)_{1 \leq i \leq n}$, où $\mathbf{x}_i = (x_i^1, \ldots, x_i^p)$ sont les valeurs des variables de stress auxquelles le $i^{\text{ème}}$ composant a été soumis, et où Y_i est le temps de survie du $i^{\text{ème}}$ composant. Rappelons que les valeurs $(x_i^1, \ldots, x_i^p)$ sont plus élevées que la normale. Grâce à ces données, on construit des estimateurs $(\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_p, \hat{\tau})$ de $(a_0, a_1, \ldots, a_p, \tau)$. Pour un niveau standard de stress $\mathbf{x} = (x^1, \ldots, x^p)$, la moyenne du logarithme du temps de survie $(\mu(\mathbf{x}))$ est estimée par $\hat{\mu}(\mathbf{x}) = \hat{a}_0 + \sum_{j=1}^{p} \hat{a}_j x^j$.

Cette méthode d'estimation ne donne pas toujours de bons résultats. Ceci peut s'expliquer par exemple lorsque le composant étudié a plusieurs défauts notés $D_1, D_2, \ldots$, et lorsque seul le défaut D_1 peut causer une panne dans les conditions standards, les autres défauts n'apparaissant que pour des niveaux de stress hors normes. Dans ce cas, on ne peut pas estimer les paramètres à partir de données obtenues avec des stress trop élevés et interpoler la relation (12) pour des stress standards. Lors d'un test accéléré, il faut augmenter suffisamment les niveaux de stress pour obtenir des informations dans des délais raisonnables, mais il ne faut pas dépasser les niveaux au delà desquels des défauts inexistants dans les conditions standards se manifestent.

Nous proposons donc de remplacer la relation (12) par une relation linéaire par morceaux:

$$\mu = \sum_{J \in M} \left\{ a_{0,J} + \sum_{j=1}^{p} a_{j,J} x^j \right\} \mathbb{1}_{\mathbf{x} \in J}$$

où M est une partition inconnue de l'ensemble $\mathcal{S}$ de toutes les valeurs possibles pour $\mathbf{x}$. Grâce au théorème 3.1 du chapitre 3, nous obtenons un critère des moindres carrés pénalisés qui permet de détecter les ruptures de μ, cad d'estimer la partition M. Pour obtenir des informations sur le temps de survie dans les conditions normales, il ne faudra utiliser que le morceau de la partition $\hat{M}$ correspondant aux plus petites valeurs de stress.

Nous considérons une collection $\mathcal{M}_n$ de partitions de $\mathcal{S}$ construites à partir d'une grille (ou quadrillage) de $\mathcal{S}$, et nous sélectionnons une partition $\hat{M}$ en minimisant un critère des moindres carrés pénalisé. Grâce au théorème 3.1, nous obtenons la forme de pénalité suivante:

$$\forall\ M \in \mathcal{M}_n \quad \text{pen}(M) = \frac{|M|}{n}\left(\alpha \log\left(\frac{N_n}{|M|}\right) + \beta\right)$$

où $|M|$ est le nombre de morceaux de la partition M et N_n est le nombre de morceaux de la grille initiale. Cette pénalité dépend de deux constantes inconnues α et β. Dans le cas où il n'y a qu'une seule variable x, nous proposons deux méthodes pour calibrer α et β à partir des données. La première méthode est basée sur l'heuristique de Massart (rappelée dans la section 4.4.1) et consiste à estimer simultanément α et β à partir des données, en ajustant le contraste associé à la "meilleure" partition de K morceaux $(\gamma_n(\hat{\mu}_K))$ sur $-\frac{1}{2}\frac{K}{n}\left(\alpha \log\left(\frac{N_n}{K}\right) + \beta\right) - \gamma$ pour une suite de K grands. La deuxième méthode consiste à calibrer la constante $c = \beta/\alpha$ à l'aide de données simulées correspondant à $\mu = 0$ et $\tau = 1$, puis à déterminer la constante multiplicative α à partir des données d'apprentissage grâce à la règle de Birgé et Massart [4]. Les deux méthodes sont évaluées et comparées sur un jeu de données simulées.

Dans les quatre chapitres suivants, je présente mes travaux en anglais.

Chapter 1

Histogram selection in non Gaussian regression

Abstract: We deal with the problem of choosing a piecewise constant estimator of a regression function s mapping $\mathcal{X}$ into $\mathbb{R}$. We consider a non Gaussian regression framework with deterministic design points, and we adopt the non asymptotic approach of model selection via penalization developed by Birgé and Massart. Given a collection of partitions of $\mathcal{X}$, with possibly exponential complexity, and the corresponding collection of piecewise constant estimators, we propose a penalized least squares criterion which selects a partition whose associated estimator performs approximately as well as the best one, in the sense that its quadratic risk is close to the infimum of the risks. The risk bound we provide is non asymptotic.

Keywords: CART, change-points detection, concentration inequalities, model selection, oracle inequalities, regression

1.1 Introduction

We consider the fixed design regression framework. We observe n pairs $(x_i, Y_i)_{1\leq i\leq n}$, where the x_i's are fixed points belonging to some set $\mathcal{X}$ and the Y_i's are real valued random variables. We suppose that:

$$Y_i = s(x_i) + \varepsilon_i, \qquad 1 \leq i \leq n, \tag{1.1}$$

where s is an unknown function mapping $\mathcal{X}$ into $\mathbb{R}$, and $(\varepsilon_i)_{1\leq i\leq n}$ are centered, independent and identically distributed random perturbations. Our aim is to get informations on s from the observations $(x_i, Y_i)_{1\leq i\leq n}$.

In order to get a simple estimator of s, we consider a partition M_0 of $\mathcal{X}$ with a large number of small cells and we minimize the least squares contrast over the class S_{M_0} of piecewise constant functions defined on the partition M_0. The resulting estimator is denoted by $\hat{s}_{M_0}$ and is called the least squares estimator over S_{M_0}. S_{M_0} is called the histogram model associated with M_0. It is a linear space with finite dimension $D_{M_0} = |M_0|$, where $|M_0|$ is the number of cells of the partition M_0. Denoting $\|.\|_n$ the Euclidean norm on $\mathbb{R}^n$ scaled by a factor $n^{-1/2}$ and denoting the same way a function $u \in \mathbb{R}^{\mathcal{X}}$ and the corresponding vector $(u(x_i))_{1\leq i\leq n} \in \mathbb{R}^n$,

the quadratic risk of $\hat{s}_{M_0}$, $\mathbb{E}\left(\|s-\hat{s}_{M_0}\|_n^2\right)$, is the sum of two terms, respectively called bias and variance:

$$\mathbb{E}\left(\|s-\hat{s}_{M_0}\|_n^2\right) = \inf_{u\in S_{M_0}} \|s-u\|_n^2 + \frac{\tau^2}{n}|M_0| \quad \text{where } \tau^2 = \mathbb{E}(\varepsilon_i^2).$$

We see in this expression of the risk of $\hat{s}_{M_0}$ that $\hat{s}_{M_0}$ behaves poorly when M_0 has a too large number of cells and that we should rather choose a partition M built from M_0 (or equivalently a histogram model $S_M \subset S_{M_0}$) which makes a better trade-off between the bias $\inf_{u\in S_M} \|s-u\|_n^2$ and the variance $\frac{\tau^2}{n}|M|$.

Our estimation procedure is as follows. We consider a collection $\mathcal{M}_n$ of partitions of $\mathcal{X}$ and the corresponding collection $(S_M)_{M\in\mathcal{M}_n}$ of histogram models. Denoting $\hat{s}_M$ the least squares estimator over the model S_M, the best model is the one which minimizes $\mathbb{E}\left(\|s-\hat{s}_M\|_n^2\right)$. Unfortunately this model depends on s. The aim of model selection is to propose a data driven criterion, whose minimizer among $(S_M)_{M\in\mathcal{M}_n}$ is an approximately best model. We select a model $S_{\hat{M}}$ by minimizing over $\mathcal{M}_n$ a penalized least squares criterion $\mathrm{crit}(M) = \|Y-\hat{s}_M\|_n^2 + \mathrm{pen}(M)$:

$$\hat{M} = \arg\min_{M\in\mathcal{M}_n} \left\{\|Y-\hat{s}_M\|_n^2 + \mathrm{pen}(M)\right\}.$$

The estimator $\hat{s}_{\hat{M}}$ is called the penalized least squares estimator. The penalty pen has to be chosen such that the model $S_{\hat{M}}$ is close to the optimal model, more precisely such that

$$\mathbb{E}\left(\|s-\hat{s}_{\hat{M}}\|_n^2\right) \le C \inf_{M\in\mathcal{M}_n} \mathbb{E}\left(\|s-\hat{s}_M\|_n^2\right). \tag{1.2}$$

The inequality (1.2) will be referred to as the oracle inequality. It bounds the risk of the penalized least squares estimator by the infimum of the risks on a given model up to a constant C. The main result of this paper determines a penalty pen which leads to an oracle type inequality.

The proposed penalty has the same form as those obtained by Birgé and Massart [4] in the Gaussian case and those obtained by Baraud, Comte and Viennet [2] in the sub-Gaussian case, for any collections of models (not only for histogram models). In this paper, the $(\varepsilon_i)_{1\le i\le n}$ are only supposed to have exponential moments around 0. We follow the same ideas and techniques as [4, 2]. The main point is to control the statistics $\chi_M^2 = \|\varepsilon_M\|_n^2$ where $\varepsilon_M = \arg\min_{u\in S_M} \|\varepsilon-u\|_n^2$. In the Gaussian case, the $\frac{n}{\tau^2}\chi_M^2$'s are χ^2 distributed. But in the non Gaussian case, it is much more difficult to study the deviations of these statistics around their expectations. Under an even milder integrability condition on the $(\varepsilon_i)_{1\le i\le n}$ (assuming that $\mathbb{E}\left(|\varepsilon_i|^p\right) < +\infty$ for some $p \ge 2$), Baraud [1] gives a polynomial concentration inequality for the χ_M^2's (see inequality (8) in the introduction). This inequality allows him to prove that penalties $\mathrm{pen}(M) = K'\frac{\tau^2}{n}D_M$, with $K' > 1$, lead to oracle type inequalities when the number of models with a given dimension D is a polynomial function of D. In order to deal with bigger collections of models, we need exponential concentration inequalities for the χ_M^2's. By writing $\chi_M = \sup_{u\in B_M} <\varepsilon, u>_n$, with $B_M = \{u\in S_M;\ \|u\|_n \le 1\}$, we can apply Bousquet's exponential concentration inequality for a supremum of an empirical process [6]. Unfortunately this general result is not sufficient here. Instead of viewing χ_M as a supremum, we can

view χ^2_M as a χ^2 like statistic and write it as a sum of squares. For histogram models, we can then build adequate exponential concentration inequalities by hand, using only Bernstein's inequality. This is the reason why we determine a penalty which we prove to lead to an oracle inequality only for histogram models.

Thanks to this penalty, given a collection $\mathcal{M}_n$ of partitions, we get an estimator $\hat{s}_{\hat{M}}$ which is simple, easy to interpret and close to the optimal one among the collection of piecewise constant estimators $(\hat{s}_M)_{M\in\mathcal{M}_n}$. Unfortunately, since the estimators $(\hat{s}_M)_{M\in\mathcal{M}_n}$ are sharply discontinuous, even the best one may not provide an accurate estimation of s.

Histogram model selection may not lead to an accurate estimation of the regression function s, but it enables to detect the change-points of s. In the framework (1.1) with $\mathcal{X} = [0,1]$ and $x_i = \frac{i}{n}$, Lebarbier [16, chapter 2] considers the collection $\mathcal{M}_n$ of all partitions with endpoints belonging to the grid $(x_i)_{1\leq i\leq n}$, and the corresponding collection $(S_M)_{M\in\mathcal{M}_n}$ of histogram models. For this collection, $|\{M \in \mathcal{M}_n;\ |M| = D\}| = \binom{n-1}{D-1}$, and therefore Baraud's result [1] does not apply to this case. Assuming the perturbations ε_i to be Gaussian, Lebarbier applies the model selection result of Birgé and Massart [3] to the collection $(S_M)_{M\in\mathcal{M}_n}$. She gets a penalty defined up to two multiplicative constants. Then she proposes a method to calibrate them according to the data and therefore gives a procedure to automatically detect the change-points of a Gaussian signal according to the data. Thanks to our result, we can propose a similar procedure without assuming the perturbations ε_i to be Gaussian.

One of the most famous statistical issues is variable selection. In the classical linear regression framework,

$$Y_i = \sum_{j=1}^{p} \beta_j x_i^j + \varepsilon_i, \qquad 1 \leq i \leq n,$$

selecting a small subset of variables $V \subset \{x^1, \ldots, x^p\}$ which explain "at best" the response Y is equivalent to choosing the "best" model S_V of functions linear in $\{x^j \in V\}$. Instead of considering linear interaction between $\left(x^1, \ldots, x^p\right)$ and Y, we can use histogram models. Sauvé and Tuleau (see chapter 2 or [23]) propose a variable selection procedure based on histogram model selection.

The CART algorithm (Classification And Regression Trees), proposed by Breiman *et al.* [7], involves histogram models. Our result allows to validate the pruning step of CART in a non Gaussian regression framework.

The paper is organized as follows. The section 1.2 presents the statistical framework and some notations. The section 1.3 gives the main result. To get this result, we have to control a χ^2 like statistic. The section 1.4 is more technical, it exposes a concentration inequality for a χ^2 like statistic and explains why the existing concentration inequality, due to Bousquet, is not sufficient. Sections 1.5 and 1.6 are devoted to the proofs.

1.2 The statistical framework

In this paper, we consider the regression framework defined by (1.1) and we look for a best or approximately best piecewise constant estimator of s. In this section, we precise the integrability condition that should satisfy the random perturbations $(\varepsilon_i)_{1\leq i\leq n}$ involved in (1.1), then we define the piecewise constant estimators of s and their risk. We give here some notations needed in the rest of the paper.

1.2.1 The random perturbations

As noted above in the introduction, we assume that the random perturbations $(\varepsilon_i)_{1\leq i\leq n}$ have finite exponential moments around 0. This assumption is equivalent to the existence of two constants $b \in \mathbb{R}_+$ and $\sigma \in \mathbb{R}_+^*$ such that

$$\forall\ \lambda \in (-1/b, 1/b) \qquad \log \mathbb{E}\left(e^{\lambda \varepsilon_i}\right) \leq \frac{\sigma^2\lambda^2}{2(1-b|\lambda|)} \tag{1.3}$$

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

Remark 1.1 *Under assumption (1.3), we have*

$$\forall\ \lambda \in (-1/2b, 1/2b) \qquad \log \mathbb{E}\left(e^{\lambda \varepsilon_i}\right) \leq \sigma^2\lambda^2$$

but we prefer inequality (1.3) to this last inequality because with the last one we loose a factor 2 *in the variance term.*

Remark 1.2 *Thanks to assumption (1.3) and Cramer-Chernoff method (see [19, section 2.1]), we can easily get concentration inequalities for any linear combination of the $(\varepsilon_i)_{1\leq i\leq n}$. First, since the $(\varepsilon_i)_{1\leq i\leq n}$ are independent, we get from inequality (1.3) similar inequalities for any linear combination $\sum_{i=1}^n \alpha_i\varepsilon_i$. Denoting $\|\alpha\|_\infty = \max_{1\leq i\leq n}|\alpha_i|$ and $v = \sigma^2\left(\sum_{i=1}^n \alpha_i^2\right)$,*

$$\forall\ \lambda \in \left(0, \frac{1}{b\|\alpha\|_\infty}\right) \qquad \log \mathbb{E}\left(e^{\lambda\sum_{i=1}^n \alpha_i\varepsilon_i}\right) \leq \frac{v\lambda^2}{2(1-b\|\alpha\|_\infty\lambda)}. \tag{1.4}$$

We denote by $\psi(\lambda)$ the right term of (1.4) and by $h(u) = 1+u-\sqrt{1+2u}$ for any $u \in \mathbb{R}_+^*$. Then, applying Cramer-Chernoff method, since for any $x > 0$

$$\sup_{0<\lambda<\frac{1}{b\|\alpha\|_\infty}} \{\lambda x - \psi(\lambda)\} = \frac{v}{b^2\|\alpha\|_\infty^2} h\left(\frac{b\|\alpha\|_\infty x}{v}\right),$$

we get for any $x > 0$:

$$\mathbb{P}\left(\sum_{i=1}^n \alpha_i\varepsilon_i \geq x\right) \quad \leq \quad \exp\left(-\frac{v}{b^2\|\alpha\|_\infty^2} h\left(\frac{b\|\alpha\|_\infty x}{v}\right)\right).$$

Finally we deduce the two following inequalities:

- *Since h is inversible with $h^{-1}(u) = u + \sqrt{2u}$,*

$$\mathbb{P}\left(\sum_{i=1}^{n} \alpha_i \varepsilon_i \geq \sqrt{2vx} + b\|\alpha\|_\infty x\right) \quad \leq \quad e^{-x}.$$

- *Since $h(u) \geq \frac{u^2}{2(1+u)}$,*

$$\mathbb{P}\left(\sum_{i=1}^{n} \alpha_i \varepsilon_i \geq x\right) \quad \leq \quad \exp\left(\frac{-x^2}{2\left(v + b\|\alpha\|_\infty x\right)}\right).$$

1.2.2 The piecewise constant estimators

For a given partition M of $\mathcal{X}$, we denote S_M the space of piecewise constant functions defined on the partition M and $\hat{s}_M$ the least squares estimator over S_M.

$$\hat{s}_M = \arg\min_{u \in S_M} \gamma_n(u) \text{ with } \gamma_n(u) = \|Y - u\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - u(x_i))^2$$

where $\|.\|_n$ denotes the Euclidean norm on $\mathbb{R}^n$ scaled by a factor $n^{-1/2}$, $Y = (Y_i)_{1 \leq i \leq n}$, and for $u \in \mathbb{R}^{\mathcal{X}}$, the vector $(u(x_i))_{1 \leq i \leq n} \in \mathbb{R}^n$ is denoted by u too. S_M is the histogram model associated with M and $\hat{s}_M$ is the piecewise constant estimator belonging to S_M which plays the role of benchmark among all the estimators in S_M.

Let now calculate the quadratic risk of $\hat{s}_M$: $\mathbb{E}\left(\|s - \hat{s}_M\|_n^2\right)$. To this end, we denote by

$s_M = \arg\min_{u \in S_M} \|s - u\|_n^2$,

$\varepsilon_M = \arg\min_{u \in S_M} \|\varepsilon - u\|_n^2$ where $\varepsilon = (\varepsilon_i)_{1 \leq i \leq n}$,

$|M|$ the number of elements of the partition M.

$\hat{s}_M$, s_M and ε_M are respectively the orthogonal projections of Y, s and ε on the space S_M according to $\|.\|_n$. Thanks to Pythagore's equality, we get that:

$$\mathbb{E}\left(\|s - \hat{s}_M\|_n^2\right) \quad = \quad \|s - s_M\|_n^2 + \mathbb{E}\left(\|\varepsilon_M\|_n^2\right).$$

For any element J of the partition M, we denote by $|J| = |\{1 \leq i \leq n;\ x_i \in J\}|$ and by $\mathbb{1}_J : x \to 1$ if $x \in J$ and 0 if $x \notin J$. Since $\left(\sqrt{\frac{n}{|J|}}\mathbb{1}_J\right)_{J \in M}$ is an orthonormal basis of $(S_M, \|.\|_n)$, we have

$$\|\varepsilon_M\|_n^2 \quad = \quad \sum_{J \in M}\left\langle \varepsilon, \sqrt{\frac{n}{|J|}}\mathbb{1}_J \right\rangle_n^2 \quad = \quad \frac{1}{n}\sum_{J \in M}\frac{\left(\sum_{x_i \in J}\varepsilon_i\right)^2}{|J|}. \tag{1.5}$$

Since $(\varepsilon_i)_{1 \leq i \leq n}$ are centered, independent and identically distributed random variables with $\mathbb{E}(\varepsilon_i^2) \leq \sigma^2$, we get that

$$\mathbb{E}\left(\|\varepsilon_M\|_n^2\right) \quad = \quad \mathbb{E}(\varepsilon_1^2)\frac{|M|}{n} \quad \leq \quad \sigma^2\frac{|M|}{n}.$$

Therefore

$$\mathbb{E}\left(\|s - \hat{s}_M\|_n^2\right) \quad = \quad \|s - s_M\|_n^2 + \mathbb{E}(\varepsilon_1^2)\frac{|M|}{n} \quad \leq \quad \|s - s_M\|_n^2 + \sigma^2\frac{|M|}{n}.$$

Remark 1.3 *In the following, the statistic $\|\varepsilon_M\|_n^2$ is denoted by χ_M^2. Thanks to the decomposition (1.5), we can see χ_M^2 as a χ^2 like statistic. If the $(\varepsilon_i)_{1\leq i\leq n}$ were Gaussian variables with variance τ^2, then the variables $\frac{n}{\tau^2}\chi_M^2$ would be $\chi^2(|M|)$-distributed.*

1.3 The main theorem

Let M_0 a partition of $\mathcal{X}$ and $\mathcal{M}_n$ a family of partitions of $\mathcal{X}$ built from M_0, i.e. for any $M \in \mathcal{M}_n$ and any element J of M, J is the union of elements of M_0. In the following theorem, we assume that the initial partition M_0 is not too fine in the sense that the elements of the partition M_0 contain a minimal number of points x_i. We measure the fineness of the partition M_0 by the number $N_{min} = \inf_{J\in M_0} |J|$ where $|J| = |\{1 \leq i \leq n;\ x_i \in J\}|$.

The ideal partition M^* minimizes the quadratic risk $\mathbb{E}\left(\|s - \hat{s}_M\|_n^2\right)$ over all the partitions $M \in \mathcal{M}_n$. Unfortunately M^* depends on the unknown regression function s and $\hat{s}_{M^*}$ can not be used as an estimator of s. The purpose of model selection is to propose a data driven criterion which selects a partition $\hat{M}$ whose associated piecewise constant estimator $\hat{s}_{\hat{M}}$ performs approximately as well as $\hat{s}_{M^*}$ in terms of risks. We select a partition $\hat{M}$ by minimizing a penalized least squares criterion $\mathrm{crit}(M) = \|Y - \hat{s}_M\|_n^2 + \mathrm{pen}(M)$ over $\mathcal{M}_n$:

$$\hat{M} = \arg\min_{M\in\mathcal{M}_n} \left\{\|Y - \hat{s}_M\|_n^2 + \mathrm{pen}(M)\right\}.$$

It remains to provide a penalty pen such that the partition $\hat{M}$ is close to the optimal partition, in the sense that the penalized least squares estiamtor $\hat{s}_{\hat{M}}$ satisfies an oracle inequality like (1.2). The following theorem determines a general form of penalty pen which leads to an oracle type inequality for any family of partitions built from a partition M_0 not too fine. We compare our result to those of Birgé and Massart and those of Baraud, and we study in more detail two particular families of partitions.

Example 1: We consider $\mathcal{X} = [0,1]$ and a grid on $[0,1]$ such that there are at least N_{min} points x_i between two consecutive grid points. For example, we can take the grid $(v_j)_{1\leq j\leq [n/N_{min}]}$ with $v_j = x_{jN_{min}}$. We define M_0 as the partition associated with the whole grid, and $\mathcal{M}_n^1$ as the family of all partitions of $[0,1]$ with endpoints belonging to the grid. $\mathcal{M}_n^1$ corresponds to the collection of partitions used by Lebarbier [16] to detect the change-points of a Gaussian signal $s : [0,1] \longrightarrow \mathbb{R}$.

Example 2: We build a partition M_0 by splitting recursively $\mathcal{X}$ and the obtained subsets in two different parts as long as each subset contains at least N_{min} points x_i. A useful representation of this construction is a tree of maximal depth, called maximal tree and denoted by T_{max}. The leaves of T_{max} are the elements of the partition M_0. Every pruned subtree of T_{max} gives a partition of $\mathcal{X}$ built from M_0. We denote by $\mathcal{M}_n^2$ this second family of partitions. $\mathcal{M}_n^2$ corresponds to the family of partitions obtained via the first step of the CART algorithm.

Theorem 1.1 *Let $b \in \mathbb{R}_+$ and $\sigma \in \mathbb{R}_+^*$ such that inequality (1.3) holds.*
Let M_0 a partition of $\mathcal{X}$ such that $N_{min} = \inf_{J\in M_0} |J|$ satisfies $N_{min} \geq 12\frac{b^2}{\sigma^2}\log n$.
Let $\mathcal{M}_n$ a family of partitions of $\mathcal{X}$ built from M_0 and $(x_M)_{M\in\mathcal{M}_n}$ a family of weights such

that

$$\sum_{M\in\mathcal{M}_n} e^{-x_M} \leq \Sigma \in \mathbb{R}_+^*.$$

Assume $\|s\|_\infty \leq R$, *with* R *a positive constant.*
Let $\theta \in (0,1)$ *and* $K > 2-\theta$ *two numbers.*

Taking a penalty satisfying

$$pen(M) \geq K\frac{\sigma^2}{n}|M| + 8\sqrt{2}(2-\theta)\frac{\sigma^2}{n}\sqrt{|M|x_M} + \left\{\left(4(2-\theta)+\frac{2}{\theta}\right)\frac{\sigma^2}{n} + \frac{4Rb}{n}\right\}x_M$$

we have

$$\begin{aligned}\mathbb{E}\left(\|s-\hat{s}_{\hat{M}}\|_n^2\right) \quad \leq \quad & \frac{2}{1-\theta}\inf_M\left\{\|s-s_M\|_n^2 + pen(M)\right\} \\ & + \frac{1}{1-\theta}\left\{8(2-\theta)\left(1+\frac{8(2-\theta)}{K+\theta-2}\right)+\frac{4}{\theta}+2\right\}\frac{\sigma^2}{n}\Sigma + \frac{12}{1-\theta}\frac{Rb}{n}\Sigma \\ & + C(b,\sigma^2,R)\frac{1\!\!1_{b\neq 0}}{n(\log n)^{3/2}}\end{aligned}$$

where $C(b,\sigma^2,R)$ *is a positive constant which depends only on* b, σ^2 *and* R.

This theorem gives the general form of the penalty function

$$\mathrm{pen}(M) = K\frac{\sigma^2}{n}|M| + \left\{\kappa_1(\theta)\frac{\sigma^2}{n}\sqrt{|M|x_M} + \left(\kappa_2(\theta)\frac{\sigma^2}{n}+\frac{4Rb}{n}\right)x_M\right\}$$

The penalty is the sum of two terms: the first one is proportional to $\frac{|M|}{n}$ and the second one depends on the complexity of the family $\mathcal{M}_n$ via the weights $(x_M)_{M\in\mathcal{M}_n}$. For $\theta \in (0,1)$ and $K > 2-\theta$, the penalized least squares estimator $\hat{s}_{\hat{M}}$ satisfies an oracle type inequality with an additional term tending to 0 like $1/n$ when $n \to +\infty$.

$$\mathbb{E}\left(\|s-\hat{s}_{\hat{M}}\|_n^2\right) \quad \leq \quad C_1\inf_M\left\{\|s-s_M\|_n^2 + \mathrm{pen}(M)\right\} + \frac{C_2}{n}$$

where the constant C_1 only depends on θ, whereas C_2 depends on s (via R), on the family of partitions (via Σ) and on the integrability condition of $(\varepsilon_i)_{1\leq i\leq n}$ (via σ^2 and b).

For the two particular families $\mathcal{M}_n$ quoted above, we calculate adequate weights $(x_M)_{M\in\mathcal{M}_n}$ and we get a simpler form of penalty. Before studying these two examples, we compare the general result with those of Birgé and Massart [4], those of Baraud, Comte and Viennet [2] and those of Baraud [1].

If b can be taken equal to zero in (1.3), then the variables $(\varepsilon_i)_{1\leq i\leq n}$ are said to be sub-Gaussian. In this case, we do not need any assumptions neither on N_{min} the minimal number of observations in each element of the partition M_0 nor on s the regression function. And taking a penalty satisfying

$$\mathrm{pen}(M) \geq K\frac{\sigma^2}{n}|M| + 8\sqrt{2}(2-\theta)\frac{\sigma^2}{n}\sqrt{|M|x_M} + \left(4(2-\theta)+\frac{2}{\theta}\right)\frac{\sigma^2}{n}x_M$$

we have

$$\mathbb{E}\left(\|s-\hat{s}_{\hat{M}}\|_n^2\right) \leq \frac{2}{1-\theta}\inf_M\left\{\|s-s_M\|_n^2+\mathrm{pen}(M)\right\} + \frac{1}{1-\theta}\left\{8(2-\theta)\left(1+\frac{8(2-\theta)}{K+\theta-2}\right)+\frac{4}{\theta}+2\right\}\frac{\sigma^2}{n}\Sigma$$

Up to some small differences in the constants (which can be improved by looking more precisely at the proof), this is the result obtained by Birgé and Massart in the Gaussian case. Using the inequality $2\sqrt{|M|x_M} \leq |M|+x_M$, we recover the result obtained by Baraud, Comte and Viennet [2] in the sub-Gaussian case.

Baraud [1] studies the non Gaussian regression framework as defined in (1.1) with a milder integrability condition on the random perturbations than ours. For a collection of histogram models $(S_M)_{M\in\mathcal{M}_n}$ whose complexity is polynomial, our theorem and those of Baraud both validate penalties $\mathrm{pen}(M)$ proportional to $|M|/n$ through an oracle type inequality with an additional term tending to 0 like $1/n$ when $n\to+\infty$. Thanks to Baraud's result, if $|\{M\in\mathcal{M}_n;\ |M|=D\}| \leq \Gamma D^r$ for some constants $\Gamma\in\mathbb{R}_+^*$ and $r\in\mathbb{N}$, one only needs to assume that the random perturbations have a finite absolute moment of order $p>2r+6$. The minimal admissible value of p increases with the degree r of the polynomial complexity. And, whatever p, having a finite absolute moment of order p seems to be not enough to deal with collections of exponential complexity. Our assumption on the exponential moments is too strong when the complexity is polynomial, but it allows us to propose a general form of penalty which is still valid when the complexity is exponential.

Let now see which form of penalty is adapted to the two collections of partitions quoted above. The complexity of the two corresponding collections of histogram models is exponential, and therefore Baraud's result does not apply to this case.

Example 1: Since $\left|\left\{M\in\mathcal{M}_n^1;|M|=D\right\}\right| = \binom{D_0-1}{D-1} \leq \left(\frac{eD_0}{D}\right)^D$, where D_0-1 is the number of grid points, taking $x_M=|M|\left(a+\log\frac{D_0}{|M|}\right)$ with $a>1$ leads to $\sum_{M\in\mathcal{M}_n^1}e^{-x_M} \leq \left(e^{a-1}-1\right)^{-1}\in\mathbb{R}_+^*$. We deduce from the above theorem that:
taking a penalty

$$\mathrm{pen}(M)=\frac{\sigma^2+Rb}{n}|M|\left(\alpha\log\frac{|M_0|}{|M|}+\beta\right)$$

with α and β big enough, we have

$$\mathbb{E}\left(\|s-\hat{s}_{\hat{M}}\|_n^2\right) \leq C_1(\alpha,\beta)\inf_M\left\{\|s-s_M\|_n^2+\frac{\sigma^2+Rb}{n}|M|\left(\log\frac{|M_0|}{|M|}+1\right)\right\}+C_2(\alpha,\beta)\frac{\sigma^2+Rb}{n} + C(b,\sigma^2,R)\frac{1\!\!1_{b\neq 0}}{n(\log n)^{3/2}}$$

Since σ^2, b and R are unknown, we consider penalties of the form $\mathrm{pen}(M)=\frac{|M|}{n}\left(\alpha'\log\frac{|M_0|}{|M|}+\beta'\right)$ and we determine the right constants α' and β' according to the data by using, for example, the same technique as Lebarbier [16]. We get a data driven criterion which selects a close

to optimal partition $\hat{M}$. The endpoints of the partition $\hat{M}$ provide estimators of the change points of the signal s.

Example 2: Thanks to Catalan inequality, $\left|\left\{M \in \mathcal{M}_n^2; |M| = D\right\}\right| \leq \frac{1}{D}\binom{2(D-1)}{D-1} \leq \frac{2^{2D}}{D}$. Thus taking $x_M = a|M|$ with $a > 2\log 2$, we get $\sum_{M\in\mathcal{M}_n^2} e^{-x_M} \leq -\log\left(1 - e^{-(a-2\log 2)}\right) \in \mathbb{R}_+^*$. We deduce from the above theorem that:
taking a penalty

$$\text{pen}(M) = \alpha\frac{\sigma^2 + Rb}{n}|M|$$

with α big enough, we have

$$\mathbb{E}\left(\|s - \hat{s}_{\hat{M}}\|_n^2\right) \quad \leq \quad C_1(\alpha)\inf_M\left\{\|s - s_M\|_n^2 + \frac{\sigma^2 + Rb}{n}|M|\right\} + C_2(\alpha)\frac{\sigma^2 + Rb}{n} + C(b, \sigma^2, R)\frac{\mathbb{1}_{b\neq 0}}{n(\log n)^{3/2}}$$

For this second example, we recommend a penalty pen(M) proportional to $\frac{|M|}{n}$. For such a penalty, the selected model satisfies an oracle inequality with an additional term tending to 0 like $1/n$ when $n \to +\infty$. This result validates the CART pruning step which involves a penalized least squares criterion with $\text{pen}(M) = \alpha'\frac{|M|}{n}$. The last step of CART consists in choosing the right parameter α' via cross-validation or test sample.

Remark 1.4 *If the points $(x_i)_{1\leq i\leq n}$ of the design are random points $(X_i)_{1\leq i\leq n}$, then with the same approach, working first conditionnally to $(X_i)_{1\leq i\leq n}$,we get a similar result. For more details see chapter 2 (or equivalently [23]).*

1.4 The key to determine an adequate form of penalty: a concentration inequality for a χ^2 like statistic

This section is more technical. First we give an expression of $\|s - \hat{s}_{\hat{M}}\|_n^2$, which allows us to see that the penalty pen(M) has to compensate the deviation of the statistic denoted by χ^2_M, in order that the penalized least squares estimator $\hat{s}_{\hat{M}}$ satisfies an oracle type inequality. The square root of this statistic is the supremum of a random process. Then we explain why Bousquet's concentration inequality for the supremum of a random process is not convenient. And finally, by viewing χ^2_M as a χ^2 like statistic and by writing it as a sum of squares, lemma 1.1 gives a concentration inequality for χ^2_M. This concentration inequality is the main point of the proof of theorem 1.1, the remaining of the proof only consists in technical details.

Let us recall that $\hat{M} = \arg\min_{M\in\mathcal{M}_n}\left\{\|Y - \hat{s}_M\|_n^2 + \text{pen}(M)\right\}$
with the penalty pen to be chosen such that

$$\mathbb{E}\left(\|s - \hat{s}_{\hat{M}}\|_n^2\right) \leq C'\inf_M\left\{\|s - s_M\|_n^2 + \sigma^2\frac{|M|}{n}\right\}.$$

According to the definition of $\hat{M}$, we have

$$\|s - \hat{s}_{\hat{M}}\|_n^2 \quad = \quad -2\left\langle\varepsilon, s - \hat{s}_{\hat{M}}\right\rangle_n - \text{pen}(\hat{M}) + \inf_{M\in\mathcal{M}_n}\left\{\|s - \hat{s}_M\|_n^2 + 2\left\langle\varepsilon, s - \hat{s}_M\right\rangle_n + \text{pen}(M)\right\}.$$

Since $\hat{s}_M = s_M + \varepsilon_M$,

$$\langle \varepsilon, s - \hat{s}_M \rangle_n = \langle \varepsilon, s - s_M \rangle_n - \|\varepsilon_M\|_n^2 \text{ and } \|s - \hat{s}_M\|_n^2 = \|s - s_M\|_n^2 + \|\varepsilon_M\|_n^2.$$

Thus

$$\begin{aligned} \|s - \hat{s}_{\hat{M}}\|_n^2 &= 2\|\varepsilon_{\hat{M}}\|_n^2 - 2\left\langle \varepsilon, s - s_{\hat{M}} \right\rangle_n - \mathrm{pen}(\hat{M}) \\ &\quad + \inf_{M \in \mathcal{M}_n} \left\{ \|s - s_M\|_n^2 - \|\varepsilon_M\|_n^2 + 2\langle \varepsilon, s - s_M \rangle_n + \mathrm{pen}(M) \right\} \end{aligned}$$

and

$$\|s - \hat{s}_{\hat{M}}\|_n^2 = \|s - s_{\hat{M}}\|_n^2 + \|\varepsilon_{\hat{M}}\|_n^2.$$

We deduce from these two last equalities that for any $\theta \in (0,1)$,

$$\begin{aligned} (1-\theta)\|s - \hat{s}_{\hat{M}}\|_n^2 &= (2-\theta)\|\varepsilon_{\hat{M}}\|_n^2 - 2\left\langle \varepsilon, s - s_{\hat{M}} \right\rangle_n - \theta\|s - s_{\hat{M}}\|_n^2 - \mathrm{pen}(\hat{M}) \\ &\quad + \inf_{M \in \mathcal{M}_n} \left\{ \|s - s_M\|_n^2 - \|\varepsilon_M\|_n^2 + 2\langle \varepsilon, s - s_M \rangle_n + \mathrm{pen}(M) \right\}. \end{aligned} \tag{1.6}$$

To get an oracle type inequality, the penalty $\mathrm{pen}(M)$ has to compensate the deviations of the statistics

$$\chi_M^2 = \|\varepsilon_M\|_n^2 \quad \text{and} \quad \langle \varepsilon, s - s_M \rangle_n$$

for all partitions $M \in \mathcal{M}_n$ simultaneously.
Thanks to assumption (1.3) and Cramer-Chernoff method (see remark 1.2), it is easy to obtain the following concentration inequality for $\langle \varepsilon, s - s_M \rangle_n$:

$$\text{for all } x > 0 \quad \mathbb{P}\left(\pm < \varepsilon, s - s_M >_n \geq \frac{\sigma}{\sqrt{n}} \|s - s_M\|_n \sqrt{2x} + \frac{b}{n} \left(\max_{1 \leq i \leq n} |s(x_i) - s_M(x_i)| \right) x \right) \leq e^{-x}.$$

If $\|s\|_\infty \leq R$ then

$$\text{for all } x > 0 \quad \mathbb{P}\left(\pm < \varepsilon, s - s_M >_n \geq \frac{\sigma}{\sqrt{n}} \|s - s_M\|_n \sqrt{2x} + \frac{2Rb}{n} x \right) \leq e^{-x}. \tag{1.7}$$

It remains to study the deviation of the statistic χ_M^2 around its expectation. As told in remark 1.3, if the perturbations $(\varepsilon_i)_{1 \leq i \leq n}$ were Gaussian variables with variance τ^2, then the variables $\frac{n}{\tau^2}\chi_M^2$ would be $\chi^2(|M|)$-distributed. Thus χ_M^2 would satisfy for any $x > 0$ the following concentration inequality:

$$\mathbb{P}\left(\chi_M^2 \geq \frac{\tau^2}{n}|M| + 2\frac{\tau^2}{n}\sqrt{|M|x} + 2\frac{\tau^2}{n}x \right) \leq e^{-x} \tag{1.8}$$

and its square root χ_M would satisfy

$$\mathbb{P}\left(\chi_M \geq \frac{\tau}{\sqrt{n}}\sqrt{|M|} + \frac{\tau}{\sqrt{n}}\sqrt{2x} \right) \leq e^{-x}. \tag{1.9}$$

Recall that $\chi_M = \|\varepsilon_M\|_n$ where ε_M is the orthogonal projection of $\varepsilon = (\varepsilon_i)_{1 \leq i \leq n}$ on S_M (more pecisely on $\{u(x_i)_{1 \leq i \leq n};\ u \in S_M\}$). According to Cauchy-Schwarz inequality, we can write χ_M as the supremum of a random process:

$$\chi_M = \|\varepsilon_M\|_n = \sup_{\substack{u \in S_M \\ \|u\|_n = 1}} < \varepsilon, u >_n = \frac{1}{n} \sup_{\substack{u \in S_M \\ \|u\|_n = 1}} \sum_{i=1}^n u_i \varepsilon_i. \tag{1.10}$$

Therefore, if the ε_i were Gaussian, we could apply the concentration inequality for the supremum of a Gaussian process due to Cirel'son, Ibragimov and Sudakov (see [19, chapter 3]). Then we would recover inequality (1.9).

Here, the expression (1.10) is still valid, but the $(\varepsilon_i)_{1\leq i\leq n}$ are not Gaussian, they are only supposed to have exponential moments around 0. Our first idea was to consider expression (1.10) and use Bousquet's concentration inequality for the supremum of an empirical process instead of the Gaussian result of Cirel'son, Ibragimov and Sudakov. Thanks to Bousquet's result [6], we have for any $x>0$ and any $\gamma>0$:

$$\mathbb{P}\left(\chi_M \geq (1+\gamma)\mathbb{E}(\chi_M) + \frac{1}{n}\sqrt{2vx} + \frac{1}{n}(2+\gamma^{-1})bcx\right) \leq e^{-x}$$

where $c = \sup_{\substack{u\in S_M \\ \|u\|_n=1}} \|u\|_\infty$ and the variance term $v = \sum_{i=1}^n \sup_{\substack{u\in S_M \\ \|u\|_n=1}} \mathrm{Var}(u_i\varepsilon_i) \leq nc^2\sigma^2$.
Thus

$$\mathbb{P}\left(\chi_M \geq (1+\gamma)\mathbb{E}(\chi_M) + \frac{\sigma}{\sqrt{n}}c\sqrt{2x} + \frac{1}{n}(2+\gamma^{-1})bcx\right) \leq e^{-x}. \tag{1.11}$$

If we compare inequality (1.9) (which corresponds to the case $b=0$ and ε_i Gaussian) with inequality (1.11), we see that the variance term v is much too large. We should have $v = \sup_u \sum_{i=1}^n \mathrm{Var}(u_i\varepsilon_i)$ instead of $v = \sum_{i=1}^n \sup_u \mathrm{Var}(u_i\varepsilon_i)$. With such a refinement, we would obtain here $v \leq n\sigma^2$ instead of $v \leq nc^2\sigma^2$ and the term $\frac{\sigma}{\sqrt{n}}c\sqrt{2x}$ in (1.11) would be replaced by $\frac{\sigma}{\sqrt{n}}\sqrt{2x}$.

Remark 1.5 *Let see which result we could get with the refinement* $v = \sup_u \sum_{i=1}^n Var(u_i\varepsilon_i) \leq n\sigma^2$.
The supremum in (1.10) is achieved with $u = \frac{\varepsilon_M}{\|\varepsilon_M\|_n}$.
Thus, denoting $\Omega_\delta = \left\{\forall J\in M_0;\ \left|\sum_{x_i\in J}\varepsilon_i\right| \leq \delta\sigma^2|J|\right\}$ *and truncating* χ_M *with* $\Omega_\delta \bigcap \left\{\chi_M \geq \frac{\sigma}{\sqrt{n}}\sqrt{2x}\right\}$, *we would get:*

$$\mathbb{P}\left(\chi_M \mathbb{1}_{\Omega_\delta} \geq (1+\gamma)\mathbb{E}(\chi_M) + \frac{\sigma}{\sqrt{n}}\sqrt{2x} + \frac{(2+\gamma^{-1})\delta b}{\sqrt{2}}\frac{\sigma}{\sqrt{n}}\sqrt{x}\right) \leq e^{-x}.$$

As Bousquet's concentration inequality is not convenient for our problem (and its refinement seems difficult to obtain), we build our own concentration inequality. Instead of considering expression (1.10) where χ_M is written as a supremum, we wiew $\chi_M^2 = \|\varepsilon_M\|_n^2$ as a χ^2 like statistic and we write it as a sum of squares (see expression (1.5)). Then we get the following lemma.

Lemma 1.1 *Let* $b\in\mathbb{R}_+$ *and* $\sigma\in\mathbb{R}_+^*$ *such that inequality (1.3) holds.*
Let M_0 *a partition of* $\mathcal{X}$ *and denote* $N_{min} = \inf_{J\in M_0}|J|$.
Let $\delta>0$ *and* $\Omega_\delta = \left\{\forall J\in M_0;\ \left|\sum_{x_i\in J}\varepsilon_i\right| \leq \delta\sigma^2|J|\right\}$
For any partition M *built from* M_0 *and for any* $x>0$

$$\mathbb{P}\left(\chi_M^2 \mathbb{1}_{\Omega_\delta} \geq \frac{\sigma^2}{n}|M| + 4\frac{\sigma^2}{n}(1+b\delta)\sqrt{2|M|x} + 2\frac{\sigma^2}{n}(1+b\delta)x\right) \leq e^{-x}$$

and

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

If $b = 0$, we do not need to truncate χ^2_M with Ω_δ and for any $x > 0$

$$\mathbb{P}\left(\chi^2_M \geq \frac{\sigma^2}{n}|M| + 4\frac{\sigma^2}{n}\sqrt{2|M|x} + 2\frac{\sigma^2}{n}x\right) \leq e^{-x}$$

In lemma 1.1, the $(\varepsilon_i)_{1\leq i\leq n}$ are only supposed to have exponential moments around 0. In this case, by truncating χ^2_M, we get a concentration inequality which differs from inequality (1.8) (corresponding to the Gaussian case) only in the multiplicative constants. The set Ω_δ on which we control the deviations of χ^2_M is very large. More precisely, if N_{min} satisfies $N_{min} \geq 2(k+1)\frac{(1+b\delta)}{\delta^2\sigma^2}\log n$ for some integer k, then

$$\mathbb{P}\left(\Omega_\delta^c\right) \leq \frac{1}{(k+1)}\frac{\delta^2\sigma^2}{(1+b\delta)}\frac{1}{n^k \log n}.$$

In theorem 1.1, we take $k = 2$ so that $\mathbb{P}\left(\Omega_\delta^c\right) = o\left(\frac{1}{n^2}\right)$ when $n \to +\infty$.

The concentration inequalities of the $\left(\chi^2_M\right)_{M\in\mathcal{M}_n}$ and $\left(\langle\varepsilon, s - s_M\rangle_n\right)_{M\in\mathcal{M}_n}$ are the key to determine the adequate form of penalty. $\langle\varepsilon, s - s_M\rangle_n$ is centered and the expectation of χ^2_M is upper bounded by $\frac{\sigma^2}{n}|M|$. The weights $(x_M)_{M\in\mathcal{M}_n}$ satisfying $\sum_{M\in\mathcal{M}_n} e^{-x_M} \leq \Sigma \in \mathbb{R}_+^*$ allow to control χ^2_M and $\langle\varepsilon, s - s_M\rangle_n$ for all $M \in \mathcal{M}_n$ simultaneously. This is the reason why, as told in section 1.3, the right penalty pen is the sum of two terms: one proportional to $\frac{|M|}{n}$ (corresponding to $\mathbb{E}(\chi^2_M)$) and a second depending on x_M.

Remark 1.6 *This lemma is based on Bernstein inequality [19, section 2.2.3]. By truncating all χ^2_M with the set Ω_δ, we get concentration inequalities which remain sharp when summing them over all partitions $M \in \mathcal{M}_n$. In the context of histogram density estimation, Castellan [8] has to control an other χ^2 like statistic. Like here, the main point is to truncate the statistic. While she concludes by applying a Talagrand inequality to the truncated statistic, we use Bernstein inequality.*

1.5 Proof of lemma 1.1

Let M a partition built from M_0 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, $\mathbb{E}\left(Z_J\right) \leq \mathbb{E}\left(\varepsilon_1^2\right) \leq \sigma^2$, and for any $k \geq 2$ we have

$$\begin{aligned}
\mathbb{E}\left(|Z_J|^k\right) &= \frac{1}{|J|^k}\mathbb{E}\left[\left\{\left|\sum_{i\in J}\varepsilon_i\right| \wedge \left(\delta\sigma^2|J|\right)\right\}^{2k}\right] \\
&= \frac{1}{|J|^k}\int_0^{+\infty} 2kx^{2k-1}\mathbb{P}\left(\left|\sum_{i\in J}\varepsilon_i\right| \wedge \left(\delta\sigma^2|J|\right) \geq x\right)dx \\
&= \frac{1}{|J|^k}\int_0^{\delta\sigma^2|J|} 2kx^{2k-1}\mathbb{P}\left(\left|\sum_{i\in J}\varepsilon_i\right| \geq x\right)dx
\end{aligned}$$

We deduce from assumption (1.3) and Cramer-Chernoff method (see remark 1.2) that for any $x > 0$

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

Thus

$$\begin{aligned}
\mathbb{E}\left(|Z_J|^k\right) &\leq \frac{1}{|J|^k}\int_0^{\delta\sigma^2|J|} 2kx^{2k-1}2\exp\left(\frac{-x^2}{2(\sigma^2|J| + bx)}\right)dx \\
&\leq \frac{4k}{|J|^k}\int_0^{+\infty} x^{2k-1}\exp\left(\frac{-x^2}{2\sigma^2|J|(1+b\delta)}\right)dx
\end{aligned}$$

Integrating part by part, we get

$$\mathbb{E}\left(|Z_J|^k\right) \leq \frac{k!}{2}\left(4\sigma^2(1+b\delta)\right)^2\left(2\sigma^2(1+b\delta)\right)^{k-2}$$

Thanks to Bernstein inequality we obtain that for any $x > 0$

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

Since $\frac{1}{n}\sum_{J\in M} Z_J = \chi^2_M$ on the set Ω_δ,

$$\mathbb{P}\left(\chi^2_M \mathbb{1}\!\mathrm{I}_{\Omega_\delta} \geq \frac{\sigma^2}{n}|M| + 4\frac{\sigma^2}{n}(1+b\delta)\sqrt{2|M|x} + 2\frac{\sigma^2}{n}(1+b\delta)x\right) \leq e^{-x}$$

Thanks to assumption (1.3), for any $J \in M_0$, we have

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

Summing these inequalities over $J \in M_0$, we get

$$\begin{aligned}
\mathbb{P}\left(\Omega_\delta^c\right) &\leq 2|M_0|\exp\left(\frac{-\delta^2\sigma^2 N_{min}}{2(1+b\delta)}\right) \\
&\leq 2\frac{n}{N_{min}}\exp\left(\frac{-\delta^2\sigma^2 N_{min}}{2(1+b\delta)}\right)
\end{aligned}$$

1.6 Proof of the theorem

Let $\theta \in (0,1)$ and $K > 2-\theta$.

According to (1.6),

$$(1-\theta)\|s-\hat{s}_{\widehat{M}}\|_n^2 = \Delta_{\widehat{M}} + \inf_{M\in\mathcal{M}_n} R_M \tag{1.12}$$

where

$$\begin{aligned}\Delta_M &= (2-\theta)\|\varepsilon_M\|_n^2 - 2<\varepsilon, s-s_M>_n -\theta\|s-s_M\|_n^2 - \mathrm{pen}(M)\\ R_M &= \|s-s_M\|_n^2 - \|\varepsilon_M\|_n^2 + 2<\varepsilon, s-s_M>_n + \mathrm{pen}(M)\end{aligned}$$

Let denote $\Omega = \left\{\forall J\in M_0;\ \left|\sum_{i\in J}\varepsilon_i\right| \le \frac{\sigma^2}{b}|J|\right\}$

Thanks to lemma 1.1,

$$\mathbb{P}\left(\Omega^c\right) \le 2\frac{n}{N_{min}}\exp\left(\frac{-\sigma^2 N_{min}}{4b^2}\right)$$

and, for any $M\in\mathcal{M}_n$ and any $x>0$,

$$\mathbb{P}\left(\|\varepsilon_M\|_n^2 \mathbb{1}_\Omega \ge \frac{\sigma^2}{n}|M| + 8\frac{\sigma^2}{n}\sqrt{2|M|x} + 4\frac{\sigma^2}{n}x\right) \le e^{-x} \tag{1.13}$$

Thanks to (1.7), we have for any $M\in\mathcal{M}_n$ and any $x>0$,

$$\mathbb{P}\Big(-<\varepsilon, s-s_M>_n \ge \frac{\sigma}{\sqrt{n}}\|s-s_M\|_n\sqrt{2x} + \frac{2Rb}{n}x\Big) \ \le\ e^{-x} \tag{1.14}$$

Setting $x = x_M + \xi$ with $\xi>0$, and summing all inequalities (1.13) and (1.14) with respect to $M\in\mathcal{M}_n$, we derive a set E_ξ such that:

- $\mathbb{P}\left(E_\xi^c\right) \le e^{-\xi}2\Sigma$
- on the set $E_\xi \bigcap \Omega$, for any M,

$$\begin{aligned}\Delta_M \ \le\ & (2-\theta)\frac{\sigma^2}{n}|M| + 8(2-\theta)\frac{\sigma^2}{n}\sqrt{2|M|(x_M+\xi)} + 4(2-\theta)\frac{\sigma^2}{n}(x_M+\xi)\\ & +2\frac{\sigma}{\sqrt{n}}\|s-s_M\|_n\sqrt{2(x_M+\xi)} + \frac{4Rb}{n}(x_M+\xi)\\ & -\theta\|s-s_M\|_n^2 - \mathrm{pen}(M)\end{aligned}$$

Using the two following inequalities

$$2\tfrac{\sigma}{\sqrt{n}}\|s-s_M\|_n\sqrt{2(x_M+\xi)} \le \theta\|s-s_M\|_n^2 + \tfrac{2}{\theta}\tfrac{\sigma^2}{n}(x_M+\xi),$$

$$8(2-\theta)\tfrac{\sigma^2}{n}\sqrt{2|M|(x_M+\xi)} \le 8\sqrt{2}(2-\theta)\tfrac{\sigma^2}{n}\sqrt{|M|x_M} + 4\sqrt{2}(2-\theta)\tfrac{\sigma^2}{n}\left(\eta|M| + \eta^{-1}\xi\right)$$

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

$$\begin{aligned}
\Delta_M &\leq (2-\theta)\frac{\sigma^2}{n}|M| + 8(2-\theta)\frac{\sigma^2}{n}\sqrt{2|M|(x_M+\xi)} \\
&\quad + \left(4(2-\theta)+\frac{2}{\theta}\right)\frac{\sigma^2}{n}(x_M+\xi) + \frac{4Rb}{n}(x_M+\xi) \\
&\quad -\text{pen}(M) \\
&\leq K\frac{\sigma^2}{n}|M| + 8\sqrt{2}(2-\theta)\frac{\sigma^2}{n}\sqrt{|M|x_M} + \left(4(2-\theta)+\frac{2}{\theta}\right)\frac{\sigma^2}{n}x_M + \frac{4Rb}{n}x_M \\
&\quad + \left\{4(2-\theta)\left(1+\frac{8(2-\theta)}{K+\theta-2}\right)+\frac{2}{\theta}\right\}\frac{\sigma^2}{n}\xi + \frac{4Rb}{n}\xi - \text{pen}(M)
\end{aligned}$$

Taking a penalty pen wich compensates for all the other terms in M, i.e.

$$\text{pen}(M) \geq K\frac{\sigma^2}{n}|M| + 8\sqrt{2}(2-\theta)\frac{\sigma^2}{n}\sqrt{|M|x_M} + \left\{\left(4(2-\theta)+\frac{2}{\theta}\right)\frac{\sigma^2}{n} + \frac{4Rb}{n}\right\}x_M$$

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

$$\Delta_{\widehat{M}} \leq \left\{4(2-\theta)\left(1+\frac{8(2-\theta)}{K+\theta-2}\right)+\frac{2}{\theta}\right\}\frac{\sigma^2}{n}\xi + \frac{4Rb}{n}\xi$$

In other words, on the set E_ξ,

$$\Delta_{\widehat{M}}\mathbb{1}_\Omega \leq \left\{4(2-\theta)\left(1+\frac{8(2-\theta)}{K+\theta-2}\right)+\frac{2}{\theta}\right\}\frac{\sigma^2}{n}\xi + \frac{4Rb}{n}\xi$$

Integrating with respect to ξ,

$$\mathbb{E}\left(\Delta_{\widehat{M}}\mathbb{1}_\Omega\right) \leq 2\left\{4(2-\theta)\left(1+\frac{8(2-\theta)}{K+\theta-2}\right)+\frac{2}{\theta}\right\}\frac{\sigma^2}{n}\Sigma + \frac{8Rb}{n}\Sigma \tag{1.15}$$

We are going now to control $\mathbb{E}\left(\inf_M R_M \mathbb{1}_\Omega\right)$.
Thanks to (1.7), for any M and any $x > 0$

$$\mathbb{P}\Big(\langle \varepsilon, s-s_M\rangle_n \geq \frac{\sigma}{\sqrt{n}}\|s-s_M\|_n\sqrt{2x} + \frac{2Rb}{n}x\Big) \leq e^{-x}$$

Thus we derive a set F_ξ such that

- $\mathbb{P}\left(F_\xi^c\right) \leq e^{-\xi}\Sigma$

- on the set F_ξ, for any M,

$$< \varepsilon, s-s_M >_n \leq \frac{\sigma}{\sqrt{n}}\|s-s_M\|_n\sqrt{2\left(x_M+\xi\right)} + \frac{2Rb}{n}\left(x_M+\xi\right)$$

It follows from definition of R_M that on the set F_ξ, for any M,

$$\begin{aligned} R_M &\leq \|s-s_M\|_n^2 + 2\frac{\sigma}{\sqrt{n}}\|s-s_M\|_n\sqrt{2\left(x_M+\xi\right)} + \frac{4Rb}{n}\left(x_M+\xi\right) + \mathrm{pen}(M) \\ &\leq 2\|s-s_M\|_n^2 + 2\frac{\sigma^2}{n}\left(x_M+\xi\right) + \frac{4Rb}{n}\left(x_M+\xi\right) + \mathrm{pen}(M) \\ &\leq 2\|s-s_M\|_n^2 + 2\mathrm{pen}(M) + 2\frac{\sigma^2}{n}\xi + \frac{4Rb}{n}\xi \end{aligned}$$

And

$$\begin{aligned} \mathbb{E}\left(\inf_M R_M \mathbb{1}_\Omega\right) &\leq 2\inf_M\left\{\|s-s_M\|_n^2 + \mathrm{pen}(M)\right\} \\ &\quad + 2\frac{\sigma^2}{n}\Sigma + \frac{4Rb}{n}\Sigma \end{aligned} \tag{1.16}$$

We conclude from (1.12), (1.15) and (1.16) that

$$\begin{aligned} (1-\theta)\mathbb{E}\left(\|s-\hat{s}_{\hat{M}}\|_n^2\mathbb{1}_\Omega\right) &\leq 2\inf_M\left\{\|s-s_M\|_n^2 + \mathrm{pen}(M)\right\} \\ &\quad + \left\{8(2-\theta)\left(1+\frac{8(2-\theta)}{K+\theta-2}\right) + \frac{4}{\theta} + 2\right\}\frac{\sigma^2}{n}\Sigma + \frac{12Rb}{n}\Sigma \end{aligned}$$

It remains to control $\mathbb{E}\left(\|s-\hat{s}_{\hat{M}}\|_n^2\mathbb{1}_{\Omega^c}\right)$, except if $b=0$ in which case it is finished.

$$\begin{aligned} \mathbb{E}\left(\|s-\hat{s}_{\hat{M}}\|_n^2\mathbb{1}_{\Omega^c}\right) &= \mathbb{E}\left(\|s-s_{\widehat{M}}\|_n^2\mathbb{1}_{\Omega^c}\right) + \mathbb{E}\left(\|\varepsilon_{\widehat{M}}\|_n^2\mathbb{1}_{\Omega^c}\right) \\ &\leq \mathbb{E}\left(\|s\|_n^2\mathbb{1}_{\Omega^c}\right) + \mathbb{E}\left(\|\varepsilon_{M_0}\|_n^2\mathbb{1}_{\Omega^c}\right) \\ &\leq R^2\mathbb{P}\left(\Omega^c\right) + \sqrt{\mathbb{E}\left(\|\varepsilon_{M_0}\|_n^4\right)}\sqrt{\mathbb{P}\left(\Omega^c\right)} \end{aligned}$$

By developing $\|\varepsilon_{M_0}\|_n^4$, since $\mathbb{E}\left(\varepsilon_i^2\right) \leq \sigma^2$ and $\mathbb{E}\left(\varepsilon_i^4\right) \leq C(b,\sigma^2)^2$, we get

$$\begin{aligned} \mathbb{E}\left(\|\varepsilon_{M_0}\|_n^4\right) &\leq \frac{\sigma^4|M_0|^2}{n^2} + \frac{C(b,\sigma^2)^2|M_0|}{n^2N_{min}} + \frac{3\sigma^4|M_0|}{n^2} \\ &\leq \frac{\sigma^4}{N_{min}^2} + \frac{C(b,\sigma^2)^2}{nN_{min}^2} + \frac{3\sigma^4}{nN_{min}} \\ &\leq \frac{C'(b,\sigma^2)^2}{N_{min}^2} \end{aligned}$$

and thus

$$\mathbb{E}\left(\|s-\hat{s}_{\hat{M}}\|_n^2\mathbb{1}_{\Omega^c}\right) \leq R^2\mathbb{P}\left(\Omega^c\right) + \frac{C'(b,\sigma^2)}{N_{min}}\sqrt{\mathbb{P}\left(\Omega^c\right)}$$

Let us recall that

$$\mathbb{P}\left(\Omega^c\right) \leq 2\frac{n}{N_{min}}\exp\left(\frac{-\sigma^2N_{min}}{4b^2}\right)$$

For $N_{min} \geq 12\frac{b^2}{\sigma^2}\log n$,

$$\mathbb{P}\left(\Omega^c\right) \leq \frac{\sigma^2}{6b^2}\frac{1}{n^2\log n}$$

and

$$\begin{aligned}\mathbb{E}\left(\|s-\hat{s}_{\hat{M}}\|_n^2 \mathbb{1}_{\Omega^c}\right) &\leq \frac{R^2\sigma^2}{6b^2}\frac{1}{n^2\log n}+\frac{\sigma^3 C'(b,\sigma^2)}{12\sqrt{6}b^3}\frac{1}{n(\log n)^{3/2}} \\ &\leq C''(b,\sigma^2,R)\frac{1}{n(\log n)^{3/2}}\end{aligned}$$

Finally we have the following result:
Taking a penalty which satisfies for all $M \in \mathcal{M}_n$

$$\mathrm{pen}(M) \geq K\frac{\sigma^2}{n}|M| + 8\sqrt{2}(2-\theta)\frac{\sigma^2}{n}\sqrt{|M|x_M} + \left\{\left(4(2-\theta)+\frac{2}{\theta}\right)\frac{\sigma^2}{n}+\frac{4Rb}{n}\right\}x_M$$

if $N_{min} \geq 12\frac{b^2}{\sigma^2}\log n$, we have

$$\begin{aligned}\mathbb{E}\left(\|s-\hat{s}_{\hat{M}}\|_n^2\right) \leq\ & \frac{2}{1-\theta}\inf_M\left\{\|s-s_M\|_n^2+\mathrm{pen}(M)\right\} \\ & +\frac{1}{1-\theta}\left\{8(2-\theta)\left(1+\frac{8(2-\theta)}{K+\theta-2}\right)+\frac{4}{\theta}+2\right\}\frac{\sigma^2}{n}\Sigma+\frac{12}{1-\theta}\frac{Rb}{n}\Sigma \\ & +C''(b,\sigma^2,R)\frac{\mathbb{1}_{b\neq 0}}{n(\log n)^{3/2}}\end{aligned}$$

Chapter 2

Variable Selection through CART

Ce chapitre présente un travail réalisé en collaboration avec Christine Tuleau [23].

Abstract: This paper deals with variable selection in the regression and binary classification frameworks. It proposes an automatic and exhaustive procedure which relies on the use of the CART algorithm and on model selection via penalization. Thanks to CART, we associate to each subset of variables a family of models which rely only on the variables belonging to the considered subset. Then, we determine a penalized criterion which selects a model and the corresponding subset of variables. The proposed penalties lead to oracle type inequalities justifying the performances of the procedure. A simulation study completes the theoretical results.

Keywords: binary classification, CART, model selection, penalization, regression, variable selection

2.1 Introduction

This paper deals with variable selection in regression and classification using the CART algorithm and the model selection approach. In both regression and classification, we have a sample of observations $\mathcal{L} = \{(X_i, Y_i);\ 1 \leq i \leq n\}$ which consists of n independent copies of a pair of random variables (X, Y). 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). We denote by 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.

Let us begin this introduction with some basic ideas focusing on the linear regression model:

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

where $X = (X^1, \ldots, X^p)$ is the vector of the p explanatory variables, Y is the response, and ε is an unobservable noise.

The well-known least squares method provides an estimator of β:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \beta_j X_i^j \right)^2$$

whose components $\hat{\beta}_1, \ldots, \hat{\beta}_p$ are usually almost all non zero. When p is large, the predictor $\hat{s} : x = (x^1, \ldots, x^p) \to \sum_{j=1}^{p} \hat{\beta}_j x^j$ may be difficult to interpret and may not give an accurate prediction. By eliminating some variables X^j, we can make interpretation easier and improve the quadratic risk of $\hat{s}$.

Ridge Regression and Lasso are penalized versions of the least squares method. Ridge Regression (see Hastie [15]) involves a L_2 penalization which produces the shrinkage of $\hat{\beta}$ but does not put any coefficients of $\hat{\beta}$ to zero. Ridge Regression gives a predictor $\hat{s}$ which is better in terms of risk, but which still involves all the initial variables. Lasso (see Tibshirani [24]) adds a L_1 penalty term to the least squares criterion. By this way, Lasso shrinks some coefficients and puts some others to zero. This last method performs variable selection. It improves both prediction accuracy and interpretation. Unfortunately, its implementation needs quadratic programming techniques.

Penalization is not the only way to perform variable selection. For example, we can cite Subset Selection (see Hastie [15]) which provides, for each $k \in \{1, \ldots, p\}$, the best subset of size k, i.e. the subset of k variables which gives smallest residual sum of squares. Then, the final subset is selected by cross validation. This method is exhaustive, and so it is difficult to use it in practice when p is large. Often, Forward or Backward Stepwise Selection (see Hastie [15]) are preferred since they are computationally efficient methods. But, they perhaps eliminate useful variables. 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 [13] which arrises the optimal model, for a moderate number p 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.* [11]. Let $\nu = X\beta$. LARS builds an estimate of ν by successive steps. It proceeds by adding, at each step, one covariate to the model, as Forward Selection. At the begining, $\nu = \nu_0 = 0$. At the first step, LARS finds the predictor X^{j_1} most correlated with the response Y and increases ν_0 in the direction of X^{j_1} until another predictor X^{j_2} has much correlation with the current residuals. So, ν_0 is replaced by ν_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.

We aim at finding a small number of variables among $\Lambda = \{X^1, X^2, \ldots, X^p\}$ which enable to explain or predict the response Y. In contrary to the methods described above, we do not consider linear interactions between $X = (X^1, \ldots, X^p)$ and Y. We propose a non linear variable selection procedure for both the regression framework and the classification one. From a

theoretical point of view, the first step of the procedure consists in applying the CART algorithm to all subsets of variables (but in practice we determine before few data-driven subsets of variables). Then, considering model selection via penalization, we select the subset which minimizes a penalized criterion. In the regression and classification frameworks, we determine a form of penalty which leads to an oracle type inequality.

Let now describe our procedure in detail. We split the sample of observations $\mathcal{L}$ in three subsamples $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$ with size n_1, n_2 and n_3 respectively. In the following, we consider the case $\mathcal{L}_1$ independent of $\mathcal{L}_2$ and the case $\mathcal{L}_1 = \mathcal{L}_2$. We apply the CART algorithm to all 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 by $T_{max}^{(M)}$, is constructed thanks to the class of admissible splits $\mathcal{S}p_M$ which involves only the variables of M. For any $M \in \mathcal{P}(\Lambda)$ and any T pruned subtree of $T_{max}^{(M)}$ (which is denoted $T \preceq T_{max}^{(M)}$), we consider the space $S_{M,T}$ of $\mathbb{L}^2_{\mathcal{Y}}(\mathbb{R}^p, \mu)$ composed of all 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 $\mathcal{L}_1$. For any (M, T), we denote by $\hat{s}_{M,T}$ the $\mathcal{L}_2$ least squares estimator of s over $S_{M,T}$.

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

Then, we select $\widehat{(M,T)}$ by minimizing a $\mathcal{L}_2$ penalized contrast:

$$\widehat{(M,T)} = \arg \min_{(M,T)} \{\gamma_{n_2}(\hat{s}_{M,T}) + \text{pen}(M,T)\}$$

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

Our purpose is to determine a penalty function pen such that the model $\widehat{(M,T)}$ is close to the optimal one, more precisely such that:

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

where l denotes the loss function. The main results of this paper give adequate penalties defined up to two multiplicative constants α and β. Therefore we get a family of estimators $\tilde{s}(\alpha, \beta)$ among which the final estimator is chosen using the test sample $\mathcal{L}_3$.

The described procedure is of course a theoretical one since, when p is too large, it is impossible in practice to take into account all the 2^p sets of variables. A solution consists in 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 by $\mathcal{P}^*$, 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 obtained 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 defined by Breiman *et al.* [7]. 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 sample, 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.* [7]. 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 [14] proved some oracle inequalities for the well-known penalty term $\left(\frac{\alpha|T|}{n}\right)$. They consider two situations that we used too in this article:

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

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

CART is an algorithm which builds a 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 too large, and on the other hand, some really important variables could be hidden by the selected ones.
A second approach is based on the Variable Importance (VI) introduced by Breiman *et al.* [7]. 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.
In this paper, we propose another approach which consists of applying CART to each subset of variables and choosing the set which minimizes an adequate penalized criterion.

2.2.2 The context

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

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

s is the best predictor according to the quadratic contrast γ. Since the distribution P is unknown, s is unknown too. Thus, in the regression and classification frameworks, we use $(X_1,Y_1),...,(X_n,Y_n)$, 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

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

and the loss function l given by (2.2) is the square of the $\mathbb{L}^2(\mathbb{R}^p,\mu)$-distance:

$$l(s,u) = \|s-u\|_\mu^2 \text{ where } \|.\|_\mu \text{ is the } \mathbb{L}^2(\mathbb{R}^p,\mu) - \text{norm}.$$

In this context, each (X_i,Y_i) satisfies

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

where $(\varepsilon_1,...,\varepsilon_n)$ is a sample such that $\mathbb{E}\left[\varepsilon_i|X_i\right] = 0$. In the following, we assume that the variables ε_i have exponential moments around 0 conditionally to X_i.

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

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

As Y and the predictors u take their values in $\{0;1\}$, we have $\gamma(u,(x,y)) = \mathbb{1}_{u(x)\neq y}$ and $l(s,u) = \mathbb{P}(Y \neq u(X)) - \mathbb{P}(Y \neq s(X)) = \mathbb{E}\left[|s(X)-u(X)||2\eta(X)-1|\right]$.

2.3 Regression

Let us consider the regression framework where the ε_i are supposed to have exponential moments around 0 conditionally to X_i. This assumption is equivalent to the existence of two constants $\sigma \in \mathbb{R}_+^*$ and $\rho \in \mathbb{R}_+$ such that

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

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

Remark 2.2 *If $\rho = 0$ in (2.3), the random variables ε_i are said to be sub-Gaussian conditionally to X_i.*

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

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

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

• $(M1)$ case :

Given the collection of models

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

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

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

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:

$$\begin{aligned}\mathbb{E}\left[\|s - \tilde{s}\|_{n_2}^2\ \middle|\mathcal{L}_1\right] \ \leq\ & C_1 \inf_{(M,T)}\left\{\inf_{u \in S_{M,T}}\|s-u\|_\mu^2 + pen(M,T)\right\} + C_2\frac{(\sigma^2+\rho R)}{n_2} \\ & + C(\rho,\sigma,R)\frac{\mathbb{1}_{\rho\neq 0}}{n_2(\log n_2)^{3/2}}\end{aligned}$$

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 is proportional to $\frac{|T|}{n_2}$. It corresponds to the penalty proposed by Breiman *et al.* [7] in their pruning algorithm and validated by Gey and Nédélec [14] for the Gaussian regression case. This proposition validates the CART pruning penalty in a more general regression framework than the Gaussian one. The second term is proportional to $\frac{|M|}{n_2}\left(1 + \log\left(\frac{p}{|M|}\right)\right)$ and is due to the variable selection. It penalizes models that are based on too much explanatory variables.

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

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

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

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

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

Remark 2.3 *In practice, since* ρ, σ *and* R *are unknown, we consider penalties of the form*

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

Remark 2.4 *If* $\rho = 0$, *the form of the penalty is*

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

the oracle bound is

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

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

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

Since σ^2 *is the single unknown parameter which appears in the penalty, instead of putting it in the constants* α' *and* β' *as proposed above, we could in practice replace it by an estimator.*

The $(M1)$ situation permits to work conditionally to the construction of the maximal trees $T_{max}^{(M)}$ and to select a model among a deterministic collection. Finding a convenient penalty to select a model among a deterministic collection is easier, but we may not always have enough

observations to split the training sample $\mathcal{L}$ in three subsamples. This is the reason why we study now the $(M2)$ situation.

• $(M2)$ case :

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

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

and to select one of them.
For technical reasons, we introduce the collection of models

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

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

Proposition 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 $\geq 1 - e^{-\xi}\Sigma - \frac{c}{n_1 \log n_1}\mathbb{1}_{\rho \neq 0}$,

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

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

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

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

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

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.

2.3.2 Final selection

Now, we have a collection of predictors

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

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

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

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

where $\mathcal{K}_M$ is the number of subtrees of $T_{max}^{(M)}$ obtained by the pruning algorithm defined by Breiman *et al.* [7]. $\mathcal{K}_M$ is very smaller than $\left|\left\{T \preceq T_{max}^{(M)}\right\}\right|$. Given the subsample $\mathcal{L}_3$, we choose the final estimator $\tilde{\tilde{s}}$ by minimizing the empirical contrast γ_{n_3} on $\mathcal{G}$.

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

The next result validates this selection.

Proposition 2.3.3

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

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

- *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* $\geq 1 - e^{-\xi} - \epsilon(n_1)$, $\forall \eta \in (0,1)$,

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

Remark 2.6 *If $\rho=0$, by integrating with respect to ξ, we get for the two methods $(M1)$ and $(M2)$ that:*
for any $\eta\in(0,1)$,

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

The conditional risk of the final estimator $\tilde{\tilde{s}}$ with respect to $\|\ \|_{n_3}$ is controlled by the minimum of the errors made by $\tilde{s}(\alpha,\beta)$. Thus the test sample selection does not 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 close to $1/2$, because in this case, the choice between the label 0 and 1 is difficult. If $\mathbb{P}(\eta(x)=1/2)\neq 0$, then the accuracy of the Bayes classifier is not really good and the comparison with s is not relevant. For this reason, we consider the margin condition introduced by Tsybakov [25]:

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

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

• $(M1)$ case :

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

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

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

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

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

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

Like in the regression case, the penalty is the sum of two terms: one proportional to $\frac{|T|}{n_2}$ and another to $\frac{|M|}{n_2}\left(1+\log\left(\frac{p}{|M|}\right)\right)$. For a given value of $|M|$, this result validates the CART pruning algorithm in the binary classification framework.

Remark 2.7 *Unfortunately, the multiplicative factors of the two terms of the penalty depend on the margin h which is difficult to estimate. Thus in practice, we consider penalties of the form*

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

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

• $(M2)$ case :

Like in the regression case, we manage to extend our result for only one subsample $\mathcal{L}_1$. But, while in the $(M1)$ method we work with the expected loss, here we need the expected loss conditionally to $\{X_i,\ (X_i,Y_i)\in\mathcal{L}_1\}$ defined by:

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

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

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

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

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

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

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

2.4.2 Final selection

With the same notations as in the subsection 2.3.2, 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:*

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

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:

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

2.5 Simulations

In this section, we illustrate by an example the theoretical procedure described in the section 2.1, and 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 [7, section 8.6, p.237]), is composed of $p = 10$ explanatory variables $X^1,\ldots,X^{10}$ such that:

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

and of the explained variable Y given by:

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

where the unobservable random variable ε is independent of $X^1,\ldots,X^{10}$ and normally distributed with mean 0 and variance 2.
The variables X^8, X^9 and X^{10} do not appear in the definition of the explained variable Y, they can be considered as observable noise.

The table 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

Table 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 $X = (X^1, \ldots, X^{10})$. The first results are related to the behaviour of the set of variables associated with the estimator $\tilde{s}(\alpha, \beta)$. More precisely, for given values of the parameters α and β of the penalty function, we look at the selected set of variables. Then, the final estimator $\tilde{\tilde{s}}$ and the associated set of variables are computed.

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

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

The behaviour of the set associated with the estimator $\tilde{s}(\alpha, \beta)$, 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 $\tilde{s}(\alpha, \beta)$.

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

Table 2.2: Sets of variables associated with the estimators $\tilde{s}(\alpha, \beta)$.

First, we notice that those results are the expected ones. Then, we see that for a fixed value of the parameter α (respectively β), the increasing of β (resp. α) results in the decreasing of the size of the selected set, as expected. Therefore, this decreasing is related to Breiman's Variable Importance since the explanatory variables disappear according to the Variable Importance Ranking (see 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 of the procedure, whose results are given by the table 2.3, returns the "good" set.

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

Table 2.3: Results of the final model selection.

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 [21] which associates Forward Selection and variable importance (VI) and whose principle is the following one. The sets of $\mathcal{P}^*$ are constructed by invoking and testing the explanatory variables according to Breiman's Variable Importance ranking. More precisely, the first set is composed of the most important variable according to VI. To construct the second one, we consider the two most important variables and we test if the addition of the second most important variable has a significant incremental influence on the response variable. If the influence is significant, the second set of $\mathcal{P}^*$ is composed of the two most importance variables. If not, we drop the second most important variable and we consider the first and the third most important variables and so on. So, at each step, we add an explanatory variable to the preceding set which is less important than the preceding ones.

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

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

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

The first advantage of this family $\mathcal{P}^*$ is that it involves, at the most p sets of variables instead of 2^p. The second one is that, if we perform our procedure restricted to the family $\mathcal{P}^*$, we obtain nearly the same results for the behavior of the set associated with $\tilde{s}$. The

only difference is that, since $\mathcal{P}^*$ does not contain the set of size 10, in the table 2.2, the set $\{1,2,3,4,5,6,7,8,9,10\}$ is replaced by $\{1,2,5,6,3,7,4\}$.

2.6 Appendix

This section presents some lemmas which are needed in the proofs of the propositions of the sections 2.3 and 2.4. The lemmas 2.1 and 2.2 give the expression of the weights needed in the model selection procedures for both the regression and the classification frameworks. The other lemmas are respectively collected in subsection 2.6.1 and subsection 2.6.2 on case they are used in the regression framework or in the classification framework.

Lemma 2.1 *The weights* $x_{M,T} = a|T| + b|M|\left(1+\log\left(\frac{p}{|M|}\right)\right)$, *with* $a > 2\log 2$ *and* $b > 1$ *two absolute constants, satisfy*

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

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

Proof 2.1 *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

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

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

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

Thus the weights $x_{M,T} = a|T|+b|M|\left(1+\log\left(\frac{p}{|M|}\right)\right)$, *with* $a > 2\log 2$ *and* $b > 1$ *two absolute constants, satisfy (2.4).* □

Lemma 2.2 *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

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

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 2.2 *The proof is quite the same as the preceding one.* □

2.6.1 Useful lemmas in the regression framework

The lemmas 2.3 and 2.4 are concentration inequalities for a sum of squared random variables whose Laplace transform are controlled. The lemma 2.3 recalls the concentration inequality obtained in chapter 1, lemma 1.1. It allows to generalize the model selection result of Birgé and Massart [4] for histogram models without assuming the observations to be Gaussian. In lemma 2.3, we consider only partitions m of $\{1,\dots,n\}$ constructed from an initial partition m_0 (i.e. for any element J of m, J is the union of elements of m_0), whereas in lemma 2.4 we consider all partitions m of $\{1,\dots,n\}$.

Lemma 2.3 *Let $\varepsilon_1,\dots,\varepsilon_n$ n independent and identically distributed random variables satisfying:*

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

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

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

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

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

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 2.3 *see chapter 1, lemma 1.1.* □

Lemma 2.4 *Let $\varepsilon_1, \ldots, \varepsilon_n$ n independent and identically distributed random variables satisfying:*

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

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

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

Let $\delta > 0$ and denote $\Omega_\delta = \left\{\forall 1 \leq i \leq n;\ |\varepsilon_i| \leq \delta\sigma^2\right\}$.
Then for any $m \in \mathcal{M}$ and any $x > 0$,

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

and

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

Proof 2.4 *The proof is exactly the same as the preceding one. The only difference is that the set Ω_δ is smaller and $N_{min} = 1$.* □

2.6.2 Useful lemmas in the classification framework

The lemma 2.5 is a concentration inequality due to Talagrand. We give here Bousquet's version [5].

Lemma 2.5 (Talagrand) *Consider n independent and identically distributed random variables $\xi_1, \ldots, \xi_n$ with values in some measurable space Θ. Let $\mathcal{F}$ be some countable class of measurable fonctions $f : \Theta \longrightarrow \mathbb{R}$ such that $|f - \mathbb{E}\left(f(\xi_1)\right)| \leq b$.*
Let $Z = \sup_{f \in \mathcal{F}} \sum_{i=1}^n \left(f(\xi_i) - \mathbb{E}(f(\xi_i))\right)$ and denote by $v = n \sup_{f \in \mathcal{F}} Var\left(f(\xi_1)\right)$.
Then, for any $x > 0$,

$$\mathbb{P}\left(Z - \mathbb{E}(Z) \geq \sqrt{2\left(v + 2b\mathbb{E}(Z)\right)x} + \frac{b}{3}x\right) \leq e^{-x}$$

and thus

$$\mathbb{P}\left(Z \geq (1+\epsilon)\mathbb{E}(Z) + \sqrt{2vx} + b\left(\frac{1}{3} + \epsilon^{-1}\right)x\right) \leq e^{-x} \quad \textit{for any } \epsilon > 0.$$

Proof 2.5 *see [5].* □

The lemma 2.6 allows to pass from local maximal inequalities to global ones.

Lemma 2.6 (A maximal inequality for weighted processes) *Let (S,d) be some countable pseudo-metric space and $u \in S$.*
Let Z be some process indexed by S and assume that $\sup_{v \in B(u,\sigma)} |Z(v) - Z(u)|$ has finite expectation for any positive number σ,
where $B(u,\sigma) = \{v \in S;\ d(u,v) \leq \sigma\}$.
Let Φ be a non negative function on $\mathbb{R}_+$ such that:

1. $\frac{\Phi(x)}{x}$ *is non increasing on* $\mathbb{R}_+^*$

2. $\forall\ \sigma \geq \sigma_* > 0\ \ \mathbb{E}\left(\sup_{v \in B(u,\sigma)} |Z(u) - Z(v)|\right) \leq \Phi(\sigma)$

Then, for any $x \geq \sigma_*$,

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

Proof 2.6 *see [18, lemma 5.1] or [20, lemma A.5.].* □

The lemma 2.7 is a maximal inequality for VC-classes and is due to Massart and Nédélec [20].

Lemma 2.7 (A maximal inequality for VC-classes) *Let* $\xi_1, \ldots, \xi_n$ *n independent and identically distributed random variables with values in* $\mathcal{X}$ *and distribution* P. *Denote* $\mathbb{P}_n$ *the empirical distribution associated with* $(\xi_1, \ldots, \xi_n)$ *and* $\nu_n = \mathbb{P}_n - P$.
Let $\mathcal{B}$ *be some countable VC-class in* $\mathcal{X}$ *with VC-entropy* $H_{\mathcal{B}} = \log\left(|\{B \bigcap \{\xi_1,\ ...,\ \xi_n\};\ B \in \mathcal{B}\}|\right)$ *and assume that* $\sigma > 0$ *is such that*

$$P(B) \leq \sigma^2,\ for\ every\ B \in \mathcal{B}.$$

Let

$$W_{\mathcal{B}} = \sup_{B \in \mathcal{B}} \{\nu_n(B)\}\ or\ \sup_{B \in \mathcal{B}} \{-\nu_n(B)\}$$

Then

$$\mathbb{E}\left(W_{\mathcal{B}}\right) \leq 2\sqrt{3}\sigma\sqrt{\frac{\mathbb{E}(H_{\mathcal{B}})}{n}}$$

provided that $\sigma \geq 4\sqrt{3}\sqrt{\frac{\mathbb{E}(H_{\mathcal{B}})}{n}}$

Proof 2.7 *see [20, lemma A.3.].* □

We use the two inequalities given in lemmas 2.6 and 2.7 to deduce the following lemma, which is useful in the proof of proposition 2.4.1.

Lemma 2.8 *Let* (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) *be independent copies of a pair of random variables* (X, Y), *where* X *takes its values in* $\mathcal{X}$ *with distribution* μ *and* Y *takes its values in* $\{0, 1\}$. *Denote by* P *the distribution of* (X, Y), *by* $\mathbb{P}_n$ *the empirical distribution associated with* $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, *and by* $\nu_n = \mathbb{P}_n - P$.
Let $S = \{\mathbb{1}_A,\ A \in \mathcal{A}\}$, *where* $\mathcal{A}$ *is a VC-class of measurable subsets of* $\mathcal{X}$, *with VC-dimension* V.
Denote by d *the* $\mathbb{L}^2(\mathcal{X}, \mu)$*-distance and, for any* $u \in S$,
$\bar{\gamma}_n(u) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Y_i \neq u(X_i)} - \mathbb{P}\left(Y \neq u(X)\right)$.
Then $\forall u \in S\ \forall x \geq 4\sqrt{3}\sqrt{\frac{\mathbb{E}(H_{\mathcal{A}})}{n}}$

$$\mathbb{E}\left(\sup_{v \in S} \frac{|\bar{\gamma}_n(u) - \bar{\gamma}_n(v)|}{d^2(u,v) + x^2}\right) \leq \frac{32\sqrt{3}}{x}\sqrt{\frac{\mathbb{E}(H_{\mathcal{A}})}{n}}$$

Proof 2.8 *Let $u \in S$.*
Let $\mathcal{A}_+ = \left\{\{(x,y) \in \mathcal{X} \times \{0,1\};\ \mathbb{1}_{y \neq v(x)} < \mathbb{1}_{y \neq u(x)}\};\ v \in S\right\}$
and $\mathcal{A}_- = \left\{\{(x,y) \in \mathcal{X} \times \{0,1\};\ \mathbb{1}_{y \neq v(x)} > \mathbb{1}_{y \neq u(x)}\};\ v \in S\right\}$.
For any $\sigma > 0$, denoting by $B(u,\sigma) = \{v \in S;\ d(u,v) \leq \sigma\}$, we get

$$\mathbb{E}\left[\sup_{v \in B(u,\sigma)} |\bar{\gamma}_n(u) - \bar{\gamma}_n(v)|\right] \leq \mathbb{E}\left[\sup_{\substack{B \in \mathcal{A}_+ \\ P(B) \leq \sigma^2}} \nu_n(B)\right] + \mathbb{E}\left[\sup_{\substack{B \in \mathcal{A}_+ \\ P(B) \leq \sigma^2}} -\nu_n(B)\right]$$
$$+\mathbb{E}\left[\sup_{\substack{B \in \mathcal{A}_- \\ P(B) \leq \sigma^2}} \nu_n(B)\right] + \mathbb{E}\left[\sup_{\substack{B \in \mathcal{A}_- \\ P(B) \leq \sigma^2}} -\nu_n(B)\right]$$

$\mathcal{A}$, $\mathcal{A}_+$ and $\mathcal{A}_-$ are VC-classes with respective VC-entropy $H_{\mathcal{A}}$, $H_{\mathcal{A}_+}$ and $H_{\mathcal{A}_-}$ satisfying $H_{\mathcal{A}_+} \leq H_{\mathcal{A}}$.
Thus, thanks to lemma 2.7,

$$\mathbb{E}\left[\sup_{v \in B(u,\sigma)} |\bar{\gamma}_n(u) - \bar{\gamma}_n(v)|\right] \leq 8\sqrt{3}\sigma\sqrt{\frac{\mathbb{E}(H_{\mathcal{A}})}{n}}$$

provided that $\sigma \geq 4\sqrt{3}\sqrt{\frac{\mathbb{E}(H_{\mathcal{A}})}{n}}$.
We conclude by applying lemma 2.6. □

The following result is a corollary of lemma 2.8 for S the class of all piecewise constant functions mapping $\mathcal{X}$ into $\{0,1\}$ defined on some partition $\mathcal{P}$ of $\mathcal{X}$.

Corollary 2.1 *Let (X_1,Y_1), (X_2,Y_2), ..., (X_n,Y_n) be independent copies of a pair of random variables (X,Y), where X takes its values in $\mathcal{X}$ with distribution μ and Y takes its values in $\{0,1\}$.*
Let S the class of all piecewise constant functions mapping $\mathcal{X}$ into $\{0,1\}$ defined on some partition $\mathcal{P}$ of $\mathcal{X}$.
Denote d the $\mathbb{L}^2(\mathcal{X},\mu)$-distance and, for any $u \in S$,
$\bar{\gamma}_n(u) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Y_i \neq u(X_i)} - \mathbb{P}(Y \neq u(X))$.
Then $\forall u \in S\ \forall x \geq 4\sqrt{3\log 2}\sqrt{\frac{|\mathcal{P}|}{n}}$

$$\mathbb{E}\left(\sup_{v \in S} \frac{|\bar{\gamma}_n(u) - \bar{\gamma}_n(v)|}{d^2(u,v) + x^2}\right) \leq \frac{32\sqrt{3\log 2}}{x}\sqrt{\frac{|\mathcal{P}|}{n}}$$

Proof 2.1 *$S = \{\mathbb{1}_A,\ A \in \mathcal{A}\}$ with $\mathcal{A}$ the class of all sets obtained by putting together some regions of the partition $\mathcal{P}$. We can thus apply lemma 2.8. It only remains to upper bound $H_{\mathcal{A}}$. The classical upper bound is given by Sauer's lemma:*

$$H_{\mathcal{A}} \leq V\left(1 + \log\left(\frac{n}{V}\right)\right) \tag{2.5}$$

where V is the VC-dimension of $\mathcal{A}$.
Here, $\mathcal{A}$ has a VC-dimension $V = |\mathcal{P}|$ and satisfy

$$H_{\mathcal{A}} \leq \log(|\mathcal{A}|) \leq |\mathcal{P}|\log(2) \tag{2.6}$$

In this case, inequality (2.5) is too pessimistic.
We conclude by applying lemma 2.8 and using inequality (2.6).

Thanks to the lemma 2.9, we see that the Hold-Out is an adaptative selection procedure for classification.

Lemma 2.9 (Hold-Out) *Assume that we observe $N+n$ independent random variables with common distribution P depending on some parameter s to be estimated. The first N observations $X' = (X'_1, \ldots, X'_N)$ are used to build some preliminary collection of estimators $(\hat{s}_m)_{m\in\mathcal{M}}$ and we use the remaining observations $(X_1, \ldots, X_n)$ to select some estimator $\hat{s}_{\hat{m}}$ among the collection defined before by minimizing the empirical contrast.*
Suppose that $\mathcal{M}$ is finite with cardinal K.
If there exists a function w such that:

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

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

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

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

Proof 2.9 *see [19, section 8.5, Hold-out for bounded contrasts].* □

2.7 Proofs

2.7.1 Regression

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

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

Following the proof of theorem 2 in [4], we get

$$(1-\theta)\|s-\tilde{s}\|_{n_2}^2 = \Delta_{\widehat{M,T}} + \inf_{(M,T)} R_{M,T} \tag{2.7}$$

where

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

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

For any M, we denote

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

Thanks to lemma 2.3, we get that for any (M,T) and any $x > 0$

$$\begin{aligned} \mathbb{P}\Bigg(\|\varepsilon_{M,T}\|_{n_2}^2 \mathbb{1}_{\Omega_M} \ &\geq \ \frac{\sigma^2}{n_2}|T| + 8\frac{\sigma^2}{n_2}\sqrt{2|T|x} + 4\frac{\sigma^2}{n_2}x \ \Big| \mathcal{L}_1 \text{ and } \{X_i;\ (X_i,Y_i) \in \mathcal{L}_2\} \Bigg) \\ &\leq e^{-x} \end{aligned} \tag{2.8}$$

and

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

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

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

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

$$\begin{aligned} \mathbb{P}\Big(- < \varepsilon, s - s_{M,T} >_{n_2} \ &\geq \ \frac{\sigma}{\sqrt{n_2}}\|s - s_{M,T}\|_{n_2}\sqrt{2x} + \frac{2\rho R}{n_2}x \ \Big| \mathcal{L}_1 \text{ and } \{X_i;\ (X_i,Y_i) \in \mathcal{L}_2\} \Big) \\ &\leq e^{-x} \end{aligned} \tag{2.9}$$

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 2.1, and summing all inequalities (2.8) and (2.9) with respect to (M,T), we derive a set E_ξ such that

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

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

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

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

Using the two following inequalities

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

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

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

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

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

$$\mathrm{pen}(M,T) \geq K\frac{\sigma^2}{n_2}|T| + \left[4(2-\theta)\left(1+\frac{8(2-\theta)}{K+\theta-2}\right)+\frac{2}{\theta}+4\frac{\rho}{\sigma^2}R\right]\frac{\sigma^2}{n_2}x_{M,T} \tag{2.10}$$

we get that, on the set E_ξ

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

Integrating with respect to ξ, we derive

$$\mathbb{E}\left[\Delta_{\widehat{M,T}}\mathbb{1}_\Omega\Big|\mathcal{L}_1\right] \leq 2\left[4(2-\theta)\left(1+\frac{8(2-\theta)}{K+\theta-2}\right)+\frac{2}{\theta}+4\frac{\rho}{\sigma^2}R\right]\frac{\sigma^2}{n_2}\Sigma(a,b) \tag{2.11}$$

We are going now to control $\mathbb{E}\left[\inf_{(M,T)} R_{M,T}\mathbb{1}_\Omega\Big|\mathcal{L}_1\right]$.
In the same way we deduced (2.9) from assumption (2.3), we get that for any (M,T) and any $x>0$

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

Thus we derive a set F_ξ such that

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

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

It follows from definition of $R_{M,T}$ and inequality (2.10) on the penalty that

$$\begin{aligned}\mathbb{E}\left[\inf_{(M,T)} R_{M,T}\mathbb{1}_\Omega\Big|\mathcal{L}_1\right] \leq 2\inf_{(M,T)}\left\{\mathbb{E}\left[\|s-s_{M,T}\|_{n_2}^2\Big|\mathcal{L}_1\right]+\mathrm{pen}(M,T)\right\}\\ +\left(2+4\frac{\rho}{\sigma^2}R\right)\frac{\sigma^2}{n_2}\Sigma(a,b)\end{aligned} \tag{2.12}$$

We conclude from (2.7), (2.11) and (2.12) that

$$(1-\theta)\mathbb{E}\left[\|s-\tilde{s}\|_{n_2}^2 \mathbb{1}_{\Omega}\middle|\mathcal{L}_1\right] \leq \quad 2\inf_{(M,T)}\left\{\mathbb{E}\left[\|s-s_{M,T}\|_{n_2}^2\middle|\mathcal{L}_1\right]+\mathrm{pen}(M,T)\right\} + \left[8(2-\theta)\left(1+\tfrac{8(2-\theta)}{K+\theta-2}\right)+\tfrac{4}{\theta}+2+12\tfrac{\rho}{\sigma^2}R\right]\tfrac{\sigma^2}{n_2}\Sigma(a,b)$$

It remains to control $\mathbb{E}\left[\|s-\tilde{s}\|_{n_2}^2 \mathbb{1}_{\Omega^c}\middle|\mathcal{L}_1\right]$, except if $\rho=0$ in which case it is finished.
After some calculations (see the proof of theorem 1.1 in chapter 1 or in [22] for more details), we get

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

and

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

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

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

Let us recall that

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

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

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

It follows that

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

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

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

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

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

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

Proof of the proposition 2.3.2:
To follow the preceding proof, we have to consider the "deterministic" bigger collection of models:

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

where $\mathcal{M}_{n_1,M}$ denote the set of trees built on the grid $\{X_i;\ (X_i,Y_i)\in\mathcal{L}_1\}$ with splits on the variables in M. By considering this bigger collection of models, we no longer have partitions built from an initial one. So, we use lemma 2.4 (with $\delta = 5\frac{\rho}{\sigma^2}\log\left(\frac{n_1}{p}\right)$) instead of lemma 2.3. The steps of the proof are the same as before. The main difference is that, the quantities are now conditionned by $\{X_i;\ (X_i,Y_i)\in\mathcal{L}_1\}$ instead of $\mathcal{L}_1$ and $\{X_i;\ (X_i,Y_i)\in\mathcal{L}_2\}$. □

Proof of the proposition 2.3.3:
It follows from the definition of $\tilde{\tilde{s}}$ that for any $\tilde{s}(\alpha,\beta)\in\mathcal{G}$

$$\left\|s-\tilde{\tilde{s}}\right\|_{n_3}^2 \leq \|s-\tilde{s}(\alpha,\beta)\|_{n_3}^2 + 2\left\langle\varepsilon,\tilde{\tilde{s}}-\tilde{s}(\alpha,\beta)\right\rangle_{n_3} \tag{2.13}$$

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

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

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

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

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

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

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

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

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

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

It follows from (2.13) that, on the set $\Omega_1 \bigcap E_\xi$, for any $\tilde{s}(\alpha,\beta)\in\mathcal{G}$ and any $\eta\in(0;1)$

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

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

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

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

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

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

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

Thanks to assumption (2.3), we get that

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

with $\epsilon(n_1) = 2n_1 \exp\left(-\frac{9\rho^2 \log^2 n_1}{2(\sigma^2+3\rho^2 \log n_1)}\right) \underset{n_1 \to +\infty}{\longrightarrow} 0$
On the set Ω_2, as for any (M,T), $\|\hat{s}_{M,T}\|_\infty \leq R + 3\rho \log n_1$, we have $M_{\alpha,\beta,\alpha',\beta'} \leq 2(R + 3\rho \log n_1)$. Thus, on the set $\Omega_2 \bigcap E_\xi$, for any $\tilde{s}(\alpha,\beta) \in \mathcal{G}$

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

It follows from (2.13) that, on the set $\Omega_2 \bigcap E_\xi$, for any $\tilde{s}(\alpha,\beta) \in \mathcal{G}$ and any $\eta \in (0;1)$

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

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

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

□

2.7.2 Classification

We denote by

- P the common distribution of the pairs (X_i, Y_i),
- $\mathbb{P}_{n_2}$ the empirical distribution associated with the sample $\mathcal{L}_2$,
- $\nu_{n_2} = \mathbb{P}_{n_2} - P$.

The contrast γ is defined by $\gamma(u,(x,y)) = (y - u(x))^2 = \mathbb{1}_{y \neq u(x)}$ for all $(x,y) \in \mathcal{X} \times \{0,1\}$ and all $u : \mathcal{X} \longrightarrow \{0,1\}$. We denote by

- $\gamma_{n_2}(u) = \mathbb{P}_{n_2}(\gamma(u,.))$,
- $\bar{\gamma}_{n_2}(u) = \nu_{n_2}(\gamma(u,.)) = \gamma_{n_2}(u) - \mathbb{E}[\gamma_{n_2}(u)]$.

The loss function l is defined by $l(s,u) = P(\gamma(u,.)) - P(\gamma(s,.))$. We denote by d the $L^2(\mathcal{X},\mu)$-distance, where μ is the common distribution of the X_i. Since $l(s,u) = \mathbb{E}\left[|s(X) - u(X)||2\eta(X) - 1|\right]$, we have

$$d^2(s,u) \leq \frac{1}{h} l(s,u). \tag{2.14}$$

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

$$l(s,\tilde{s}) \leq l(s,s_{M,T}) + \bar{\gamma}_{n_2}(s_{M,T}) - \bar{\gamma}_{n_2}(\tilde{s}) + \text{pen}(M,T) - \text{pen}(\widehat{M,T})$$

For any $M' \in \mathcal{P}(\Lambda)$ and any $T' \preceq T_{max}^{(M')}$, we consider some positive number $y_{M',T'}$ to be chosen later and we denote:

$$\bullet w_{M',T'}(u) = (d(s, s_{M,T}) + d(s,u))^2 + y_{M',T'}^2 \quad \forall u \in S_{M',T'}$$
$$\bullet V_{M',T'} = \sup_{u \in S_{M',T'}} \frac{|\bar{\gamma}_{n_2}(u) - \bar{\gamma}_{n_2}(s_{M,T})|}{w_{M',T'}(u)}$$

Then

$$l(s,\tilde{s}) \quad \leq \quad l(s, s_{M,T}) + w_{\widehat{M,T}}(\tilde{s}) V_{\widehat{M,T}} + \mathrm{pen}(M,T) - \mathrm{pen}(\widehat{M,T}) \tag{2.15}$$

To control $V_{\widehat{M,T}}$, we consider $V_{M',T'}$ for all possible values of (M',T'), and we upper bound them simultaneously.
Let $M' \in \mathcal{P}(\Lambda)$ and $T' \preceq T_{max}^{(M')}$.

$$V_{M',T'} \quad = \quad \sup_{u \in S_{M',T'}} \left| \nu_{n_2} \left(\frac{\gamma(u,.) - \gamma(s_{M,T},.)}{w_{M',T'}(u)} \right) \right|$$

For any $u \in S_{M',T'}$, which is a finite set,

$$\left| \frac{\gamma(u,.) - \gamma(s_{M,T},.)}{w_{M',T'}(u)} \right| \quad \leq \quad \frac{1}{{y_{M',T'}}^2}$$

and

$$\begin{aligned} Var_P \left(\frac{\gamma(u,.) - \gamma(s_{M,T},.)}{w_{M',T'}(u)} \right) \quad &\leq \quad \frac{d^2(u, s_{M,T})}{\left(d^2(u, s_{M,T}) + y_{M',T'}^2 \right)^2} \\ &\leq \quad \frac{1}{4 y_{M',T'}^2} \end{aligned}$$

Thus, by Talagrand's inequality (lemma 2.5), we get, for any $x > 0$,

$$\mathbb{P} \left(V_{M',T'} \geq \frac{5}{2} \mathbb{E}(V_{M',T'}) + \frac{1}{y_{M',T'}} \sqrt{\frac{x}{2n_2}} + \frac{2}{y_{M',T'}^2} \frac{x}{n_2} \;\Big|\; \mathcal{L}_1 \right) \leq e^{-x}$$

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 2.1, and summing all the obtained inequalities with respect to (M',T'), we derive a set Ω_ξ such that

- $\mathbb{P} \left(\Omega_\xi^c | \mathcal{L}_1 \right) \leq e^{-\xi} \Sigma(a,b)$
- on the set Ω_ξ, whatever $M' \in \mathcal{P}(\Lambda)$ and $T' \preceq T_{max}^{(M')}$,

$$V_{M',T'} < \frac{5}{2} \mathbb{E}(V_{M',T'}) + \frac{1}{y_{M',T'}} \sqrt{\frac{x_{M',T'} + \xi}{2n_2}} + \frac{2}{y_{M',T'}^2} \frac{x_{M',T'} + \xi}{n_2} \tag{2.16}$$

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

It remains to control $\mathbb{E}[V_{M',T'}]$.
We choose $u_{M',T'}$ such that $d(s,u_{M',T'}) = \inf_{u\in S_{M',T'}} d(s,u)$

$$\mathbb{E}[V_{M',T'}] \leq \underbrace{\mathbb{E}\left[\sup_{u\in S_{M',T'}}\left(\frac{|\bar{\gamma}_{n_2}(u)-\bar{\gamma}_{n_2}(u_{M',T'})|}{w_{M',T'}(u)}\right)\right]}_{(E1)} + \underbrace{\mathbb{E}\left[\frac{|\bar{\gamma}_{n_2}(u_{M',T'})-\bar{\gamma}_{n_2}(s_{M,T})|}{\inf_{u\in S_{M',T'}}\left\{w_{M',T'}(u)\right\}}\right]}_{(E2)} \tag{2.17}$$

For any $u \in S_{M',T'}$

$$w_{M',T'}(u) \geq \frac{1}{4}d^2(u,u_{M',T'}) + y^2_{M',T'}$$

Thus, choosing $y_{M',T'} \geq 2\sqrt{3\log 2}\sqrt{\frac{|T'|}{n_2}}$, we get from corollary 2.1:

$$\begin{aligned}(E1) &\leq 4\mathbb{E}\left[\sup_{u\in S_{M',T_{\prime}}}\frac{|\bar{\gamma}_{n_2}(u)-\bar{\gamma}_{n_2}(u_{M',T'})|}{d^2(u,u_{M',T'})+4y^2_{M',T'}}\right]\\ &\leq \frac{64\sqrt{3\log 2}}{y_{M',T'}}\sqrt{\frac{|T'|}{n_2}}\end{aligned} \tag{2.18}$$

For any $u \in S_{M',T'}$

$$w_{M',T'}(u) \geq 2y_{M',T'}d(s_{M,T},u_{M',T'})$$

Thus,

$$\begin{aligned}(E2) &\leq \frac{\sqrt{Var_P\left[\gamma(u_{M',T'},.)-\gamma(s_{M,T},.)\right]}}{2y_{M',T'}d(s_{M,T},u_{M',T'})\sqrt{n_2}}\\ &\leq \frac{1}{2y_{M',T'}\sqrt{n_2}}\end{aligned} \tag{2.19}$$

We deduce from inequalities (2.17),(2.18) and (2.19) that:

$$\mathbb{E}[V_{M',T'}] \leq \frac{64\sqrt{3\log 2}}{y_{M',T'}}\sqrt{\frac{|T'|}{n_2}} + \frac{1}{2y_{M',T'}\sqrt{n_2}}$$

It follows from the inequality (2.16) and from the control of $\mathbb{E}[V_{M',T'}]$ that, on the set Ω_ξ, whatever $M' \in \mathcal{P}(\Lambda)$ and $T' \preceq T^{(M')}_{max}$,

$$\begin{aligned}V_{M',T'} \leq\ & \frac{5}{2y_{M',T'}\sqrt{n_2}}\left(64\sqrt{3\log 2}\sqrt{|T'|}+\frac{1}{2}\right) + \frac{1}{y_{M',T'}}\sqrt{\frac{x_{M',T'}+\xi}{2n_2}}\\ &+\frac{2}{y^2_{M',T'}}\frac{x_{M',T'}+\xi}{n_2}\end{aligned}$$

We choose

$$\begin{aligned} y_{M',T'} &= \frac{5L}{\sqrt{n_2}}\left(64\sqrt{3\log 2}\sqrt{|T'|}+\frac{1}{2}\right) \\ &\quad +2\sqrt{2}L\sqrt{\frac{x_{M',T'}+\xi}{n_2}} \\ &\geq 2\sqrt{3\log 2}\sqrt{\frac{|T'|}{n_2}} \end{aligned}$$

with $L>1$ (to be chosen later),
so that, on the set Ω_ξ, whatever $M' \in \mathcal{P}(\Lambda)$ and $T' \preceq T_{max}^{(M')}$,

$$V_{M',T'} \leq \frac{1}{L} \tag{2.20}$$

It follows from the choice of $y_{M',T'}$ and from inequality (2.14) that,

$$\begin{aligned} w_{\widehat{M,T}}(\tilde{s}) &\leq \frac{2}{h}\left(l(s,s_{M,T})+l(s,\tilde{s})\right)+\frac{100L^2}{n_2}\left(64^2 3\log 2|\hat{T}_{\hat{M}}|+\frac{1}{4}\right) \\ &\quad +16L^2\frac{x_{\widehat{M,T}}+\xi}{n_2} \end{aligned} \tag{2.21}$$

From (2.15), (2.20) and (2.21), we obtain that, on the set Ω_ξ,

$$\begin{aligned} \left(1-\frac{2}{Lh}\right)l\left(s,\tilde{s}\right) &\leq \left(1+\frac{2}{Lh}\right)l\left(s,s_{M,T}\right)+\mathrm{pen}(M,T) \\ &\quad +100.64^2.3.\log 2L\frac{|\hat{T}_{\hat{M}}|}{n_2}+16L\frac{x_{\widehat{M,T}}}{n_2}-\mathrm{pen}(\widehat{M,T}) \\ &\quad +\frac{L\left(25+16\xi\right)}{n_2} \end{aligned}$$

So, we take a penalty function pen such that :

$$\mathrm{pen}(M,T) \geq 100.64^2.3.\log 2L\frac{|T|}{n_2}+16L\frac{x_{M,T}}{n_2}\ \forall\ (M,T)$$

We choose L such that $\frac{1+\frac{2}{Lh}}{1-\frac{2}{Lh}}=C$, ie $L=\frac{2}{h}\frac{C+1}{C-1}$

Then, on the set Ω_ξ,

$$l\left(s,\tilde{s}\right) \leq C\left\{l\left(s,s_{M,T}\right)+\mathrm{pen}(M,T)\right\}+\frac{(C+1)^2}{C-1}\left(25+16\xi\right)\frac{1}{n_2 h}$$

Integrating this inequality with respect to ξ leads to:

$$\mathbb{E}\left[l\left(s,\tilde{s}\right)|\mathcal{L}_1\right] \leq C\left\{l\left(s,s_{M,T}\right)+\mathrm{pen}(M,T)\right\}+\frac{(C+1)^2}{C-1}\left(25+16\Sigma(a,b)\right)\frac{1}{n_2 h}$$

Thus,

$$\mathbb{E}\left[l\left(s,\tilde{s}\right)|\mathcal{L}_1\right] \leq C \inf_{m\in\mathcal{P}(\Lambda)} \inf_{T\preceq T_{max}^{(m)}} \left\{l\left(s,S_{M,T}\right)+\text{pen}(M,T)\right\} + \frac{(C+1)^2}{C-1}\left(25+16\Sigma(a,b)\right)\frac{1}{n_2 h}$$

Finally, we have the following result: Let $C>1$.
If $\text{pen}(M,T) > \frac{C+1}{C-1}\left((2^{15}5^2 3+64a)\log 2\frac{|T|}{n_2 h}+32b\frac{|M|}{n_2 h}\left(1+\log\left(\frac{p}{|M|}\right)\right)\right)$
then

$$\begin{aligned}\mathbb{E}\left(l(s,\tilde{s})\ |\mathcal{L}_1\right) &\leq C \inf_{M\in\mathcal{P}(\Lambda)} \inf_{T\preceq T_{max}^{(M)}} \left\{l(s,S_{M,T})+\text{pen}(M,T)\right\} \\ &\quad + \frac{(C+1)^2}{C-1}\frac{(25+16\Sigma(a,b))}{n_2 h}\end{aligned}$$

□

Proof of the proposition 2.4.2:
This proof is quite simlilar to the preceding one. We just need to replace $w_{M',T'}(u)$ and $V_{M',T'}$ by

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

And like the proof of proposition 2.3.2, we change the conditionnement. □

Proof of the proposition 2.4.3:
This result is obtained by a direct application of the lemma 2.9 which is given in the subsection 2.6. □

Chapter 3

Piecewise polynomial estimation of a regression function

Abstract: We deal with the problem of choosing a piecewise polynomial estimator of a regression function s mapping $[0,1]^p$ into $\mathbb{R}$. In a first part of this paper, we consider some collection of piecewise polynomial models. Each model is defined by a partition M of $[0,1]^p$ and a series of degrees $\underline{\mathtt{d}} = (d_J)_{J\in M} \in \mathbb{N}^M$. We propose a penalized least squares criterion which selects a model whose associated piecewise polynomial estimator performs approximately as well as the best one, in the sense that its quadratic risk is close to the infimum of the risks. The risk bound we provide is non asymptotic. In a second part, we apply this result to tree-structured collections of partitions, which look like the one constructed in the first step of the CART algorithm. And we propose an extension of the CART algorithm to build a piecewise polynomial estimator of a regression function.

Keywords: CART, concentration inequalities, MARS, model selection, oracle inequalities, polynomial estimation, regression

3.1 Introduction

We observe n real variables $(Y_i)_{1\leq i\leq n}$. Each observation Y_i corresponds to the value of an unknown function $s : [0,1]^p \to \mathbb{R}$ at a point $\mathbf{x}_i = (x_i^1, \ldots, x_i^p) \in [0,1]^p$, plus a random perturbation ε_i.

$$Y_i = s(\mathbf{x}_i) + \varepsilon_i, \qquad 1 \leq i \leq n, \tag{3.1}$$

and the variables $(\varepsilon_i)_{1\leq i\leq n}$ are supposed to be centered, independent and identically distributed (i.i.d.). Our aim is to estimate the regression function s from the observations $(Y_i)_{1\leq i\leq n}$.

We look for an estimator $\hat{s}$ of s such that $\hat{s}$ is close to s in the sense that its quadratic risk is small when the number n of observations is large. Given a model S, i.e. a class of functions mapping $[0,1]^p$ into $\mathbb{R}$, a classical method of estimating s consists in minimizing the quadratic contrast (as well called least squares contrast) over the model S. The resulting estimator is denoted by $\hat{s}_S$ and is called the least squares estimator over the model S. Denoting $\|.\|_n$ the Euclidean norm on $\mathbb{R}^n$ scaled by a factor $n^{-1/2}$ and denoting the same way a function u

mapping $[0,1]^p$ into $\mathbb{R}$ and the corresponding vector $(u(\mathbf{x}_i))_{1\leq i\leq n}\in\mathbb{R}^n$, the quadratic risk of $\hat{s}_S$, $\mathbb{E}\left(\|s-\hat{s}_S\|_n^2\right)$, is the sum of two terms, respectively called bias and variance:

$$\mathbb{E}\left(\|s-\hat{s}_S\|_n^2\right)=\inf_{u\in S}\|s-u\|_n^2+\frac{\tau^2}{n}\dim_{\mathbb{R}^n}(S)\ \text{ where } \tau^2=\mathbb{E}(\varepsilon_i^2).$$

So our work consists in finding a model S which makes a good trade-off between the bias $\inf_{u\in S}\|s-u\|_n^2$ and the variance $\frac{\tau^2}{n}\dim_{\mathbb{R}^n}(S)$. This is the subject matter of model selection via penalization developped by Birgé and Massart [3, 4]. In a very general Gaussian framework, they propose a penalized least squares criterion to select a model $S_{\hat{m}}$ among a given countable collection of models $(S_m)_{m\in\mathfrak{M}_n}$. They justify their criterion with an oracle type inequality. Their result is in particular valid when we observe a vector $Y=(Y_i)_{1\leq i\leq n}$ as defined by (3.1) assuming the random perturbations $(\varepsilon_i)_{1\leq i\leq n}$ to be i.i.d. Gaussian centered variables. We adopt the same approach as Birgé and Massart, but we do not assume the $(\varepsilon_i)_{1\leq i\leq n}$ to be Gaussian. We only suppose that the $(\varepsilon_i)_{1\leq i\leq n}$ have exponential moments around 0. As piecewise polynomial functions have good approximation properties, we consider here models of piecewise polynomial functions.

Our estimation procedure is the following one. Let $\mathfrak{M}_n$ be a finite collection of pairs $m=(M,\underline{\mathtt{d}})$ where M is partition of $[0,1]^p$ and $\underline{\mathtt{d}}=(d_J)_{J\in M}\in\mathbb{N}^M$. We consider the corresponding collection of piecewise polynomial models $(S_m)_{m\in\mathfrak{M}_n}$, i.e. for each $m=(M,\underline{\mathtt{d}})\in\mathfrak{M}_n$, S_m is the class of all piecewise polynomial functions defined on the partition M and with various degree d_J on each region $J\in M$. Each S_m is a linear space with finite dimension denoted by D_m. Denoting $\hat{s}_m$ the least squares estimator over S_m, the best model is the one which minimizes $\mathbb{E}\left(\|s-\hat{s}_m\|_n^2\right)$. Unfortunately this model depends on s. The aim of model selection is to propose a data driven criterion, whose minimizer among $(S_m)_{m\in\mathfrak{M}_n}$ is an approximately best model. We select a model $S_{\hat{m}}$ by minimizing over $\mathfrak{M}_n$ a penalized least squares criterion $\mathrm{crit}(m)=\|Y-\hat{s}_m\|_n^2+\mathrm{pen}(m)$:

$$\hat{m}=\arg\min_{m\in\mathfrak{M}_n}\left\{\|Y-\hat{s}_m\|_n^2+\mathrm{pen}(m)\right\}.$$

The estimator $\hat{s}_{\hat{m}}$ is called the penalized least squares estimator (PLSE). The penalty pen has to be chosen such that the model $S_{\hat{m}}$ is close to the optimal model, more precisely such that

$$\mathbb{E}\left(\|s-\hat{s}_{\hat{m}}\|_n^2\right)\leq C\inf_{m\in\mathfrak{M}_n}\mathbb{E}\left(\|s-\hat{s}_m\|_n^2\right). \tag{3.2}$$

The inequality (3.2) will be referred to as the oracle inequality. It bounds the risk of the penalized least squares estimator by the infimum of the risks on a given model up to a constant C. The main result of this paper determines a form of penalty pen which leads to an oracle type inequality.

As noted above, this work is already done in [3, 4] for more general collections of models in a Gaussian framework. In chapter 1 (or equivalently in [22]), we relax the Gaussian assumption in the regression framework (3.1) for collections of histogram models. Here we generalize the result of chapter 1 for collections of piecewise polynomial models. We propose the following form of penalty:

$$\mathrm{pen}(m)=K\frac{\sigma^2}{n}D_m+\kappa_1\frac{\sigma^2}{n}\sqrt{D_m x_m}+\kappa_2\frac{\sigma^2+Rb}{n}x_m$$

with weights $(x_m)_{m\in\mathfrak{M}_n}$ satisfying $\sum_m e^{-x_m} \leq \Sigma$, $\Sigma \in \mathbb{R}_+^*$.
In the same regression framework as ours (with an even milder integrability condition on the $(\varepsilon_i)_{1\leq i\leq n}$), Baraud [1] validates penalties pen(m) proportional to $\frac{D_m}{n}$, like those of Mallows' C_p, for "small" collection of models: the number of models with a given dimension D can be a polynomial function of D but not an exponential function of D. In the Gaussian regression framework, in a situation where the number of models with a given dimension D is $\binom{N}{D}$, where N is a large parameter (N grows to infinity with n), Birgé and Massart [4] prove that Mallows' C_p can lead to terrible results. For "small" collections of piecewise polynomial models we recommend a penalty proportional to $\frac{D_m}{n}$, and for larger collections of models we propose to add a correction via the weights x_m.

After choosing a collection $\mathfrak{M}_n$, in practice we have to find adequate weights x_m, precise the development of pen(m), and calibrate the unknown constants involved in pen(m) according to the data. For $p = 1$ (i.e. $s : [0,1] \longrightarrow \mathbb{R}$), this work is done in [10] by Comte and Rozenholc. They obtain an algorithm which automatically determines a partition, a series of degrees, and computes a piecewise polynomial estimator of s. The form of the penalty used in their algorithm is theoretically validated by [2] when the perturbations are sub-Gaussian. Thanks to our result, it is validated in a more general case.

In a second part of this chapter, we apply our result to tree-structured collections of partitions of $[0,1]^p$ associated with uniform degrees $d \in \{0, 1, \ldots, d_{\max}\}$. We get a penalty pen$(m)$ which is proportional to the dimension D_m of the linear model S_m. Then we propose an extension of the CART algorithm [7] to build a piecewise polynomial estimator of a regression function $s : [0,1]^p \longrightarrow \mathbb{R}$. The procedure selects a partition and a degree which is enforced to be the same on each subregion of the partition. We give the results obtained on a simulated toy example with $p = 1$, but the same procedure works whatever p.
Friedman [12] proposes an other extension of CART called MARS (Multivariate Adaptive Regression Splines) which uses splines of order 1 to build a continous estimator of the regression function s. In contrary to MARS, our piecewise polynomial estimator is not enforced to be continuous at subregion boundaries.

The paper is organized as follows. The section 3.2 presents the statistical framework. We take advantage of this section to define notations needed in the rest of the paper. The section 3.3 gives the main result. In section 3.4, we first recall the different steps of the CART algorithm. Then we apply the result of section 3.3 to tree-structured collections of partitions adapted from CART and suitable for piecewise polynomial estimation. And finally we propose an extension of the CART algorithm to build a piecewise polynomial estimator of a regression function. To get the model selection result of section 3.3, like in chapter 1, we have to control a χ^2 like statistic. The section 3.5 is more technical, it exposes a concentration inequality for the χ^2 like statistic. Sections 3.6 and 3.7 are devoted to the proofs.

3.2 The statistical framework

In this paper, we consider the regression framework already defined in the introduction by (3.1). We assume that the i.i.d. random perturbations $(\varepsilon_i)_{1\leq i\leq n}$ have exponential moments

around 0 and that the points $(\mathbf{x}_i)_{1\leq i\leq n}$ of the design are "well distributed" in $[0,1]^p$.

The first assumption can be expressed by the existence of two non negative constants b and σ such that

$$\forall\ \lambda \in (-1/b, 1/b) \qquad \log \mathbb{E}\left(e^{\lambda\varepsilon_i}\right) \leq \frac{\sigma^2\lambda^2}{2(1-b|\lambda|)} \tag{3.3}$$

σ^2 is necessarly greater than $\mathbb{E}(\varepsilon_i^2)$ and can be chosen as close to $\mathbb{E}(\varepsilon_i^2)$ as we want, but at the price of a larger b. If $b=0$ in (3.3) the variables ε_i are said to be sub-Gaussian.

The second assumption is expressed through the empirical distribution function F_n associated with the points $(\mathbf{x}_i)_{1\leq i\leq n}$:

$$F_n : \mathbf{x} = (x^1, \ldots, x^p) \longrightarrow \frac{1}{n}\sum_{i=1}^{n} 1\!\!1_{\{x_i^j \leq x^j\ \forall 1\leq j\leq p\}} \tag{3.4}$$

We assume that F_n is close to the uniform distribution function on $[0,1]^p$, denoted by F. We measure the distance between F_n and F by $\|F_n - F\|_\infty$ where $\|.\|_\infty$ denotes the sup-norm.

For any $d \in \mathbb{N}$, we denote by $\mathbb{R}_d[\mathbf{x}] = \mathbb{R}_d[x^1, \ldots, x^p]$ the space of polynomial functions on $[0,1]^p$ with degree d.
For a given $m = (M, \underline{\mathtt{d}})$ where M is a partition of $[0,1]^p$ and $\underline{\mathtt{d}} = (d_J)_{J\in M}$ is a series of non negative integers, we denote

- S_m the space of piecewise polynomial functions defined on the partition M and with various degree d_J on each region $J \in M$.

$$S_m = \left\{ \sum_{J\in M} f_J 1\!\!1_J;\ \forall\ J \in M\ f_J \in \mathbb{R}_{d_J}[\mathbf{x}] \right\} \tag{3.5}$$

- $\hat{s}_m$ the least squares estimator of s over S_m.

$$\hat{s}_m = \arg\min_{u\in S_m} \|Y-u\|_n^2$$

  where $\|.\|_n$ denotes the Euclidean norm on $\mathbb{R}^n$ scaled by a factor $n^{-1/2}$ and, for a function u, the vector $(u(\mathbf{x}_i))_{1\leq i\leq n} \in \mathbb{R}^n$ is denoted u too.

S_m is the piecewise polynomial model associated with $m = (M, \underline{\mathtt{d}})$. It is a linear space with finite dimension. We denote by D_m the dimension of S_m, and by $K_m = |M|$ the number of elements of the partition M.

$$D_m = \sum_{J\in M} C(d_J, p) \text{ where } C(d_J,p) = \dim\ \mathbb{R}_{d_J}[x^1,\ldots,x^p] = \sum_{k=0}^{d_J} \binom{p+k-1}{p-1} = \binom{p+d_J}{p}.$$

$\hat{s}_m$ is the piecewise polynomial estimator belonging to S_m which plays the role of benchmark among all the estimators in S_m.

Denoting $s_m = \arg\min_{u\in S_m} \|s-u\|_n^2$, the quadratic risk of the estimator $\hat{s}_m$ is

$$\mathbb{E}\left(\|s-\hat{s}_m\|_n^2\right) = \|s-s_m\|_n^2 + \mathbb{E}(\varepsilon_1^2)\frac{\dim_{\mathbb{R}^n}(S_m)}{n} \leq \|s-s_m\|_n^2 + \sigma^2\frac{D_m}{n}$$

3.3 The main theorem

As told in the introduction, we consider a finite collection $\mathfrak{M}_n$ of pairs $m = (M, \underline{\mathsf{d}})$ with M a partition of $[0,1]^p$ and $\underline{\mathsf{d}} = (d_J)_{J\in M} \in \mathbb{N}^M$, and the corresponding collection of piecewise polynomial models $(S_m)_{m\in\mathfrak{M}_n}$ as defined by (3.5).
We select a model $S_{\hat{m}}$ by minimizing a penalized criterion $\mathrm{crit}(m) = \|Y - \hat{s}_m\|_n^2 + \mathrm{pen}(m)$ over $\mathfrak{M}_n$, i.e.

$$\hat{m} = \arg\min_{m\in\mathfrak{M}_n} \left\{ \|Y - \hat{s}_m\|_n^2 + \mathrm{pen}(m) \right\}.$$

It remains to provide a penalty pen such that the model $\hat{m}$ is close to the optimal model, in the sense that the PLSE $\hat{s}_{\hat{m}}$ satisfies an oracle inequality like (3.2).

For simplicity, let us start with the case $p = 1$ and the equispaced design $x_i = \frac{i}{n}$. Then, the regression function s is a signal mapping $[0,1]$ into $\mathbb{R}$, and the observations Y_i correspond to the values of s at the equispaced points $x_i = \frac{i}{n}$ perturbed by noise. We consider here two standard choices for the collection of models $(S_m)_{m\in\mathfrak{M}_n}$:

(R) The collection of piecewise polynomial models built on regular partitions with various degree belonging to $\{0, 1, \ldots, d\}$ on each subregion. A regular partition is a partition whose elements have all the same measure with respect to the Lebesgue measure μ. We keep only the regular partitions which have less than $\mathcal{K}_n = \left\lfloor \frac{n}{(\log n)^2} \right\rfloor$ elements (in order that each subregion contains at least $\simeq (\log n)^2$ design points). We denote by $\mathcal{M}_n^r$ the family of regular partitions on $[0,1]$ with size smaller than $\mathcal{K}_n = \left\lfloor \frac{n}{(\log n)^2} \right\rfloor$, and by $\mathfrak{M}_n^{r,d}$ the family of pairs $m = (M, \underline{\mathsf{d}})$ with $M \in \mathcal{M}_n^r$ and $\underline{\mathsf{d}} = (d_J)_{J\in M} \in \{0, 1, \ldots, d\}^M$.

(IR) The classical collection of irregular piecewise polynomial models associated with the collection $\mathfrak{M}_n^{ir,d}$ defined below. We consider a sub-series $\left(x_{\phi(j)}\right)_{1\le j\le \mathcal{K}_n}$ of $(x_i)_{1\le i\le n}$ such that: for any $1 \le j < \mathcal{K}_n$, $\phi(j+1) - \phi(j) \ge (\log n)^2$. We define $\mathcal{M}_n^{ir}$ as the family of all partitions with endpoints belonging to the grid $\left(x_{\phi(j)}\right)_{1\le j\le \mathcal{K}_n}$. (By taking this new grid instead of the initial grid $\left(x_i = \frac{i}{n}\right)_{1\le i\le n}$, we ensure that each subregion contains at least $\simeq (\log n)^2$ design points.) Then $\mathfrak{M}_n^{ir,d}$ is the family of pairs $m = (M, \underline{\mathsf{d}})$ with $M \in \mathcal{M}_n^{ir}$ and $\underline{\mathsf{d}} = (d_J)_{J\in M} \in \{0, 1, \ldots, d\}^M$.

In the collection of regular partitions $\mathcal{M}_n^r$, for any $1 \le K \le \mathcal{K}_n$, there is only one partition with size K. Since the degrees of the polynomials are variable in the set $\{0, 1, \ldots, d\}$, there is nevertheless a great number of models $m \in \mathfrak{M}_n^{r,d}$.
If we denote by $\mathfrak{M}_n^r$ the collection of all pairs $m = (M, \underline{\mathsf{d}})$ with M a regular partition of $[0,1]$ and $\underline{\mathsf{d}} = (d_J)_{J\in M} \in \mathbb{N}^M$ with no restriction on the degrees d_J, then we have for any $D \ge 1$ and any $1 \le K \le D$,

$$\begin{aligned} \left|\{m \in \mathfrak{M}_n^r;\ D_m = D \text{ and } K_m = K\}\right| &= \left| \left\{ (d_1, \ldots, d_K) \in \mathbb{N}^K;\ \sum_{j=1}^{K} (d_j + 1) = D \right\} \right| \\ &= \binom{D-1}{K-1} \end{aligned}$$

And,

$$|\{m \in \mathfrak{M}_n^r;\ D_m = D\}| = \sum_{K=1}^{D} \binom{D-1}{K-1} = 2^{D-1}$$

There are too many models in order to apply Baraud's result [1].

Given the collection $\mathfrak{M}_n^{r,d}$, the proposition 3.1 gives a form of penalty function which allows to select a close to optimal model $\hat{m}$, in the sense that the penalized least squares estimator $\hat{s}_{\hat{m}}$ satisfies an oracle type inequality. This proposition is a consequence of theorem 3.1.

Proposition 3.1 *Let $b \in \mathbb{R}_+$ and $\sigma \in \mathbb{R}_+^*$ such that inequality (3.3) holds. Let R a positive constant such that $\|s\|_\infty \leq R$.*
We consider the collection of regular piecewise polynomial models associated with $\mathfrak{M}_n^{r,d}$.
Then there exists a universal constant α_0 such that for any $\alpha \geq \alpha_0$:
taking

$$pen(m) \geq \alpha(d+1)(\sigma^2 + (d+1)Rb)\frac{D_m}{n},$$

we have

$$\mathbb{E}\left(\|s - \hat{s}_{\hat{m}}\|_n^2\right) \leq C_1(\alpha)\inf_m \left\{\|s - s_m\|_n^2 + pen(m)\right\} + \frac{C_2(\alpha, b, \sigma^2, R, d)}{n}.$$

For the collection of regular piecewise polynomial models, the proposition 3.1 recommends penalties $\text{pen}(m)$ proportional to $\frac{D_m}{n}$.

Since we do not know the constants α_0, b, σ^2 and R, we consider in practice penalties of the form $\text{pen}(m) = \alpha' \frac{D_m}{n}$. Then we calibrate the multiplicative factor α' according to the data, using for example Massart's heuristic [19, section 8.5.2, paragraph "Some heuristics"] or Birgé and Massart's rule [4, section 4] (see chapter 4, sections 4.4.1 and 4.5.2).

In contrary to $\mathcal{M}_n^r$, in the collection of irregular partitions $\mathcal{M}_n^{ir}$, there are many partitions with a given size $K \in \{1, \ldots, \mathcal{K}_n\}$. More precisely, $\left|\left\{M \in \mathcal{M}_n^{ir};\ |M| = K\right\}\right| = \binom{\mathcal{K}_n - 1}{K-1}$.

The proposition 3.2 gives a result similar to proposition 3.1 for the collection $\mathfrak{M}_n^{ir,d}$ whose complexity is bigger than those of $\mathfrak{M}_n^{r,d}$. It is also a consequence of theorem 3.1.

Proposition 3.2 *Let $b \in \mathbb{R}_+$ and $\sigma \in \mathbb{R}_+^*$ such that inequality (3.3) holds. Let R a positive constant such that $\|s\|_\infty \leq R$.*
We consider the collection of irregular piecewise polynomial models associated with $\mathfrak{M}_n^{ir,d}$.
Then there exists two universal constants α_0 and β_0 such that for any $\alpha \geq \alpha_0$ and any $\beta \geq \beta_0$:
taking

$$pen(m) \geq (d+1)(\sigma^2 + (d+1)Rb)\left(\alpha\frac{1}{n}\log\binom{\mathcal{K}_n - 1}{K_m - 1} + \beta\frac{D_m}{n}\right),$$

we have

$$\mathbb{E}\left(\|s - \hat{s}_{\hat{m}}\|_n^2\right) \leq C_1(\alpha, \beta)\inf_m \left\{\|s - s_m\|_n^2 + pen(m)\right\} + \frac{C_2(\alpha, \beta, b, \sigma^2, R, d)}{n}.$$

For the collection of irregular piecewise polynomial models, the proposition 3.2 recommends penalties

$$\text{pen}(m) = \alpha' \frac{1}{n} \log \binom{\mathcal{K}_n - 1}{K_m - 1} + \beta' \frac{D_m}{n}. \tag{3.6}$$

We can also choose a more elaborate form of penalty, like those used by Comte and Rozenholc [10]: for any $m = (M, \underline{\mathtt{d}})$ with $K_m = |M| = K$ and $\underline{\mathtt{d}} = (d_j)_{1 \leq j \leq K}$

$$\text{pen}_{CR}(m) = \frac{1}{n} \left[c_1 \log \binom{\mathcal{K}_n - 1}{K - 1} + c_2 (\log K)^{c_3} + c_4 \sum_{j=1}^{K} (d_j + 1) + c_5 \sum_{j=1}^{K} [\log(d_j + 1)]^{c_6} \right].$$

In [10], they theoretically validate this penalty when the perturbations $(\varepsilon_i)_{1 \leq i \leq n}$ are sub-Gaussian thanks to the model selection result of [2]. Proposition 3.2 allows to validate it when the $(\varepsilon_i)_{1 \leq i \leq n}$ are only supposed to have exponential moments around 0.

Using the inequality

$$\binom{\mathcal{K}_n - 1}{K - 1} \leq \left(\frac{e \mathcal{K}_n}{K} \right)^K,$$

we recover the penalty $\text{pen}(m) = \frac{D_m}{n} \left(\alpha'' \log \left(\frac{\mathcal{K}_n}{K_m} \right) + \beta'' \right)$ proposed by Lebarbier [16] for collection of piecewise constant models defined on irregular partitions (which corresponds here to $d = 0$), and used in chapter 4 for collection of piecewise affine models (which corresponds to $d = 1$). We give in chapter 4 two methods to calibrate the constants α'' and β''. The same methods could be used to calibrate the constants α' and β' in (3.6).

Let us now go back to the more general issue of estimating a function $s : [0,1]^p \longrightarrow \mathbb{R}$ when we observe its values (perturbed by noise) at some points $(\mathbf{x}_i)_{1 \leq i \leq n}$ not necessarily equispaced. Thanks to theorem 3.1, we can get results similar to proposition 3.1 and proposition 3.2 when $s : [0,1]^p \longrightarrow \mathbb{R}$ and the design points $\mathbf{x}_i$ are "well distributed" in $[0,1]^p$ (instead of $s : [0,1] \longrightarrow \mathbb{R}$ and $\mathbf{x}_i = \frac{i}{n}$).

For $p = 1$, in the two preceding propositions, we consider partitions of $[0,1]$ composed of segments containing at least $\simeq (\log n)^2$ design points. Now, we consider some collection $\mathcal{M}_n$ of partitions of $[0,1]^p$ composed of hyperrectangle axis oriented regions, which all contain a minimal number of design points.

We consider two types of collection of partitions $\mathcal{M}_n$: "small" collections $\mathcal{M}_n$ in which the number of partitions with a given size K is bounded by a polynomial function of K, and possibly bigger collections $\mathcal{M}_n$ in which all partitions M are built from an initial partition M_0, which is not too fine in the sense that each element of M_0 contains a minimal number of design points. More precisely, we assume that $\mathcal{M}_n$ satisfies one of the two following assumptions:

(A1) there exists $\Gamma \in \mathbb{R}_+^*$ and $a \in \mathbb{N}$ such that: for any $K \geq 1$,

$$|\{M \in \mathcal{M}_n;\ |M| = K\}| \leq \Gamma K^a,$$

and then we denote by $N_{min} = \inf_{M \in \mathcal{M}_n} \inf_{J \in M} |J|$ where $|J| = |\{1 \leq i \leq n;\ \mathbf{x}_i \in J\}|$,

(A2) there exists some partition M_0 such that: for any $M \in \mathcal{M}_n$ and any element J of M, J is the union of elements of M_0,
and then we denote by $N_{min} = \inf_{J \in M_0} |J|$ ($\leq \inf_{M \in \mathcal{M}_n} \inf_{J \in M} |J|$ with equality if $M_0 \in \mathcal{M}_n$).

The assumption (A2) allows to deal with large collections of partitions.

Remark 3.1 *The collection $\mathcal{M}_n^r$ considered in proposition 3.1 satisfies (A1), and the collection $\mathcal{M}_n^{ir}$ considered in proposition 3.2 satisfies (A2).*

Given a collection $\mathfrak{M}_n$ of pairs $m = (M, \underline{\mathtt{d}})$ with $M \in \mathcal{M}_n$ and $\underline{\mathtt{d}} = (d_J)_{J \in M} \in \{0, 1, \ldots, d\}^M$, the theorem 3.1 gives a general form of penalty involving weights $(x_m)_{m \in \mathfrak{M}_n}$, and which leads to an oracle type inequality.
The constants $A(d,p)$, $B(2d,p)$, $C(d,p)$ and $C''(d,p)$ only depend on the maximal degree d of the polynomials and on the number of variables p. $C(d,p) = \binom{p+d}{p}$. $A(d,p)$ and $B(2d,p)$ are defined in lemma 3.4. $C''(d,p) = 1 + \sqrt{2}C(d,p)A(d,p)$ is defined in remark 3.2. The constant $\kappa(a) = 8(2a+5)$ if (A1) and 24 if (A2).

Theorem 3.1 *Let $b \in \mathbb{R}_+$ and $\sigma \in \mathbb{R}_+^*$ such that inequality (3.3) holds.*
Let $d \in \mathbb{N}$, and $\mathcal{M}_n$ a collection of partitions of $[0,1]^p$ composed of hyperrectangle axis oriented regions and satisfying assumption (A1) or (A2) with $N_{min} \geq \kappa(a) A(d,p)^2 \frac{b^2}{\sigma^2} \log n$.
Assume the points $(\mathbf{x}_i)_{1 \leq i \leq n}$ to be distributed such that

$$\|F_n - F\|_\infty \leq \frac{\kappa(a) A(d,p)^2}{2^p (4B(2d,p)+1)} \frac{b^2}{\sigma^2} \frac{\log n}{n}.$$

Let $\mathfrak{M}_n \subset \{m = (M, \underline{\mathtt{d}});\ M \in \mathcal{M}_n$ and $\underline{\mathtt{d}} = (d_J)_{J \in M}$ with $d_J \leq d\}$, and $(x_m)_{m \in \mathfrak{M}_n}$ a family of weights such that for some $\Sigma \in \mathbb{R}_+^*$

$$\sum_{m \in \mathfrak{M}_n} e^{-x_m} \leq \Sigma. \tag{3.7}$$

Assume $\|s\|_\infty \leq R$, with R a positive constant.
Let $\theta \in (0,1)$ and $K > 2 - \theta$ two numbers.
Taking a penalty satisfying

$$\begin{aligned} pen(m) \quad \geq \quad & K \frac{\sigma^2}{n} D_m + 8\sqrt{2}(2-\theta) C(d,p) \frac{\sigma^2}{n} \sqrt{D_m x_m} \\ & + \left[\left(4(2-\theta) C(d,p) + \frac{2}{\theta} \right) \frac{\sigma^2}{n} + 2C''(d,p) \frac{Rb}{n} \right] x_m \end{aligned} \tag{3.8}$$

we have

$$\begin{aligned} \mathbb{E}\left(\|s - \hat{s}_{\hat{m}}\|_n^2 \right) \quad \leq \quad & \frac{2}{1-\theta} \inf_m \left\{ \|s - s_m\|_n^2 + pen(m) \right\} \\ & + \frac{1}{1-\theta} \left[8(2-\theta) C(d,p) \left(1 + 8C(d,p) \frac{(2-\theta)}{K + \theta - 2} \right) + \frac{4}{\theta} + 2 \right] \frac{\sigma^2}{n} \Sigma \\ & + \frac{6C''(d,p)}{1-\theta} \frac{Rb}{n} \Sigma + C(b, \sigma^2, R, d, p) \frac{1\!\!1_{b \neq 0}}{n (\log n)^{3/2}} \end{aligned}$$

where $C(b, \sigma^2, R, d, p)$ is a positive constant which depends only on b, σ^2, R, d and p.

This theorem gives the general form of the penalty function

$$\text{pen}(m) = K\frac{\sigma^2}{n}D_m + \left(\kappa_1(\theta,d,p)\frac{\sigma^2}{n}\sqrt{D_m x_m} + \kappa_2(\theta,d,p)\frac{\sigma^2+Rb}{n}x_m\right)$$

The penalty is the sum of two terms: the first one is proportional to $\frac{D_m}{n}$ and the second one depends on the complexity of the family $\mathfrak{M}_n$ via the weights $(x_m)_{m\in\mathfrak{M}_n}$. For $\theta \in (0,1)$ and $K > 2-\theta$, the PLSE $\hat{s}_{\hat{m}}$ satisfies an oracle type inequality

$$\mathbb{E}\left(\|s-\hat{s}_{\hat{m}}\|_n^2\right) \quad \leq \quad C_1 \inf_m \left\{\|s-s_m\|_n^2 + \text{pen}(m)\right\} + \frac{C_2}{n}$$

where the constant C_1 only depends on θ, whereas C_2 depends on s (via R), on d the maximal degree of the polynomials, on p the number of variables, on the family $\mathfrak{M}_n$ (via Σ) and on the integrability condition of $(\varepsilon_i)_{1\leq i\leq n}$ (via σ^2 and b).

Given a collection $\mathcal{M}_n$ of partitions and a collection $\mathfrak{M}_n$ of pairs $m = (M, \underline{\mathtt{d}})$ with $M \in \mathcal{M}_n$, in order to get a convenient penalty, we have to find weights (x_m) satisfying (3.7) and to replace their expression in (3.8).
Let us denote by

$$\mathcal{C}(K) = \log\left|\{M \in \mathcal{M}_n;\ |M| = K\}\right|.$$

Then the weights $x_m = \mathcal{C}(K_m) + \mathtt{a}D_m$ with $\mathtt{a} > \log 2$ satisfy inequality (3.7). And applying theorem 3.1, we get that for any $\alpha \geq \alpha_0$ and any $\beta \geq \beta_0$ a penalty satisfying:

$$\text{pen}(m) \geq C(d,p)(\sigma^2 + A(d,p)Rb)\left(\alpha\frac{\mathcal{C}(K_m)}{n} + \beta\frac{D_m}{n}\right)$$

leads to an oracle type inequality. This result generalizes propositions 3.1 and 3.2. In practice, we can take

$$\text{pen}(m) = \alpha'\frac{\mathcal{C}(K_m)}{n} + \beta'\frac{D_m}{n}$$

or more elaborate forms.

If we draw a grid in $[0,1]^p$ such that each cell of the grid contains at least N_{min} design points, and if $\mathcal{M}_n$ is the collection of all partitions obtained by removing some axis of the grid, then $\left|\{M \in \mathcal{M}_n;\ |M| = K\}\right| \leq \binom{\mathcal{K}_n}{K}$ where $\mathcal{K}_n$ is the size of the partition associated with the complete grid. In this case we can take penalties of the form

$$\text{pen}(m) = \alpha'\frac{1}{n}\log\binom{\mathcal{K}_n}{K_m} + \beta'\frac{D_m}{n}.$$

3.4 CART extension to piecewise polynomial estimation

In this section, we first give an overview of CART in our regression framework. Then, in view of the form of penalty given by theorem 3.1 for a tree-structured collection of partitions (like the one built in the first step of CART), we explain how CART can be extended to produce a piecewise polynomial estimator.

3.4.1 An overview of CART

CART is an algorithm which builds a piecewise constant estimator of a regression function or a classifier from a training sample $\mathcal{L} = (\mathbf{x}_i, Y_i)_{1 \leq i \leq n}$. We focus here on the regression framework with fixed design points as defined in the introduction by (3.1). In order to produce a piecewise constant estimator of s from the points $(\mathbf{x}_i, Y_i)_{1 \leq i \leq n}$, CART proceeds in three steps.

In the first step, CART builds a fine partition of $[0,1]^p$ by splitting $[0,1]^p$ and the obtained regions in two hyperrectangle axis oriented subregions as long as they contain a minimal number N_{min} of design points. Each split is determined by minimizing the empirical least squares criterion corresponding to histogram estimation. A useful representation of this recursive construction is a tree of maximal depth, denoted by T_{max}. This step is thus called the construction of a maximal tree.
REMARK: The leaves of T_{max} form a fine partition M_0 of $[0,1]^p$. If T is a pruned subtree of T_{max} (i.e. a subtree with same root as T_{max}), then the leaves of T form a partition M_T of $[0,1]^p$ built from M_0. We write $T \preceq T_{max}$ if T is a pruned subtree of T_{max}. At the end of this first step, we have a collection $(M_T)_{T \preceq T_{max}}$ of partitions of $[0,1]^p$ built from M_0.

The second step, called pruning step, consists in selecting some pruned subtrees of T_{max} by minimizing a penalized least squares criterion. In this step, we have in mind that the least squares histogram estimator associated with a too deep tree overfits the training data. This is the reason why Breiman *et al* propose to select the smallest pruned subtree of T_{max} minimizing the criterion

$$\mathrm{crit}_0(T) = R_0(T) + \alpha \frac{|M_T|}{n},$$

where $R_0(T)$ is the mean of the residual squares of the least squares histogram estimator associated with T. They prove the following result:

Proposition 3.3 *There exists*

- $\alpha_0 = 0 < \alpha_1 < \ldots < \alpha_K$ *a finite increasing series of parameters*
- $T_0 = T_{max} \succeq T_1 \succeq \ldots \succeq T_K$ *a nested series of pruned subtrees of* T_{max} *with the tree* T_K *reduced to its root*

such that, for any $\alpha_k \leq \alpha < \alpha_{k+1}$, T_k *is the smallest pruned subtree of* T_{max} *minimizing the criterion* $crit_0(T) = R_0(T) + \alpha \frac{|M_T|}{n}$.

REMARK: At the end of the pruning step, we have $K+1$ candidates left. As $(T_0, T_1, \ldots, T_K)$ is a nested series of $K+1$ different pruned subtrees, there is at most one tree for a given number of leaves. And in general, we observe that many sizes are not represented. Thus K is very smaller than $|\{T \preceq T_{max}\}|$. The pruning step reduces the number of subtrees a lot.

In the last step, a final tree $\hat{T}$ is selected among the nested series $(T_0, T_1, \ldots, T_K)$ via test sample or cross validation. As K is small, we are able to calculate an estimation of the risk of every histogram estimator associated with a tree of the nested series. We choose $\hat{T}$ which has a minimal (or nearly minimal) estimated risk and s is estimated by the least squares histogram estimator associated with $\hat{T}$. In the following, this final selection is made via test

sample.

As noted above, at the end of the first step of the CART algorithm, we have a collection $\mathcal{M}_n^{CART} = (M_T)_{T \preceq T_{max}}$ of partitions built from an initial partition of $[0,1]^p$. In the classical CART algorithm, the same training sample $\mathcal{L}$ is used to build the maximal tree T_{max} and to prune it, and the collection $\mathcal{M}_n^{CART}$ depends on $\mathcal{L}$. Like in [14] and in chapter 2 (or [23]), we consider a slightly modified version of CART, where the first and second steps are respectively done with a sample $\mathcal{L}_1$ and an other sample $\mathcal{L}_2$ independent of $\mathcal{L}_1$. This modified version of CART is easier to study since all steps are done with independent samples. Unfortunately, in practice, when the number of observations is small, we can not use two different samples in the two first steps.

3.4.2 CART extension to piecewise polynomial estimation

The first step of CART gives a collection $(M_T)_{T \preceq T_{max}}$ of partitions built from an initial one. T_{max} is constructed in order that the piecewise constant estimator defined on its leaves best fit the data. Thus the collection $(M_T)_{T \preceq T_{max}}$ is well adapted to piecewise constant estimation. In order to get a collection of partitions adapted to piecewise polynomial estimation, we build for each $0 \leq d \leq d_{max}$ a maximal tree T_{max}^d whose associated piecewise polynomial estimator with degree d best fit the data. The construction of T_{max}^d is modelled on those of T_{max}. The only difference between T_{max}^d and T_{max} is that each split of T_{max}^d is determined by minimizing the empirical least squares criterion corresponding to piecewise polynomial estimation with degree d instead of histogram estimation. For every d, we consider the partitions $(M_T)_{T \preceq T_{max}^d}$ and the models of piecewise polynomial functions defined on M_T with the same degree d on each subregion. These models correspond to pairs $m = (M_T, \underline{\mathtt{d}})$ with $\underline{\mathtt{d}} = (d, \ldots, d)$ and $T \preceq T_{max}^d$. For short, we denote $m = (T, d)$ instead of $m = (M_T, (d, \ldots, d))$.
We get the collection of models

$$\mathfrak{M}_n = \bigcup_{d=0}^{d_{max}} \left\{ (T,d);\ T \preceq T_{max}^d \right\}$$

In order to apply theorem 3.1 to the collection of models $\mathfrak{M}_n$, we have to build T_{max}^d with a sample $\mathcal{L}_1$ and then work with a second independent sample $\mathcal{L}_2$. In this case, the collection $\mathfrak{M}_n$ is deterministic conditionally to $\mathcal{L}_1$.

Thanks to Catalan inequality,

$$\left| \left\{ T \preceq T_{max}^d;\ |M_T| = K \right\} \right| \leq \frac{1}{K} \binom{2(K-1)}{K-1} \leq \frac{2^{2K}}{4K} \ ,$$

Thus, taking $x_m = LD_m = L|M_T|C(d,p)$ with $L > 2\log 2$ for any $m = (T,d) \in \mathfrak{M}_n$, we have

$$\begin{aligned} \sum_{m \in \mathfrak{M}_n} e^{-x_m} &\leq \sum_{d=0}^{d_{max}} \sum_{K=1}^{n} \frac{2^{2K}}{4K} e^{-LKC(d,p)} \leq \sum_{K=1}^{+\infty} \frac{2^{2K}}{4K} \frac{e^{-LK}}{1-e^{-LK}} \\ &\leq \sum_{K=1}^{+\infty} \frac{\left(4e^{-L}\right)^K}{3K} < +\infty \end{aligned}$$

We deduce from theorem 3.1 that taking a penalty

$$\text{pen}(m) = \gamma C(d_{max}, p)(\sigma^2 + bRA(d_{max}, p))\frac{D_m}{n}$$

with γ big enough, we have

$$\begin{aligned}\mathbb{E}\left(\|s - \hat{s}_{\hat{m}}\|_n^2\right) &\leq C_1(\gamma)\inf_m \left\{\|s - s_m\|_n^2 + C(d_{max}, p)(\sigma^2 + bRA(d_{max}, p))\frac{D_m}{n}\right\} \\ &\quad + \frac{C_2(\gamma, b, \sigma^2, R, d_{max}, p)}{n}\end{aligned}$$

As we do not know the constants b, σ^2 and R, we consider the following form of penalty:

$$\text{pen}(m) = \alpha\frac{D_m}{n}$$

The corresponding penalized criterion is

$$\text{crit}(m) = \|Y - \hat{s}_m\|_n^2 + \alpha\frac{D_m}{n}$$

In order to find the model $\hat{m} = \arg\min_{m \in \mathfrak{M}_n} \text{crit}(m)$, we proceed in two steps:

- First, for every degree d, we select a subtree $\hat{T}_d$ of T_{max}^d by minimizing among all $T \preceq T_{max}^d$ the criterion

  $$\text{crit}_d(T) = R_d(T) + \alpha C(d, p)\frac{|M_T|}{n}$$

  where $R_d(T) = \|Y - \hat{s}_{(T,d)}\|_n^2$ measures the lack of fit of the piecewise polynomial estimator with degree d associated with T. $R_d(T)$ is the analogue of $R_0(T)$. Like in the CART pruning criterion $\text{crit}_0(T)$, the additional term is proportional to the number of leaves of T. Thus we can get the tree $\hat{T}_d$ by the analogue of the CART pruning procedure for piecewise polynomial estimation with degree d.

- Then we choose a degree $\hat{d}$ by

  $$\hat{d} = \arg\min_{0 \leq d \leq d_{max}} \left\{R_d(\hat{T}_d) + \alpha C(d, p)\frac{|M_{\hat{T}_d}|}{n}\right\}.$$

Finally $\hat{m} = \left(\hat{T}_{\hat{d}}, \hat{d}\right)$.

The trees $(\hat{T}_d)_{0 \leq d \leq d_{max}}$, the degree $\hat{d}$ and the model $\hat{m}$ depend on the parameter α. Thus we should rather denote them by $(\hat{T}_d(\alpha))_{0 \leq d \leq d_{max}}$, $\hat{d}(\alpha)$ and $\hat{m}(\alpha)$.

By following the proof of Breiman *et al*, we get a result similar to proposition 3.3 for the criterion crit_d. We denote by $(\alpha_{0,d}, \alpha_{1,d}, \ldots, \alpha_{K_d,d})$ and $(T_{0,d}, T_{1,d}, \ldots, T_{K_d,d})$ the corresponding series of parameters and subtrees. For any $\alpha_{k,d} \leq \alpha < \alpha_{k+1,d}$, $T_{k,d}$ is the smallest subtree of T_{max}^d minimizing the criterion $\text{crit}_d(T) = R_d(T) + \alpha C(d, p)\frac{|M_T|}{n}$. In other words, for any $\alpha_{k,d} \leq \alpha < \alpha_{k+1,d}$, $\hat{T}_d(\alpha) = T_{k,d}$. The algorithm which builds these two series is the same as

the CART pruning algorithm except that R_0 is replaced by R_d.

Then, by putting together all series of parameters $(\alpha_{0,d}, \alpha_{1,d}, \ldots, \alpha_{K_d,d})$, $0 \leq d \leq d_{max}$, we get a series $\alpha_0 = 0 < \alpha_1 < \ldots < \alpha_K$ with $K = \sum_{d=0}^{d_{max}} K_d$. For any $1 \leq i \leq K$, we find $\hat{T}_d(\alpha_i)$ in the set of subtrees $\{T_{0,d}, T_{1,d}, \ldots, T_{K_d,d}\}$ and we calculate

$$\hat{d}(\alpha_i) = \arg \min_{0 \leq d \leq d_{max}} \left\{ R_d(\hat{T}_d(\alpha_i)) + \alpha_i C(d,p) \frac{|M_{\hat{T}_d(\alpha_i)}|}{n} \right\}.$$

At this stage, we have a collection of models $\left\{ \hat{m}(\alpha_i) = (\hat{T}_{\hat{d}}(\alpha_i), \hat{d}(\alpha_i));\ 1 \leq i \leq K \right\}$ whose cardinal is smaller than K. (In the following example, $K = 46$ and we obtain 20 different models.) Some models $\hat{m}(\alpha)$ may not belong to the set $\{\hat{m}(\alpha_i);\ 1 \leq i \leq K\}$, but we observe in practice that it is quite a rare situation, and we believe that we can forget them with no regret. (In the following example, only one model $\hat{m}(\alpha)$ corresponding to large α's is forgotten.)

We choose the final model among $\{\hat{m}(\alpha_i);\ 1 \leq i \leq K\}$ via the test sample $\mathcal{L}_3$.

A simulated example:

We simulate $n = 1000$ variables $(Y_i)_{1 \leq i \leq n}$ defined by:

$$Y_i = s(x_i) + \varepsilon_i, \qquad 1 \leq i \leq n,$$

with $(\varepsilon_i)_{1 \leq i \leq n}$ independent Gumbel variables renormalized such that they have mean 0 and variance 1, $x_i = \frac{i}{n}$, and $s : [0,1] \longrightarrow \mathbb{R}$

$$s(x) = \begin{cases} 10 & \text{if } 0 \leq x \leq 0.2, \\ 7.5 - 12.5x & \text{if } 0.2 < x \leq 0.6, \\ 0 & \text{if } 0.6 < x \leq 1. \end{cases}$$

We build $(T^d_{max})_{0 \leq d \leq 3}$ the maximal trees corresponding to the degrees $0 \leq d \leq 3$ by the analogue of the CART growing procedure using a first sample $\mathcal{L}_1$.
Then, for each degree d, we compute the series $(\alpha_{0,d}, \alpha_{1,d}, \ldots, \alpha_{K_d,d})$ and $(T_{0,d}, T_{1,d}, \ldots, T_{K_d,d})$ via the analogue of the CART pruning procedure using a second sample $\mathcal{L}_2$.
We denote by $\alpha_0 = 0 < \alpha_1 < \ldots < \alpha_K$ the series obtained by concatenating the precedings series of parameters. Here we obtain $K = 46$.
And we calculate $\hat{d}(\alpha_i) = \arg \min_{0 \leq d \leq 3} \left\{ R_d(\hat{T}_d(\alpha_i)) + \alpha_i C(d,p) \frac{|M_{\hat{T}_d(\alpha_i)}|}{n} \right\}$. We obtain only 20 different models $\hat{m}(\alpha_i)$. (Only one model with degree 0 and 5 subregions is hidden. It correspond to $\hat{m}(\alpha)$ with $\alpha \simeq 65$.)

The table 3.1 gives the values of α_i, $\hat{d}(\alpha_i)$ and $|\hat{T}_{\hat{d}}(\alpha_i)|$ for $19 \leq i \leq 38$.
Moreover

- for any $1.1420 \leq \alpha < 1.1607$, the selected model $\hat{m}(\alpha)$ is always the same model with degree $d = 2$ associated with a partition with 7 subregions,
- for any $1.1607 \leq \alpha \leq 1.6816$, the selected model $\hat{m}(\alpha)$ is always the same model with degree $d = 2$ associated with a partition with 6 subregions,

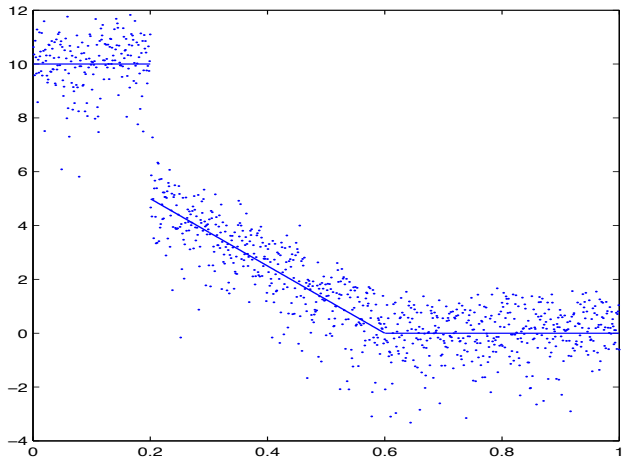

Figure 3.1: plot of s

α_i	1.0624	1.1420	1.1463	1.1607	1.1762	1.2975	1.3026	1.5743	1.6816	1.9360
$\hat{d}(\alpha_i)$	2	2	2	2	2	2	2	2	2	3
$\lvert\hat{T}_{\hat{d}}(\alpha_i)\rvert$	9	7	7	6	6	6	6	6	6	3
α_i	2.0420	2.0603	2.1260	2.8613	2.9912	3.7294	6.9802	9.0454	9.8319	36.5189
$\hat{d}(\alpha_i)$	3	3	3	2	2	2	1	1	1	1
$\lvert\hat{T}_{\hat{d}}(\alpha_i)\rvert$	3	3	3	3	3	3	3	3	3	3

Table 3.1: Results of the model selection procedure

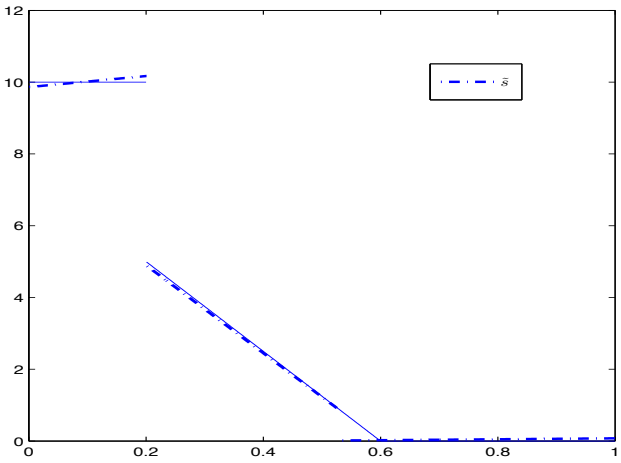

Figure 3.2: plot of $\tilde{s}$ and s

- for any $1.6816 < \alpha < 1.9360$, $\hat{m}(\alpha) = \hat{m}(1.6816)$ or $\hat{m}(1.9360)$,
- for any $1.9360 \leq \alpha \leq 2.1260$, the selected model $\hat{m}(\alpha)$ is always the same model with degree $d = 3$ associated with a partition with 3 subregions, whose limits are 0.2010 and 0.5810,
- for any $2.1260 < \alpha < 2.8613$, $\hat{m}(\alpha) = \hat{m}(2.1260)$ or $\hat{m}(2.8613)$,
- for any $2.8613 \leq \alpha \leq 3.7294$, the selected model $\hat{m}(\alpha)$ is always the same model with degree $d = 2$ associated with a partition with 3 subregions, whose limits are 0.2010 and 0.5810,
- for any $3.7294 < \alpha < 6.9802$, $\hat{m}(\alpha) = \hat{m}(3.7294)$ or $\hat{m}(6.9802)$,
- for any $6.9802 \leq \alpha \leq 36.5189$ (and even for $6 \leq \alpha \leq 50$), the selected model $\hat{m}(\alpha)$ is always the same model with degree $d = 1$ associated with a partition with 3 subregions, whose limits are 0.2010 and 0.5340.

Using a test sample $\mathcal{L}_3$, we finally select the model with degree 1 and 3 subregions, whose limits are 0.2010 and 0.5340. The corresponding estimator $\tilde{s}$ is represented on figure 3.2. His $\mathcal{L}_3-$empirical risk is 1.0653. We get the right number of subregions and the selected degree (which is enforced to be the same on each subregion) equals the maximal degree of s.

In practice, when the number of data is not big enough, we can not split the data in 3 samples $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$. We try the same procedure as before by using the same sample $\mathcal{L}_1 = \mathcal{L}_2$ for the construction of the maximal trees $(T^d_{max})_{0 \leq d \leq 3}$ and for their pruning. We get the following results.
By putting together the series of parameters obtained by pruning all T^d_{max}, we get a series $(\alpha_i)_{0 \leq i \leq K}$ with size $K = 42$. We obtain only 12 different models $\hat{m}(\alpha_i)$ and for any $\alpha > 0$ $\hat{m}(\alpha)$ belongs to the set $\{\hat{m}(\alpha_i);\ 1 \leq i \leq K\}$.

α_i	2.5740	2.5830	2.6721	2.8321	3.0063	3.1198	3.2613	3.3234	3.3651	3.5246
$\hat{d}(\alpha_i)$	0	0	0	0	0	0	0	0	0	0
$\lvert\hat{T}_{\hat{d}}(\alpha_i)\rvert$	13	13	13	13	13	13	12	12	12	12
α_i	3.5625	4.0520	5.1410	8.0859	8.4452	10.4688	11.0017	12.7039	25.1121	27.8815
$\hat{d}(\alpha_i)$	0	0	0	0	1	1	1	1	1	1
$\lvert\hat{T}_{\hat{d}}(\alpha_i)\rvert$	11	10	10	10	3	3	3	3	3	3

Table 3.2: Results of the model selection procedure using a single sample $\mathcal{L}_1 = \mathcal{L}_2$

We write in table 3.2 a part of the series of α_i, $\hat{d}(\alpha_i)$ and $|\hat{T}_{\hat{d}}(\alpha_i)|$. We see on this table that the model with degree 1 and 3 subregions corresponds to larger values of α when the maximal trees (and thus the collection of models $\mathfrak{M}_n$) are constructed with the same sample as well used for the pruning step. When the collection of models $\mathfrak{M}_n$ is random, the number of potential models is bigger and the penalty has to be bigger too. From a theoretical point of view, we have to consider a deterministic collection $\mathfrak{M}_n^0$ which contains $\mathfrak{M}_n$. The weights $(x_m)_{m\in\mathfrak{M}_n^0}$ and thus the penalty will be larger.

In addition to the informations given in table 3.2, we have:

- for any $2.5740 \leq \alpha < 3.2613$, the selected model $\hat{m}(\alpha)$ is always the same model with degree $d = 0$ associated with a partition with 13 subregions,
- for any $3.2613 \leq \alpha < 3.5625$, the selected model $\hat{m}(\alpha)$ is always the same model with degree $d = 0$ associated with a partition with 12 subregions,
- for any $3.5625 \leq \alpha < 4.0520$, the selected model $\hat{m}(\alpha)$ is always the same model with degree $d = 0$ associated with a partition with 11 subregions,
- for any $4.0520 \leq \alpha \leq 8.0859$, the selected model $\hat{m}(\alpha)$ is always the same model with degree $d = 0$ associated with a partition with 10 subregions,
- for any $8.0859 < \alpha < 8.4452$, $\hat{m}(\alpha) = \hat{m}(8.0859)$ or $\hat{m}(8.4452)$,
- for any $8.4452 \leq \alpha \leq 27.8815$ (and even for $8.1 \leq \alpha \leq 53$), the selected model $\hat{m}(\alpha)$ is always the same model with degree $d = 1$ associated with a partition with 3 subregions, whose limits are 0.2010 and 0.5780,
- for larger α, the selected model $\hat{m}(\alpha)$ correspond to the degree $d = 0$ and partitions of size 4, 3, 2 and 1.

By using a test sample $\mathcal{L}_3$, we finally selects the model with degree 1 and 3 subregions, whose limits are 0.2010 and 0.5780. The corresponding estimator $\tilde{s}_2$ is represented on figure 3.3. His $\mathcal{L}_3-$empirical risk is 1.0569. Thus $\tilde{s}_2$ is a bit better than $\tilde{s}$. In practice, we can make this last step of final selection by cross validation instead of test sample.

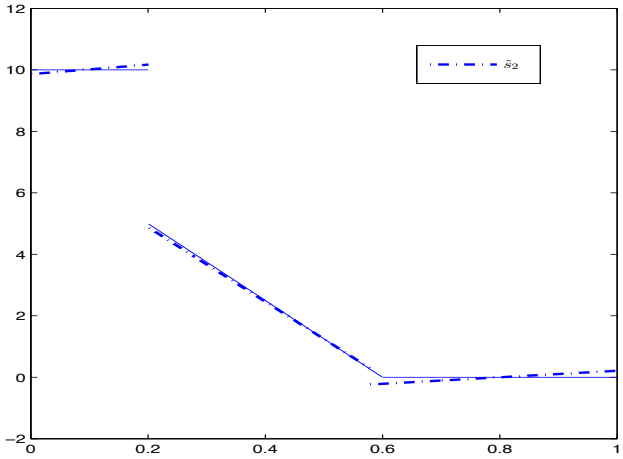

Figure 3.3: plot of $\tilde{s}_2$ and s

3.4.3 Some comments on MARS

The algorithm called MARS (Multivariate Adaptive Regression Splines), proposed by Friedman [12], as well builds a piecewise polynomial estimator of a regression function. This algorithm can be viewed as an other extension of CART, which no longer relies on tree-structured partitions (see Appendix). The main difference with our approach is that the estimator obtained with MARS is continuous with continuous first derivatives.

The first step of MARS is a forward stepwise procedure which looks like the CART growing procedure. It produces a large family of tensor product spline basis functions: $\{B_1, B_2, \ldots, B_N\}$, with $B_1 = 1$ and

$$B_i(\mathbf{x}) = \Pi_{k=1}^{K_i} \left[s_{i,k} \left(x^{v(i,k)} - t_{i,k} \right) \right]_+$$

where $s_{i,k} = \pm$, $v(i,k) \in \{1, 2, \ldots, p\}$ is the number of a variable and $t_{i,k} \in \mathbb{R}$ is a knot location.

The second step of MARS is a backward stepwise procedure which looks like the CART pruning procedure. It removes basis functions that no longer contribute sufficiently to the accuracy of the fit. In the MARS approach, a model S_m is the linear space spanned by a subset $m \subset \{B_1, B_2, \ldots, B_N\}$ which contains B_1. The second step of MARS selects a model $\hat{m}$ by minimizing the criterion

$$GCV(m) = \|Y - \hat{s}_m\|_n^2 \left(1 - \frac{D_m + c|m|}{n} \right)^{-2}$$

where $\hat{s}_m$ is the least squares estimator of s over S_m, $D_m = \dim(S_m)$ and $|m|$ is the number of functions in the set m. Friedman recommends $c = 3$ in the general case and $c = 2$ in the additive case, which corresponds to the case where all basis functions B_i are enforced to have

only $K_i = 1$ term. When n grows to $+\infty$, $\left(1 - \frac{D_m + c|m|}{n}\right)^{-2} \sim 1 + 2\frac{D_m + c|m|}{n}$. Thus

$$\begin{aligned} GCV(m) &\simeq \|Y - \hat{s}_m\|_n^2 + 2\|Y - \hat{s}_m\|_n^2 \frac{D_m + c|m|}{n} \\ &\simeq \|Y - \hat{s}_m\|_n^2 + 2\hat{\tau}^2 \frac{D_m + c|m|}{n} \end{aligned}$$

where $\hat{\tau}^2 = \|Y - \hat{s}_m\|_n^2$ is viewed as an estimator of the variance τ^2. Thus the criterion $GCV(m)$ is close to a penalized least squares criterion with a penalty term

$$\mathrm{pen}_{GCV}(m) = 2\hat{\tau}^2 \frac{D_m + c|m|}{n}$$

For piecewise polynomial models of the form (3.5), we recommend a penalty

$$\mathrm{pen}(m) = K\frac{\sigma^2}{n} D_m + \left(\kappa_1(\theta, d, p)\frac{\sigma^2}{n}\sqrt{D_m x_m} + \kappa_2(\theta, d, p)\frac{\sigma^2 + Rb}{n} x_m\right) \tag{3.9}$$

with x_m such that

$$\sum_m e^{-x_m} \leq \Sigma$$

Theorem 3.1 is not valid for the models S_m involved in MARS. Since Birgé and Massart [4] recommend the same form of penalty as (3.9) in a Gaussian framework for any countable collection of models, we believe that a convenient penalty can be obtained here again by choosing convenient weights and replacing them in expression (3.9).

If there is only one covariable x (i.e. $p = 1$), then the first step of MARS gives a family of basis functions of the form:

$$\mathcal{F}_1 = \{1, (x - t_1)_+, (t_1 - x)_+, \ldots, (x - t_K)_+, (t_K - x)_+\}$$

The linear space spanned by $\mathcal{F}_1$ is the space of all piecewise affine functions defined on the partition with subregion boundaries located at $t_1, \ldots, t_K$, which are continuous everywhere. We denote by m every subset of $\mathcal{F}_1$ which contains the function 1, and by S_m the linear space spanned by m.

Lemma 3.1 *Let* $\mathcal{G}_1 = \{m \subset \mathcal{F}_1;\ 1 \in m\}$.

$$|\{m \in \mathcal{G}_1;\ D_m = N \text{ and } |m| = N\}| = \begin{cases} 2^{N-1}\binom{K}{N-1} & \text{if } N = 1 \text{ or } 2, \\ 2^{N-1}\binom{K}{N-1} + 2^{N-3}(N-2)\binom{K}{N-2} & \text{if } 3 \leq N \leq K+1, \\ K2^{K-1} & \text{if } N = K+2. \end{cases}$$

For any $4 \leq D \leq K + 2$ *and any* $D + 1 \leq N \leq 2D - 3$,

$$|\{m \in \mathcal{G}_1;\ D_m = D \text{ and } |m| = N\}| = 2^{2D-N-3}\binom{K}{D-2}\binom{D-2}{2D-N-3}.$$

And for any $1 \leq D \leq K+2$,

$$|\{m \in \mathcal{G}_1;\ D_m = D\}| \ \leq\ 3^{D-1}\binom{K+1}{D-1}.$$

It follows from lemma 3.1 that we can take $x_m = 2|m| + D_m \log\frac{K+2}{D_m}$ or $x_m = D_m\left(2.1 + \log\frac{K+2}{D_m}\right)$. Thus penalties of the form $\text{pen}(m) = \frac{\alpha D_m \log\frac{K+2}{D_m} + \beta|m|}{n}$ or $\text{pen}(m) = \frac{\alpha D_m\left(\log\frac{K+2}{D_m}+\beta\right)}{n}$ are convenient for the second step of MARS. Up to the factor $\log\frac{K+2}{D_m}$, the proposed penalty $\text{pen}(m)$ has the same form as $\text{pen}_{GCV}(m)$.

If we consider p variables $\mathbf{x} = (x^1, \ldots, x^p)$ and if, during the first step of MARS, B_1 is the single function admissible for splitting, then the first step of MARS gives a family of basis functions of the form:

$$\begin{aligned}\mathcal{F}_p^{ad} = \quad & \{1, (x^1 - t_1^1)_+, (t_1^1 - x^1)_+, \ldots, (x^1 - t_{K_1}^1)_+, (t_{K_1}^1 - x^1)_+, \\ & \ldots, (x^p - t_1^p)_+, (t_1^p - x^p)_+, \ldots, (x^p - t_{K_p}^p)_+, (t_{K_p}^p - x^p)_+\}\end{aligned}$$

and we derive from lemma 3.1 the following lemma:

Lemma 3.2 *Let* $\mathcal{G}_p^{ad} = \left\{m \subset \mathcal{F}_p^{ad};\ 1 \in m\right\}$.

$$\begin{aligned}|\{m \in \mathcal{G}_p^{ad};\ D_m = D\}| \ &\leq \sum_{\substack{(D_1,\ldots,D_p)\\ \sum_i D_i - (p-1) = D}} \Pi_{i=1}^p 3^{D_i-1}\binom{K_i+1}{D_i-1} \\ &\leq\ 3^{D-1}\binom{D_{max}-1}{D-1}\end{aligned}$$

where $D_{max} = (\sum_{i=1}^p K_i) + p + 1 = dim(\mathcal{F}_p^{ad})$.

Thus we suggest a penalty of the form $\text{pen}(m) = \frac{\alpha D_m\left(\log\frac{K+2}{D_m}+\beta\right)}{n}$.

3.5 The key to determine an adequate form of penalty: a concentration inequality for a χ^2 like statistic

This section is more technical. First we recall an expression of $\|s - \hat{s}_{\hat{m}}\|_n^2$, already used in [4] and in chapter 1, which allows us to see that the penalty $\text{pen}(m)$ has to compensate the deviation of a χ^2 like statistic, denoted χ_m^2, in order that the PLSE $\hat{s}_{\hat{m}}$ satisfies an oracle type inequality. Then, in lemma 3.3, we give a concentration inequality for χ_m^2. This concentration inequality is the main point of the proof of theorem 3.1, the remaining of the proof only consists in technical details.

As noted in (chapter 1, section 1.4), we get from the definition of $\hat{m}$ that, for any $\theta \in (0,1)$,

$$\begin{aligned}(1-\theta)\|s - \hat{s}_{\hat{m}}\|_n^2 \ &= \ (2-\theta)\|\varepsilon_{\hat{m}}\|_n^2 - 2\langle\varepsilon, s - s_{\hat{m}}\rangle_n - \theta\|s - s_{\hat{m}}\|_n^2 - \text{pen}(\hat{m}) \\ &\quad + \inf_{m \in \mathfrak{M}_n}\left\{\|s - s_m\|_n^2 - \|\varepsilon_m\|_n^2 + 2\langle\varepsilon, s - s_m\rangle_n + \text{pen}(m)\right\}\end{aligned} \tag{3.10}$$

where, for any model m, $s_m = \arg\min_{u \in S_m} \|s-u\|_n^2$ and $\varepsilon_m = \arg\min_{u \in S_m} \|\varepsilon - u\|_n^2$.
To get an oracle type inequality, the penalty $\text{pen}(m)$ has to compensate, for all model $m = (M, \underline{\mathrm{d}}) \in \mathfrak{M}_n$ simultaneously, the deviations of the statistics

- $\chi_m^2 = \|\varepsilon_m\|_n^2 = \sum_{J \in M} \|\Pi_{J,d_J}\varepsilon\|_n^2$
  where Π_{J,d_J} denotes the orthogonal projection of $\mathbb{R}^n$ on the space corresponding to polynomial functions with degree smaller than d_J and support in J.
- $\langle \varepsilon, s - s_m \rangle_n$

Thanks to assumption (3.3), it is easy to obtain the following concentration inequality for $\langle \varepsilon, s - s_m \rangle_n$:
for any $x > 0$,

$$\mathbb{P}\left(\pm < \varepsilon, s - s_m >_n \geq \frac{\sigma}{\sqrt{n}} \|s - s_m\|_n \sqrt{2x} + \frac{b}{n}\left(\max_{1 \leq i \leq n} |s(\mathbf{x}_i) - s_m(\mathbf{x}_i)| \right) x \right) \quad \leq \quad e^{-x}.$$

The main point is the study of the deviations of the χ^2 like statistic χ_m^2 around its expectation. The following lemma gives a concentration inequality for χ_m^2.

Lemma 3.3 *Let $b \in \mathbb{R}_+$ and $\sigma \in \mathbb{R}_+^*$ such that inequality (3.3) holds.*
Let $d \in \mathbb{N}$, and $\mathcal{M}_n$ a collection of partitions of $[0,1]^p$ composed of hyperrectangle axis oriented regions and satisfying assumption (A1) or (A2). We denote by $N_{min} = \inf_{M \in \mathcal{M}_n} \inf_{J \in M} |J|$.
Let $\delta > 0$ and $\Omega_\delta = \bigcap_{M \in \mathcal{M}_n} \left\{ \forall J \in M;\ \|\Pi_{J,d}\varepsilon\|_n \leq \delta \sigma^2 \sqrt{\frac{|J|}{n}} \right\}$
If the points $(\mathbf{x}_i)_{1 \leq i \leq n}$ are distributed such that

$$\|F_n - F\|_\infty \leq \frac{\inf_{M \in \mathcal{M}_n} \inf_{J \in M} \mu(J)}{2^{p+2} B(2d,p)},$$

then for any partition $M \in \mathcal{M}_n$, for any $(d_J)_{J \in M} \in \mathbb{N}^M$ with all $d_J \leq d$, and for any $x > 0$

$$\mathbb{P}\left(\chi_m^2 \mathbb{1}_{\Omega_\delta} \geq \frac{\sigma^2}{n} D_m + 4C(d,p)\frac{\sigma^2}{n}(1 + b\delta C'(d,p))\sqrt{2D_m x} + 2C(d,p)\frac{\sigma^2}{n}(1 + b\delta C'(d,p))x \right) \leq e^{-x}$$

and

$$\mathbb{P}\left(\Omega_\delta^c\right) \leq \begin{cases} 2\Gamma C(d,p)\left(\frac{n}{N_{min}}\right)^{a+1} \exp\left(\frac{-\delta^2\sigma^2 N_{min}}{2C(d,p)(1+b\delta C'(d,p))}\right) & \text{if (A1) is satisfyed,} \\ 2C(d,p)\frac{n}{N_{min}} \exp\left(\frac{-\delta^2\sigma^2 N_{min}}{2C(d,p)(1+b\delta C'(d,p))}\right) & \text{if (A2) is satisfyed.} \end{cases}$$

If $b = 0$, we do not need to truncate χ_m^2 with Ω_δ and for any $x > 0$

$$\mathbb{P}\left(\chi_m^2 \geq \frac{\sigma^2}{n} D_m + 4C(d,p)\frac{\sigma^2}{n}\sqrt{2D_m x} + 2C(d,p)\frac{\sigma^2}{n} x \right) \leq e^{-x}$$

The constants $A(d,p)$, $B(2d,p)$, $C(d,p)$ and $C'(d,p)$ only depend on d and p. $A(d,p)$ and $B(2d,p)$ are defined in lemma 3.4, $C(d,p) = \dim \mathbb{R}_d[x^1, \ldots, x^p]$ and $C'(d,p) = \sqrt{\frac{2}{C(d,p)}} A(d,p)$.

Remark 3.2 *Under assumptions of lemma 3.3,*

- *if* $N_{min} \geq 2(a+k+1)C(d,p)\frac{(1+b\delta C'(d,p))}{\delta^2\sigma^2}\log n$,

$$\mathbb{P}\left(\Omega_\delta^c\right) \leq \frac{\Gamma}{(a+k+1)}\frac{\delta^2\sigma^2}{(1+b\delta C'(d,p))}\frac{1}{n^k\log n}$$

with $a=0$ *and* $\Gamma=1$ *if assumption (A2) is satisfyed instead of (A1).*

- *if* $\|s\|_\infty \leq R$, *then, for any model* m, $\max_{1\leq i\leq n}|s(\mathbf{x}_i)-s_m(\mathbf{x}_i)| \leq RC''(d,p)$
where $C''(d,p)=1+\sqrt{2}C(d,p)A(d,p)$, *and thus, for any model* m *and for any* $x>0$,

$$\mathbb{P}\left(\pm < \varepsilon, s-s_m >_n \geq \frac{\sigma}{\sqrt{n}}\|s-s_m\|_n\sqrt{2x}+\frac{RbC''(d,p)}{n}x\right) \quad \leq \quad e^{-x} \qquad (3.11)$$

3.6 Proof of lemma 3.3

To prove lemma 3.3, we consider for any hyperrectangle axis oriented subregion $J \subset [0,1]^p$ and any $d_J \in \mathbb{N}$, a family $(\phi_{1,J}, \phi_{2,J}, \ldots, \phi_{D_J,J})$ of polynomial functions with support in J and degree $\leq d_J$ such that the corresponding vectors $((\phi_{1,J}(\mathbf{x}_i))_{1\leq i\leq n}, (\phi_{2,J}(\mathbf{x}_i))_{1\leq i\leq n}, \ldots, (\phi_{D_J,J}(\mathbf{x}_i))_{1\leq i\leq n})$ form an orthonormal basis of the linear subspace of $\mathbb{R}^n$ corresponding to polynomial functions with support in J and degree $\leq d_J$, according to $< .,. >_J$ the Euclidean scalar product of $\mathbb{R}^n$ scaled by a factor $|J|^{-1}$. In other words, the $(\phi_{k,J})_{1\leq k\leq D_J}$ are polynomial functions with support in J and degree $\leq d_J$ such that:

- $\frac{1}{|J|}\sum_{\mathbf{x}_i\in J}\phi_{k,J}(\mathbf{x}_i)\phi_{l,J}(\mathbf{x}_i) = \mathbb{1}_{k=l}$,
- for any polynomial function g with support in J and degree $\leq d_J$, the vector $g(\mathbf{x}_i)_{1\leq i\leq n}$ is a linear combination of $((\phi_{1,J}(\mathbf{x}_i))_{1\leq i\leq n}, (\phi_{2,J}(\mathbf{x}_i))_{1\leq i\leq n}, \ldots, (\phi_{D_J,J}(\mathbf{x}_i))_{1\leq i\leq n})$.

The number of functions D_J depends on (d_J,p) and on the $x_i \in J$ but is always upper bounded by $C(d_J,p)$.
Thanks to the function $\phi_{k,J}$, we can write

$$\|\varepsilon_m\|_n^2 = \sum_{J\in M}\|\Pi_{J,d_J}\varepsilon\|_n^2 = \sum_{J\in M}\sum_{k=1}^{D_J} < \varepsilon, \sqrt{\frac{n}{|J|}}\phi_{k,J} >_n^2 .$$

We need to upper bound the sup-norms $\sup_{\mathbf{x}\in J}|\phi_{k,J}(\mathbf{x})|$ for any $0\leq k\leq D_J$ and any region J by a constant which depends only on (d,p) and not on J nor on the points $\mathbf{x}_i \in J$.

Let us recall that the Legendre polynomials, denoted by $(L_j)_{j\geq 0}$, form an orthogonal basis of $\left(\mathbb{R}[x], L^2([-1,1])\right)$ and satisfy $\sup_{x\in[-1,1]}|L_j(x)|=1$ and $\int_{-1}^1 L_j(x)^2dx = \left(j+\frac{1}{2}\right)^{-1}$. Thus $\left(P_j=\sqrt{j+\frac{1}{2}}L_j\right)_{j\geq 0}$ are orthonormal polynomials of $\left(\mathbb{R}[x], L^2([-1,1])\right)$ and

$$\sup_{x\in[-1,1]}|P_j(x)| = \sqrt{j+\frac{1}{2}}.$$

And therefore for any $P \in \mathbb{R}_d[x]$,

$$\sup_{x\in[-1,1]} |P(x)| \leq \frac{d+1}{\sqrt{2}} \left(\int_{-1}^{1} P(x)^2 dx \right)^{1/2}.$$

The lemma 3.5 and corollary 3.1 state that we have a similar result when we replace $\mathbb{R}_d[x]$ by $\mathbb{R}_d[x^1, \ldots, x^p]$ and the $L^2([-1,1])$ scalar product by a l^2 discret scalar product associated with a set of points $\mathbf{x}_i = (x_i^1, \ldots, x_i^p)$ "well-distributed" in a hyperrectangle axis oriented region like $[0,1]^p$.
The lemma 3.4 defines the constants A and B involved in lemma 3.5 and corollary 3.1.

Lemma 3.4 *For any $g \in \mathcal{C}^p([0,1]^p)$, we define*

$$\begin{aligned} N(g) &= \left\| \frac{\partial^p g}{\partial x_1 \ldots \partial x_p} \right\|_\infty + \sum_{k=1}^{p} \left\| \frac{\partial^{p-1} g}{\partial x_1 \ldots \partial x_{k-1} \partial x_{k+1} \ldots \partial x_p} (\square, \ldots, \square, 1, \square, \ldots, \square) \right\|_\infty \\ &\quad + \ldots + \sum_{j=1}^{p} \left\| \frac{\partial g}{\partial x_j} (1, \ldots, 1, \square, 1, \ldots, 1) \right\|_\infty + |g(1, \ldots, 1)| . \end{aligned}$$

N is a norm on $\mathcal{C}^p([0,1]^p)$.
Since $\mathbb{R}_d[\mathbf{x}] = \mathbb{R}_d[x^1, \ldots, x^p] \subset \mathcal{C}^p([0,1]^p)$ is a linear space with finite dimension, there exists two positive constants $A(d,p)$ and $B(d,p)$ such that, for any $f \in \mathbb{R}_d[\mathbf{x}]$,

- $\|f\|_\infty \leq A(d,p) \sqrt{\int f(\mathbf{x})^2 d\mathbf{x}}$
- $N(f) \leq B(d,p) \int |f(\mathbf{x})| d\mathbf{x}$

$A(d,p)$ and $B(d,p)$ only depend on d and p.

Lemma 3.5 *Let $(\mathbf{x}_i)_{1\leq i\leq n}$ n points in $[0,1]^p$. Denote F the uniform distribution function on $[0,1]^p$ and F_n the empirical distribution function associated with $(\mathbf{x}_i)_{1\leq i\leq n}$ as defined by (3.4) in section 3.2.*
If

$$\sup_{\mathbf{x}\in[0,1]^p} |F_n(\mathbf{x}) - F(\mathbf{x})| \leq \frac{1}{2B(2d,p)}$$

then, for any $f \in \mathbb{R}_d[\mathbf{x}] = \mathbb{R}_d[x^1, \ldots, x^p]$,

$$\sup_{\mathbf{x}\in[0,1]^p} |f(\mathbf{x})| \leq \sqrt{2} A(d,p) \sqrt{\frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}_i)^2}$$

where $A(d,p)$ and $B(2d,p)$ are the two positive constants defined in lemma 3.4.

Corollary 3.1 *Let $(\mathbf{x}_i)_{1\leq i\leq n}$ n points in $[0,1]^p$ and J an hyperrectangle axis oriented subregion of $[0,1]^p$. Denote $|J| = |\{1 \leq i \leq n;\ \mathbf{x}_i \in J\}|$, F_J the uniform distribution function on J, and $F_{n,J}$ the empirical distribution function associated with the points $\mathbf{x}_i \in J$.*
If

$$\sup_{\mathbf{x}\in J} |F_{n,J}(\mathbf{x}) - F_J(\mathbf{x})| \leq \frac{1}{2B(2d,p)} \tag{3.12}$$

then, for any $f \in \mathbb{R}_d[\mathbf{x}]$,

$$\sup_{\mathbf{x}\in J}|f(\mathbf{x})| \leq \sqrt{2}A(d,p)\sqrt{\frac{1}{|J|}\sum_{\mathbf{x}_i\in J} f(\mathbf{x}_i)^2} \tag{3.13}$$

where $A(d,p)$ *and* $B(2d,p)$ *are the two positive constants defined in lemma 3.4.*

Remark 3.3 *What is important in this result is that the constant* $\sqrt{2}A(d,p)$*, which appears in inequality (3.13), does not depend neither on the region* J *nor on the number* $|J|$ *of points. In order that assumption (3.12) is satisfied for any hyperrectangle region* $J \in M$*,* $M \in \mathcal{M}_n$*, and for the points* $\mathbf{x}_i \in J$*, we suppose that*

$$\|F_n - F\|_\infty \leq \frac{\inf_{M\in\mathcal{M}_n}\inf_{J\in M}\mu(J)}{2^{p+2}B(2d,p)}$$

Then, thanks to corollary 3.1, for any hyperrectangle region $J \in M$*,* $M \in \mathcal{M}_n$*, and any* $d_j \leq d$*, we have:*

$$\forall\ 1 \leq k \leq D_J \quad \|\phi_{k,J}\|_\infty \leq \sqrt{2}A(d,p)$$

Let first prove lemma 3.5. Then we use its corollary to prove lemma 3.3.

Proof of lemma 3.5

Let F the uniform distribution function on $[0,1]^p$ and F_n the empirical distribution function associated with $(\mathbf{x}_i)_{1\leq i\leq n}$ as defined by (3.4) in section 3.2.
For any function $f \in \mathcal{C}^p([0,1]^p)$, by integrating part by part with respect to each variable x^j the term $\int\ldots\int(F_n(x^1,\ldots,x^p) - x^1\ldots x^p)\frac{\partial^p f}{\partial x^1\ldots\partial x^p}(x^1,\ldots,x^p)dx^1\ldots dx^p$, we get that:

$$\begin{aligned}
&\int\ldots\int f(x^1,\ldots,x^p)dx^1\ldots dx^p - \frac{1}{n}\sum_{i=1}^n f(x_i^1,\ldots,x_i^p)\\
= \quad &(-1)^{p-1}\int\ldots\int(F_n(x^1,\ldots,x^p) - x^1\ldots x^p)\frac{\partial^p f}{\partial x^1\ldots\partial x^p}(x^1,\ldots,x^p)dx^1\ldots dx^p\\
+ \quad &(-1)^{p-2}\sum_{k=1}^p\int\ldots\int(F_n(\ldots,1,\ldots) - \prod_{j\neq k}x^j)\frac{\partial^{p-1}f}{(\partial x^j)_{j\neq k}}(\ldots,1,\ldots)dx^1\ldots dx^{k-1}dx^{k+1}\ldots dx^p\\
+ \quad &\ldots\\
+ \quad &\sum_{j=1}^p\int(F_n(1,\ldots,1,x^j,1,\ldots,1) - x^j)\frac{\partial f}{\partial x^j}(1,\ldots,1,x^j,1,\ldots,1)dx^j.
\end{aligned}$$

Thus

$$\left|\int\ldots\int f(x^1,\ldots,x^p)dx^1\ldots dx^p - \frac{1}{n}\sum_{i=1}^n f(x_i^1,\ldots,x_i^p)\right| \leq \|F_n - F\|_\infty N(f) \tag{3.14}$$

where N is the norm on $\mathcal{C}^p([0,1]^p)$ defined in lemma 3.4.
Let $P \in \mathbb{R}_d[\mathbf{x}]$. According to (3.14) with $f = P^2$ and to lemma 3.4, we get that

$$\left|\int P(\mathbf{x})^2 d\mathbf{x} - \frac{1}{n}\sum_{i=1}^n P(\mathbf{x}_i)^2\right| \leq \|F_n - F\|_\infty N(P^2) \leq \frac{1}{2B(2d,p)}N(P^2) \leq \frac{1}{2}\int P(\mathbf{x})^2 d\mathbf{x}$$

and thus

$$\int P(\mathbf{x})^2 d\mathbf{x} \leq 2\frac{1}{n}\sum_{i=1}^{n} P(\mathbf{x}_i)^2$$

It follows from lemma 3.4 that

$$\|P\|_\infty \leq \sqrt{2}A(d,p)\sqrt{\frac{1}{n}\sum_{i=1}^{n} P(\mathbf{x}_i)^2}$$

□

We are now able to prove lemma 3.3.

Proof of lemma 3.3
Let $m = (M, \underline{\mathsf{d}})$ where $M \in \mathcal{M}_n$ and $\underline{\mathsf{d}} = (d_J)_{J\in M} \in \{0, 1, \ldots, d\}^M$.
Then

$$\chi_m^2 = \|\varepsilon_m\|_n^2 = \sum_{J\in M} \|\Pi_{J,d_J}\varepsilon\|_n^2.$$

As, for any $J \in M$ and any $d_J \leq d$,

$$\|\Pi_{J,d_J}\varepsilon\|_n^2 \leq \|\Pi_{J,d}\varepsilon\|_n^2$$

on the set Ω_δ,

$$\|\Pi_{J,d_J}\varepsilon\|_n \leq \delta\sigma^2\sqrt{\frac{|J|}{n}}.$$

Let denote, for any $J \in M$,

$$Z_J = \|\Pi_{J,d_J}\varepsilon\|_n^2 \wedge \left(\delta^2\sigma^4\frac{|J|}{n}\right)$$

$(Z_J)_{J\in M}$ are independent random variables such that:

- on the set Ω_δ $\chi_m^2 = \sum_{J\in M} Z_J$,
- $\mathbb{E}(Z_J) \leq \mathbb{E}\left(\|\Pi_{J,d_J}\varepsilon\|_n^2\right) = \frac{\tau^2}{n}D_J \leq C(d_J,p)\frac{\sigma^2}{n}$,
- and for any $k \geq 2$

$$\mathbb{E}\left(|Z_J|^k\right) = \int_0^{\delta\sigma^2\sqrt{\frac{|J|}{n}}} 2kt^{2k-1}\mathbb{P}\left(\|\Pi_{J,d_J}\varepsilon\|_n \geq t\right)dt \tag{3.15}$$

In order to upper bound $\mathbb{E}\left(|Z_J|^k\right)$, we first upper bound each $\mathbb{P}\left(\|\Pi_{J,d_J}\varepsilon\|_n \geq t\right)$. We use the functions $\phi_{k,J}$ defined above to decompose $\|\Pi_{J,d_J}\varepsilon\|_n^2$ as follows:

$$\|\Pi_{J,d_J}\varepsilon\|_n^2 = \sum_{k=1}^{D_J} < \varepsilon, \sqrt{\frac{n}{|J|}}\phi_{k,J} >_n^2$$

Then

$$\begin{aligned}\mathbb{P}\left(\|\Pi_{J,d_J}\varepsilon\|_n \geq t\right) &= \mathbb{P}\left(\sum_{k=1}^{D_J} < \varepsilon, \sqrt{\frac{n}{|J|}}\phi_{k,J} >_n^2 \geq t^2\right) \\ &\leq \sum_{j=1}^{D_J} \mathbb{P}\left(< \varepsilon, \sqrt{\frac{n}{|J|}}\phi_{k,J} >_n^2 \geq \frac{t^2}{D_J}\right) \\ &\leq \sum_{j=1}^{D_J} \mathbb{P}\left(\left|< \varepsilon, \sqrt{\frac{n}{|J|}}\phi_{k,J} >_n\right| \geq \frac{t}{\sqrt{D_J}}\right) \qquad (3.16)\end{aligned}$$

Thanks to assumption (3.3), we get that for any $0 < t \leq \delta\sigma^2\sqrt{\frac{|J|}{n}}$,

$$\begin{aligned}\mathbb{P}\left(\left|< \varepsilon, \sqrt{\frac{n}{|J|}}\phi_{k,J} >_n\right| \geq \frac{t}{\sqrt{D_J}}\right) &\leq 2\exp\left(\frac{-t^2}{2D_J\left(\frac{\sigma^2}{n} + \frac{b}{n}\sqrt{\frac{n}{|J|}}\|\phi_{k,J}\|_\infty\frac{t}{\sqrt{D_J}}\right)}\right) \\ &\leq 2\exp\left(\frac{-t^2}{2D_J\frac{\sigma^2}{n}\left(1 + \frac{b\delta}{\sqrt{D_J}}\|\phi_{k,J}\|_\infty\right)}\right)\end{aligned}$$

$D_J \leq C(d_J, p) \leq C(d, p)$ and according to corollary 3.1 we get that $\|\phi_{k,J}\|_\infty \leq \sqrt{2}A(d,p)$. Thus, for any $0 < t \leq \delta\sigma^2\sqrt{\frac{|J|}{n}}$,

$$\mathbb{P}\left(\left|< \varepsilon, \sqrt{\frac{n}{|J|}}\phi_{k,J} >_n\right| \geq \frac{t}{\sqrt{D_J}}\right) \leq 2\exp\left(\frac{-t^2}{2C(d,p)\frac{\sigma^2}{n}\left(1 + b\delta C'(d,p)\right)}\right)$$

where $C'(d,p) = \sqrt{\frac{2}{C(d,p)}}A(d,p)$.

It follows from inequality (3.16) that, for any $0 < t \leq \delta\sigma^2\sqrt{\frac{|J|}{n}}$,

$$\mathbb{P}\left(\|\Pi_{J,d_J}\varepsilon\|_n \geq t\right) \leq 2C(d_J,p)\exp\left(\frac{-t^2}{2C(d,p)\frac{\sigma^2}{n}\left(1 + b\delta C'(d,p)\right)}\right) \qquad (3.17)$$

and equality (3.15) gives:

$$\mathbb{E}\left(|Z_J|^k\right) \leq \int_0^{+\infty} 4C(d_J,p)kt^{2k-1}\exp\left(\frac{-t^2}{2C(d,p)\frac{\sigma^2}{n}\left(1 + b\delta C'(d,p)\right)}\right)dt$$

Integrating part by part, and summing with respect to $J \in M$, we get

$$\sum_{J\in M}\mathbb{E}\left(|Z_J|^k\right) \leq \frac{k!}{2}D_m\left(4C(d,p)\frac{\sigma^2}{n}(1 + b\delta C'(d,p))\right)^2\left(2C(d,p)\frac{\sigma^2}{n}(1 + b\delta C'(d,p))\right)^{k-2}$$

Thanks to Bernstein inequality we obtain that for any $x > 0$

$$\mathbb{P}\left(\sum_{J\in M} Z_J \geq \frac{\sigma^2}{n}D_m + 4C(d,p)\frac{\sigma^2}{n}(1 + b\delta C'(d,p))\sqrt{2D_m x} + 2C(d,p)\frac{\sigma^2}{n}(1 + b\delta C'(d,p))x\right) \leq e^{-x}$$

Since $\chi^2_m = \sum_{J\in M} Z_J$ on the set Ω_δ,

$$\mathbb{P}\left(\chi^2_m \mathbb{1}_{\Omega_\delta} \geq \frac{\sigma^2}{n}D_m + 4C(d,p)\frac{\sigma^2}{n}(1+b\delta C'(d,p))\sqrt{2D_m x} + 2C(d,p)\frac{\sigma^2}{n}(1+b\delta C'(d,p))x\right) \leq e^{-x}$$

It remains to upper bound $\mathbb{P}(\Omega^c_\delta)$.
Thanks to inequality (3.17), for any $J \in M$, $M \in \mathcal{M}_n$, we have

$$\begin{aligned}\mathbb{P}\left(\|\Pi_{J,d}\varepsilon\|_n \geq \delta\sigma^2\sqrt{\frac{|J|}{n}}\right) &\leq 2C(d,p)\exp\left(\frac{-\delta^2\sigma^2|J|}{2C(d,p)(1+b\delta C'(d,p))}\right)\\ &\leq 2C(d,p)\exp\left(\frac{-\delta^2\sigma^2 N_{min}}{2C(d,p)(1+b\delta C'(d,p))}\right)\end{aligned}$$

If $\mathcal{M}_n$ satisfies (A1), then summing these inequalities over $J \in M$ and over $M \in \mathcal{M}_n$, we get

$$\mathbb{P}\left(\Omega^c_\delta\right) \leq 2\Gamma C(d,p)\left(\frac{n}{N_{min}}\right)^{a+1}\exp\left(\frac{-\delta^2\sigma^2 N_{min}}{2C(d,p)(1+b\delta C'(d,p))}\right).$$

If $\mathcal{M}_n$ satisfies (A2), then $\Omega_\delta = \left\{\forall J \in M_0;\ \|\Pi_{J,d}\varepsilon\|_n \leq \delta\sigma^2\sqrt{\frac{|J|}{n}}\right\}$ and we only need to sum the preceding inequalities over $J \in M_0$. We then get

$$\mathbb{P}\left(\Omega^c_\delta\right) \leq 2C(d,p)\frac{n}{N_{min}}\exp\left(\frac{-\delta^2\sigma^2 N_{min}}{2C(d,p)(1+b\delta C'(d,p))}\right).$$

□

3.7 Proof of the theorem

The proof is exactly the same as the proof of theorem 1.1 in chapter 1. The only difference is that we use the concentration inequality of lemma 3.3 instead of lemma 1.1, and that we need again the upper bound of corollary 3.1.
For simplicity, we write the proof for the case (A2) where all partitions are built from an initial one denoted by M_0.

Let $\theta \in (0,1)$ and $K > 2-\theta$.

According to (3.10),

$$(1-\theta)\|s-\hat{s}_{\widehat{m}}\|^2_n = \Delta_{\widehat{m}} + \inf_{m\in\mathfrak{M}_n} R_m \tag{3.18}$$

where

$$\begin{aligned}\Delta_m &= (2-\theta)\|\varepsilon_m\|^2_n - 2<\varepsilon, s-s_m>_n -\theta\|s-s_m\|^2_n - \mathrm{pen}(m)\\ R_m &= \|s-s_m\|^2_n - \|\varepsilon_m\|^2_n + 2<\varepsilon, s-s_m>_n + \mathrm{pen}(m)\end{aligned}$$

Let denote $\Omega = \bigcap_{M\in\mathcal{M}_n}\left\{\forall J\in M;\ \|\Pi_{J,d}\varepsilon\|_n \leq \frac{\sigma^2}{bC'(d,p)}\sqrt{\frac{|J|}{n}}\right\} = \left\{\forall J\in M_0;\ \|\Pi_{J,d}\varepsilon\|_n \leq \frac{\sigma^2}{bC'(d,p)}\sqrt{\frac{|J|}{n}}\right\}$

Thanks to lemma 3.3,

$$\mathbb{P}\left(\Omega^c\right) \leq 2C(d,p)\frac{n}{N_{min}}\exp\left(\frac{-\sigma^2 N_{min}}{4C(d,p)b^2C'(d,p)^2}\right)$$

and, for any $m \in \mathfrak{M}_n$ and any $x > 0$,

$$\mathbb{P}\left(\|\varepsilon_m\|_n^2 \mathbb{1}_\Omega \geq \frac{\sigma^2}{n}D_m + 8C(d,p)\frac{\sigma^2}{n}\sqrt{2D_m x} + 4C(d,p)\frac{\sigma^2}{n}x\right) \leq e^{-x} \tag{3.19}$$

Thanks to inequality (3.11), we have for any $m \in \mathfrak{M}_n$ and any $x > 0$,

$$\mathbb{P}\left(- < \varepsilon, s - s_m >_n \geq \frac{\sigma}{\sqrt{n}}\|s - s_m\|_n\sqrt{2x} + \frac{RbC''(d,p)}{n}x\right) \quad \leq \quad e^{-x} \tag{3.20}$$

Setting $x = x_m + \xi$ with $\xi > 0$, and summing all inequalities (3.19) and (3.20) with respect to $m \in \mathfrak{M}_n$, we derive a set E_ξ such that:

- $\mathbb{P}\left(E_\xi^c\right) \leq e^{-\xi}2\Sigma$
- on the set $E_\xi \bigcap \Omega$, for any m,

$$\begin{aligned}\Delta_m \quad \leq \quad & (2-\theta)\frac{\sigma^2}{n}D_m + 8(2-\theta)C(d,p)\frac{\sigma^2}{n}\sqrt{2D_m(x_m+\xi)} + 4(2-\theta)C(d,p)\frac{\sigma^2}{n}(x_m+\xi) \\ & +2\frac{\sigma}{\sqrt{n}}\|s-s_m\|_n\sqrt{2(x_m+\xi)} + \frac{2C''(d,p)Rb}{n}(x_m+\xi) \\ & -\theta\|s-s_m\|_n^2 - \mathrm{pen}(m)\end{aligned}$$

Using the two following inequalities
$2\frac{\sigma}{\sqrt{n}}\|s-s_m\|_n\sqrt{2(x_m+\xi)} \leq \theta\|s-s_m\|_n^2 + \frac{2}{\theta}\frac{\sigma^2}{n}(x_m+\xi)$,
$8(2-\theta)C(d,p)\frac{\sigma^2}{n}\sqrt{2D_m(x_m+\xi)} \leq 8\sqrt{2}(2-\theta)C(d,p)\frac{\sigma^2}{n}\sqrt{D_m x_m} + 4\sqrt{2}(2-\theta)C(d,p)\frac{\sigma^2}{n}\left(\eta D_m + \eta^{-1}\xi\right)$
with $\eta = \frac{1}{4\sqrt{2}C(d,p)}\frac{K+\theta-2}{2-\theta} > 0$, we deduce that on the set $E_\xi \bigcap \Omega$, for any m,

$$\begin{aligned}\Delta_m \quad \leq \quad & (2-\theta)\frac{\sigma^2}{n}D_m + 8(2-\theta)C(d,p)\frac{\sigma^2}{n}\sqrt{2D_m(x_m+\xi)} \\ & + \left(4(2-\theta)C(d,p) + \frac{2}{\theta}\right)\frac{\sigma^2}{n}(x_m+\xi) + 2C''(d,p)\frac{Rb}{n}(x_m+\xi) \\ & -\mathrm{pen}(m) \\ \leq \quad & K\frac{\sigma^2}{n}D_m + 8\sqrt{2}(2-\theta)C(d,p)\frac{\sigma^2}{n}\sqrt{D_m x_m} + \left(4(2-\theta)C(d,p) + \frac{2}{\theta}\right)\frac{\sigma^2}{n}x_m + 2C''(d,p)\frac{Rb}{n}x_m \\ & + \left[4(2-\theta)C(d,p)\left(1 + 8C(d,p)\frac{(2-\theta)}{K+\theta-2}\right) + \frac{2}{\theta}\right]\frac{\sigma^2}{n}\xi + 2C''(d,p)\frac{Rb}{n}\xi - \mathrm{pen}(m)\end{aligned}$$

Taking a penalty pen wich compensates for all the other terms in m, i.e.

$$\mathrm{pen}(m) \geq K\frac{\sigma^2}{n}D_m + 8\sqrt{2}(2-\theta)C(d,p)\frac{\sigma^2}{n}\sqrt{D_m x_m} + \left[\left(4(2-\theta)C(d,p) + \frac{2}{\theta}\right)\frac{\sigma^2}{n} + 2C''(d)\frac{Rb}{n}\right]x_m$$

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

$$\Delta_{\widehat{m}} \leq \left[4(2-\theta)C(d,p)\left(1+8C(d,p)\frac{(2-\theta)}{K+\theta-2}\right)+\frac{2}{\theta}\right]\frac{\sigma^2}{n}\xi + 2C''(d,p)\frac{Rb}{n}\xi$$

In other words, on the set E_ξ,

$$\Delta_{\widehat{m}}\mathbb{1}_\Omega \leq \left[4(2-\theta)C(d,p)\left(1+8C(d,p)\frac{(2-\theta)}{K+\theta-2}\right)+\frac{2}{\theta}\right]\frac{\sigma^2}{n}\xi + 2C''(d,p)\frac{Rb}{n}\xi$$

Integrating with respect to ξ,

$$\mathbb{E}\left(\Delta_{\widehat{m}}\mathbb{1}_\Omega\right) \leq \left[8(2-\theta)C(d,p)\left(1+8C(d,p)\frac{(2-\theta)}{K+\theta-2}\right)+\frac{4}{\theta}\right]\frac{\sigma^2}{n}\Sigma + 4C''(d,p)\frac{Rb}{n}\Sigma. \quad (3.21)$$

We are going now to control $\mathbb{E}\left(\inf_m R_m \mathbb{1}_\Omega\right)$.
Thanks to inequality (3.11), for any m and any $x>0$

$$\mathbb{P}\Big(\langle \varepsilon, s-s_m\rangle_n \geq \frac{\sigma}{\sqrt{n}}\|s-s_m\|_n\sqrt{2x}+\frac{RbC''(d,p)}{n}x\Big) \quad \leq \quad e^{-x}$$

Thus we derive a set F_ξ such that

- $\mathbb{P}\left(F_\xi^c\right) \leq e^{-\xi}\Sigma$

- on the set F_ξ, for any m,

$$< \varepsilon, s-s_m >_n \quad \leq \quad \frac{\sigma}{\sqrt{n}}\|s-s_m\|_n\sqrt{2\left(x_m+\xi\right)}+\frac{RbC''(d,p)}{n}\left(x_m+\xi\right)$$

It follows from definition of R_m that on the set F_ξ, for any m,

$$\begin{aligned} R_m &\leq \|s-s_m\|_n^2 + 2\frac{\sigma}{\sqrt{n}}\|s-s_m\|_n\sqrt{2\left(x_m+\xi\right)} + 2C''(d,p)\frac{Rb}{n}\left(x_m+\xi\right) + \mathrm{pen}(m) \\ &\leq 2\|s-s_m\|_n^2 + 2\frac{\sigma^2}{n}\left(x_m+\xi\right) + 2C''(d,p)\frac{Rb}{n}\left(x_m+\xi\right) + \mathrm{pen}(m) \\ &\leq 2\|s-s_m\|_n^2 + 2\mathrm{pen}(m) + 2\frac{\sigma^2}{n}\xi + 2C''(d,p)\frac{Rb}{n}\xi \end{aligned}$$

And

$$\begin{aligned} \mathbb{E}\left(\inf_m R_m \mathbb{1}_\Omega\right) \quad \leq \quad & 2\inf_m\left\{\|s-s_m\|_n^2+\mathrm{pen}(m)\right\} \\ & +2\frac{\sigma^2}{n}\Sigma + 2C''(d,p)\frac{Rb}{n}\Sigma \end{aligned} \quad (3.22)$$

We conclude from (3.18), (3.21) and (3.22) that

$$\begin{aligned} (1-\theta)\mathbb{E}\left(\|s-\hat{s}_{\hat{m}}\|_n^2\mathbb{1}_\Omega\right) \quad \leq \quad & 2\inf_m\left\{\|s-s_m\|_n^2+\mathrm{pen}(m)\right\} \\ & + \left[8(2-\theta)C(d,p)\left(1+8C(d,p)\frac{(2-\theta)}{K+\theta-2}\right)+\frac{4}{\theta}+2\right]\frac{\sigma^2}{n}\Sigma + 6C''(d,p)\frac{Rb}{n}\Sigma \end{aligned}$$

It remains to control $\mathbb{E}\left(\|s-\hat{s}_{\hat{m}}\|_n^2 1\!\!1_{\Omega^c}\right)$, except if $b=0$ in which case it is finished.

$$\begin{aligned}\mathbb{E}\left(\|s-\hat{s}_{\hat{m}}\|_n^2 1\!\!1_{\Omega^c}\right) &= \mathbb{E}\left(\|s-s_{\widehat{m}}\|_n^2 1\!\!1_{\Omega^c}\right)+\mathbb{E}\left(\|\varepsilon_{\widehat{m}}\|_n^2 1\!\!1_{\Omega^c}\right)\\ &\leq \mathbb{E}\left(\|s\|_n^2 1\!\!1_{\Omega^c}\right)+\mathbb{E}\left(\|\varepsilon_{m_0}\|_n^2 1\!\!1_{\Omega^c}\right)\\ &\leq R^2\mathbb{P}\left(\Omega^c\right)+\sqrt{\mathbb{E}\left(\|\varepsilon_{m_0}\|_n^4\right)}\sqrt{\mathbb{P}\left(\Omega^c\right)}\end{aligned}$$

where $m_0=(M_0,\underline{d_0})$ and $\underline{d_0}=(d,d,\ldots,d)$.

By developping $\|\varepsilon_{m_0}\|_n^4=\left(\sum_{J\in M_0}\sum_{k=1}^{D_{0,J}}<\varepsilon,\sqrt{\frac{n}{|J|}}\phi_{k,J}>_n^2\right)^2$ and using $\|\phi_{k,J}\|_\infty\leq\sqrt{2}A(d,p)$ (cf corollary 3.1) and $|M_0|\leq\frac{n}{N_{min}}$, we get

$$\mathbb{E}\left(\|\varepsilon_{m_0}\|_n^4\right) \leq \frac{C(d,p,b,\sigma^2)^2}{N_{min}^2}$$

Thus we have

$$\mathbb{E}\left(\|s-\hat{s}_{\hat{m}}\|_n^2 1\!\!1_{\Omega^c}\right) \leq R^2\mathbb{P}\left(\Omega^c\right)+\frac{C(d,p,b,\sigma^2)}{N_{min}}\sqrt{\mathbb{P}\left(\Omega^c\right)}$$

Let us recall that

$$\mathbb{P}\left(\Omega^c\right)\leq 2C(d,p)\frac{n}{N_{min}}\exp\left(\frac{-\sigma^2 N_{min}}{4C(d,p)b^2C'(d,p)^2}\right)$$

For $N_{min}\geq 12C(d,p)C'(d,p)^2\frac{b^2}{\sigma^2}\log n$,

$$\mathbb{P}\left(\Omega^c\right)\leq\frac{\sigma^2}{6b^2C'(d,p)^2}\frac{1}{n^2\log n}$$

and

$$\mathbb{E}\left(\|s-\hat{s}_{\hat{m}}\|_n^2 1\!\!1_{\Omega^c}\right) \leq C(b,\sigma^2,R,d,p)\frac{1}{n(\log n)^{3/2}}$$

Finally we have the following result:
Taking a penalty which satisfies for all $m\in\mathfrak{M}_n$

$$\mathrm{pen}(m)\geq K\frac{\sigma^2}{n}D_m+8\sqrt{2}(2-\theta)C(d,p)\frac{\sigma^2}{n}\sqrt{D_m x_m}+\left[\left(4(2-\theta)C(d,p)+\frac{2}{\theta}\right)\frac{\sigma^2}{n}+2C''(d,p)\frac{Rb}{n}\right]x_m$$

we have

$$\begin{aligned}\mathbb{E}\left(\|s-\hat{s}_{\hat{m}}\|_n^2\right) \leq\ & \frac{2}{1-\theta}\inf_m\left\{\|s-s_m\|_n^2+\mathrm{pen}(m)\right\}\\ &+\frac{1}{1-\theta}\left[8(2-\theta)C(d,p)\left(1+8C(d,p)\frac{(2-\theta)}{K+\theta-2}\right)+\frac{4}{\theta}+2\right]\frac{\sigma^2}{n}\Sigma\\ &+\frac{6C''(d,p)}{1-\theta}\frac{Rb}{n}\Sigma+C(b,\sigma^2,R,d,p)\frac{1\!\!1_{b\neq 0}}{n(\log n)^{3/2}}\end{aligned}$$

□

Chapter 4

Application to Accelerating Life Test

Ce chapitre présente un travail réalisé en collaboration avec Marc Lavarde.

Abstract: The Accelerating Life Test log-linear model assumes that some specified percentile of life has a log-linear relationship with the stress. Since this model does not always lead to good results, we consider here piecewise log-linear models associated with a collection $\mathcal{M}_n$ of partitions of the stress set. We determine a penalized least squares criterion which selects a close to optimal partition, in the sense that it satisfies an oracle type inequality. The theoretical result gives a penalty defined up to multiplicative constants. Then we propose two methods to calibrate the constants according to the data.

Keywords: Accelerating Life Test, model selection, regression, Weibull log-linear model

4.1 Introduction

Accelerating Life Test (ALT) are used in reliability engineering to obtain timely information on the times-to-failure distributions of components and systems. Rapid change in technology, shorter periods in product development, and higher reliabilities of products, make accelerated tests even more useful and important in today's industries. Test units are subjected to higher than usual levels of stress, such as temperature, voltage, force, humidity, vibration, dust, and use-rate. Then, the test results are extrapolated from the test conditions to the usual use conditions, via a physically reasonable statistical model.

An ALT model specifies the time-to-failure distribution (for example the lognormal or the Weibull distribution) and the relationship between the stress variables (for example the temperature and the voltage) and the parameters of the distribution. Given a time-to-failure distribution, the log-linear model assumes that some specified percentile of life t_q has a log-linear relationship with the stress, i.e.

$$\log t_q = c_0 + \sum_{j=1}^{p} a_j x^j$$

where the x^j are known functions of one or more basic engineering stresses, which are selected according to the failure mechanism. The percentile t_q is chosen according to the assumed

underlying distribution of the life-time. Many well-known and commonly used ALT models, like Arrhenius model and Inverse Power model are special cases of the log-linear model. The table 4.1 recalls these classical examples of Life-Stress relationships. In the following, we consider the Weibull log-linear model.

Model	Relationship	covariate transformation
Arrhenius	$t_q = Ae^{\frac{B}{T}}$	$x = \frac{1}{T}$
Inverse Power	$t_q = \frac{1}{KV^n}$	$x = \log V$
Combination 1	$t_q = \frac{Ce^{\frac{B}{T}}}{V^n}$	$x^1 = \frac{1}{T}$ and $x^2 = \log V$
Combination 2	$t_q = Ae^{\frac{\phi}{T} + \frac{b}{H}}$	$x^1 = \frac{1}{T}$ and $x^2 = \frac{1}{H}$

Table 4.1: Some common Life-Stress relationships.

The Weibull log-linear model is one of the most widely used ALT models in reliability engineering. In this model, we consider that the time-to-failure follows a Weibull distribution with parameters η and β:

$$f(t) = \frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1} e^{-\left(\frac{t}{\eta}\right)^{\beta}} 1\!\!1_{t>0}.$$

The Weibull scale parameter η is the percentile $t_{63.2}$ and is supposed to be a log-linear function of the transformed stress variables:

$$\log \eta = \log t_{63.2} = c_0 + \sum_{j=1}^{p} a_j x^j$$

The Weibull shape parameter β is assumed to be constant or at least independent of stresses. We denote by $\mu = \log \eta - \frac{\gamma}{\beta}$ and by $\tau = \frac{1}{\beta}\sqrt{\frac{\pi^2}{6}}$, where $\gamma = 0.5772\ldots$ is Euler's constant. Then denoting by Y the logarithm of the life-time, we have:

$$\begin{aligned} Y &= \mu + Z \\ \mu &= a_0 + \sum_{j=1}^{p} a_j x^j \end{aligned} \tag{4.1}$$

where Z is an unobservable random variable with mean 0 and variance τ^2, and $(a_0, a_1, \ldots, a_p, \tau)$ are unknown parameters. $\beta Z - \gamma$ is a standard extreme value variable, i.e. $\beta Z - \gamma \sim SEV(0,1)$ (it has mean $-\gamma$ and variance $\frac{\pi^2}{6}$).

Given a model, after the accelerated test, the next step is to fit the test results to the model, i.e. to estimate the model parameters from the test data $(\mathbf{x}_i, Y_i)_{1\leq i\leq n}$ where $\mathbf{x}_i = (x_i^1, \ldots, x_i^p)$ are the values of the transformed stress variables to which the i^{th} component is subjected and Y_i is the logarithm of its time-to-failure. Recall that all $\mathbf{x}_i$ correspond to higher than usual levels of stress. The estimators $(\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_p, \hat{\tau})$ of $(a_0, a_1, \ldots, a_p, \tau)$ allow to understand the relationship between the stress and the time-to-failure. For a usual level of stress $\mathbf{x} = (x^1, \ldots, x^p)$, $\hat{\mu}(\mathbf{x}) = \hat{a}_0 + \sum_{j=1}^{p} \hat{a}_j x^j$ is an estimator of the mean of the life-time logarithm.

In some situations, the classical log-linear models described before lead to bad results. One possible explanation is the following: the component may have several defects denoted by $D_1, D_2, \ldots$. The defect D_1 causes the failure in usual levels of stress. The other defects are less serious and never cause the failure in usual levels of stress. They only break under very high levels of stress. So estimating the parameters $(a_0, a_1, \ldots, a_p, \tau)$ according to test data with too high levels of stress, and extrapolating the relationship (4.1) for usual levels of stress, may lead to bad results. During the test, the value of the stress variables have to be large enough to get timely information, but should not overstep the threshold up to which the other defects $D_2, \ldots$ are detected. Our aim is first to determine the right region of stress, and then to estimate the unknown parameters from the observations corresponding to stress values belonging to this region.

The weaknesses of the log-linear model and the possible existence of defects causing failures only under over-stress conditions lead us to suggest a piecewise log-linear model. We replace the relationship (4.1) between μ and the transformed stress variables x^j by the following one:

$$\mu = \sum_{J \in M} \left\{ a_{0,J} + \sum_{j=1}^{p} a_{j,J} x^j \right\} \mathbb{1}_{\mathbf{x} \in J}$$

where M is an unknown partition of the transformed stress set $\mathcal{S}$. We consider a collection $\mathcal{M}_n$ of partitions of $\mathcal{S}$, and for each $M \in \mathcal{M}_n$ we denote by S_M the space of piecewise linear functions defined on the partition M. We determine a penalized least squares criterion which selects a close to optimal partition $\hat{M}$. In order to get informations on the time-to-failure under usual stress levels, we recommend to use only the region J of $\hat{M}$ which corresponds to the smallest values of stresses. The other regions are over-stress regions.

In ALT, estimation is usually performed via maximum likelihood method. Unfortunately, for the Weibull model, there exists no analytic solutions to the maximization of the likelihood. We can use one of the algorithm implemented to determine the maximum likelihood estimator or move to the least squares estimator. One of the advantage of the least squares criterion is that it does not depend on the underlying distribution. In this paper, we use a penalized least squares criterion to detect the change points in the behavior of the logarithm of the life-time. We select a partition $\hat{M}$ and, at the same time, we estimate the parameters $(a_{0,J}, a_{1,J}, \ldots, a_{p,J})_{J \in \hat{M}}$ by a least squares method. If the maximum likelihood method is preferred, after getting the partition $\hat{M}$ from our penalized least squares criterion, one can remove to the maximum likelihood criterion to estimate $(a_{0,J}, a_{1,J}, \ldots, a_{p,J})_{J \in \hat{M}}$.

The main result of this paper is a consequence of theorem 3.1 (chapter 3). It determines a penalized least squares criterion which allows to select a close to optimal partition of the stress set. This penalized criterion does not depend on the underlying distribution of the life-time. We only assume that the random variable Z has exponential moments around 0, which is the case when the life-time is Weibull distributed. The theoretical result defines a penalty up to multiplicative constants. In a second part of this paper, we propose two methods to calibrate the constants according to the data. We explain their heuristics and evaluate their performances on simulated data.

The paper is organized as follows. The section 4.2 presents the statistical framework and some notations. The section 4.3 gives a simplified version of theorem 3.1 and the form of penalty to be used to select a partition $\hat{M}$ of the stress set. We end this section with some indications to compute $\hat{M}$. The sections 4.4 and 4.5 present two methods to calibrate the unknown constants of the penalty.

4.2 The statistical framework

For simplicity, the transformed stress set $\mathcal{S}$ will be taken to be $[0,1]^p$.

We observe n pairs $(\mathbf{x}_i, Y_i)_{1\leq i\leq n}$, where

- $\mathbf{x}_i = (x_i^1, \ldots, x_i^p) \in [0,1]^p$ are the p values of the transformed stress variables to which the i^{th} test unit is subjected,
- Y_i is the observed life-time logarithm of the i^{th} test unit.

We assume that

$$Y_i \quad = \quad \mu(\mathbf{x}_i) + Z_i$$

where $(Z_i)_{1\leq i\leq n}$ are independent and identically distributed (i.i.d.) unobservable random variables with mean 0, variance τ^2 and exponential moments around 0. μ and τ are unknown parameters. μ is a function of the transformed stress variables and τ is a positive constant.

We consider piecewise linear models for the parameter μ. Let $\mathcal{M}_n$ some collection of partitions of $\mathcal{S} = [0,1]^p$. For any partition $M \in \mathcal{M}_n$, we denote by S_M the space of piecewise linear functions defined on M:

$$S_M = \left\{ \mathbf{x} = (x^1, \ldots, x^p) \to \sum_{J\in M} \left(a_{0,J} + \sum_{j=1}^{p} a_{j,J} x^j \right) \mathbb{1}_{\mathbf{x}\in J};\ a_{j,J} \in \mathbb{R} \right\} \tag{4.2}$$

S_M is the piecewise linear model associated with the partition M. It is a linear space with dimension $D_M = (p+1)|M|$. We do not assume that μ belongs to one of the models $(S_M)_{M\in\mathcal{M}_n}$. The $(S_M)_{M\in\mathcal{M}_n}$ are only approximations of the reality.

For any partition $M \in \mathcal{M}_n$, we denote by $\hat{\mu}_M$ the least squares estimator of μ over S_M.

$$\hat{\mu}_M = \arg\min_{u\in S_M} \gamma_n(u) \quad \text{where } \gamma_n(u) = \|Y-u\|_n^2$$

$\|.\|_n$ denotes the Euclidean norm on $\mathbb{R}^n$ scaled by a factor $n^{-1/2}$ and, for a function u, the vector $(u(\mathbf{x}_i))_{1\leq i\leq n} \in \mathbb{R}^n$ is denoted u too.
$\hat{\mu}_M$ is the piecewise linear estimator belonging to S_M which plays the role of benchmark among all the estimators in S_M. Denoting $\mu_M = \arg\min_{u\in S_M} \|\mu-u\|_n^2$, the quadratic risk of the estimator $\hat{\mu}_M$ is

$$\mathbb{E}\left(\|\mu-\hat{\mu}_M\|_n^2\right) = \|\mu-\mu_M\|_n^2 + \tau^2 \frac{\dim_{\mathbb{R}^n}(S_M)}{n} \leq \|\mu-\mu_M\|_n^2 + \tau^2 \frac{(p+1)|M|}{n}$$

The risk $\mathbb{E}\left(\|\mu - \hat{\mu}_M\|_n^2\right)$ measures the prediction accuracy obtained with the partition M (and more precisely with the least squares estimator $\hat{\mu}_M$).

As $(Z_i)_{1\leq i\leq n}$ are assumed to have exponential moments around 0, there exists two positive constants b and σ such that

$$\forall\ \lambda \in (-1/b, 1/b) \qquad \log \mathbb{E}\left(e^{\lambda Z_i}\right) \leq \frac{\sigma^2\lambda^2}{2(1-b|\lambda|)} \tag{4.3}$$

σ^2 is necessarily greater than τ^2 and can be chosen as close to τ^2 as we want, but at the price of a larger b.

In the following lemma, we define two positive constants $A(p)$ and $B(p)$ which only depend on p and which appear in theorem 4.1. Let $\mathcal{L}_p$ the space of linear functions on $[0,1]^p$.

$$\mathcal{L}_p = \left\{ \mathbf{x} = (x^1, \ldots, x^p) \to a_0 + \sum_{j=1}^{p} a_j x^j;\ (a_0, a_1, \ldots, a_p) \in \mathbb{R}^p \right\}$$

We consider three norms on $\mathcal{L}_p$: the supnorm $\|.\|_\infty$, the $L^2([0,1]^p)$-norm and the norm N_1 defined as follows: for any $f : (x^1, \ldots, x^p) \to a_0 + \sum_{j=1}^p a_j x^j$, $N_1(f) = \sum_{j=0}^p |a_j|$.

Lemma 4.1 *Since $\mathcal{L}_p$ is a linear space with dimension $p+1$, there exists two positive constants $A(p)$ and $B(p)$ such that, for any $f \in \mathcal{L}_p$,*

- $\|f\|_\infty \leq A(p)\sqrt{\int f(\mathbf{x})^2 d\mathbf{x}}$
- $2N_1(f) \leq \sqrt{B(p)}\sqrt{\int f(\mathbf{x})^2 d\mathbf{x}}$

$A(p)$ and $B(p)$ only depend on p.

4.3 The main theorem

Let M_0 a partition of $\mathcal{S} = [0,1]^p$ composed of hyperrectangle axis oriented regions and $\mathcal{M}_n$ a family of partitions composed of hyperrectangle axis oriented regions built from those of M_0, i.e. for any $M \in \mathcal{M}_n$ and any element J of M, J is the union of elements of M_0. We consider the corresponding collection of piecewise linear models $(S_M)_{M\in\mathcal{M}_n}$ as defined by (4.2). We denote by $N_{min} = \inf_{J\in M_0} |J|$ where $|J| = |\{1 \leq i \leq n;\ \mathbf{x}_i \in J\}|$. In order to estimate the $p+1$ parameters $(a_{0,J}, a_{1,J}, \ldots, a_{p,J})$ for any region J, we need $N_{min} \geq p+1$.

The ideal partition M^* is the one which minimizes the quadratic risk $\mathbb{E}\left(\|\mu - \hat{\mu}_M\|_n^2\right)$ over all $M \in \mathcal{M}_n$. It is not necessarily the true partition (when a true partition exists), but the partition which allows to get the best prediction accuracy. Our aim is not to determine all the true change points of μ, but to determine stress regions which lead to good predictions. Unfortunately M^* depends on the unknown function μ, and therefore we have to estimate it. We adopt the non asymptotic approach of model selection via penalization. We select a

partition $\hat{M}$ by minimizing a penalized least squares criterion $\text{crit}(M) = \gamma_n(\hat{\mu}_M) + \text{pen}(M)$ over $\mathcal{M}_n$:

$$\hat{M} = \arg \min_{M \in \mathcal{M}_n} \left\{ \gamma_n(\hat{\mu}_M) + \text{pen}(M) \right\}.$$

It remains to provide a penalty pen such that the partition $\hat{M}$ is close to the optimal partition M^*, in the sense that the PLSE $\hat{\mu}_{\hat{M}}$ satisfies an oracle inequality like (4). The following theorem determines a general form of penalty pen which leads to an oracle type inequality for any family of partitions built from a partition M_0 not too fine, assuming the points $(\mathbf{x}_i)_{1 \leq i \leq n}$ to be well distributed. This result is a consequence of theorem 3.1 (chapter 3) when the degree of the polynomials are fixed to be $d = 1$.

Theorem 4.1 *Let $b \in \mathbb{R}_+$ and $\sigma \in \mathbb{R}_+^*$ such that inequality (4.3) holds.*
Assume the points $(\mathbf{x}_i)_{1 \leq i \leq n}$ to be distributed such that

$$\|F_n - F\|_\infty \leq \frac{24A(p)^2}{2^p(4B(p)+1)} \frac{b^2}{\sigma^2} \frac{\log n}{n}$$

where F_n is the empirical distribution function associated with $(\mathbf{x}_i)_{1 \leq i \leq n}$, F is the uniform distribution function on $[0,1]^p$, and $\|.\|_\infty$ is the sup-norm on $[0,1]^p$.
Let M_0 a partition of $[0,1]^p$ composed of hyperrectangle axis oriented regions and assume that $N_{min} = \inf_{J \in M_0} |J|$ satisfies

$$N_{min} \geq 24A(p)^2 \frac{b^2}{\sigma^2} \log n$$

Let $\mathcal{M}_n$ a family of partitions of $[0,1]^p$ composed of hyperrectangle axis oriented regions built from those of M_0, and $(x_M)_{M \in \mathcal{M}_n}$ a family of weights such that

$$\sum_{M \in \mathcal{M}_n} e^{-x_M} \leq \Sigma \in \mathbb{R}_+^* \tag{4.4}$$

Assume $\|\mu\|_\infty \leq R$, with R a positive constant.
Let $\theta \in (0,1)$ and $K > 2 - \theta$ two numbers.
Taking a penalty satisfying

$$\begin{aligned} pen(M) \quad \geq \quad & K(p+1)\frac{\sigma^2}{n}|M| + 8\sqrt{2}(2-\theta)(p+1)\frac{\sigma^2}{n}\sqrt{(p+1)|M|x_M} \\ & + \left[\left(4(2-\theta)(p+1) + \frac{2}{\theta} \right) \frac{\sigma^2}{n} + 2A'(p)\frac{Rb}{n} \right] x_M \end{aligned} \tag{4.5}$$

we have

$$\begin{aligned} \mathbb{E}\left(\|\mu - \hat{\mu}_{\hat{M}}\|_n^2 \right) \quad \leq \quad & \frac{2}{1-\theta} \inf_M \left\{ \|\mu - \mu_M\|_n^2 + pen(M) \right\} \\ & + \frac{1}{1-\theta} \left[8(2-\theta)(p+1) \left(1 + 8(p+1)\frac{(2-\theta)}{K+\theta-2} \right) + \frac{4}{\theta} + 2 \right] \frac{\sigma^2}{n}\Sigma + \frac{6A'(p)}{1-\theta}\frac{Rb}{n}\Sigma \\ & + C(b,\sigma^2,R,p)\frac{1\!\!1_{b \neq 0}}{n(\log n)^{3/2}} \end{aligned}$$

where $C(b,\sigma^2,R,p)$ is a positive constant which depends only on b, σ^2, R, p,
and $A'(p) = 1 + \sqrt{2}(p+1)A(p)$.

The theorem 4.1 gives the general form of the penalty function

$$\text{pen}(M) = K(p+1)\frac{\sigma^2}{n}|M| + \left(\kappa_1(\theta,p)\frac{\sigma^2}{n}\sqrt{|M|x_M} + \kappa_2(\theta,p)\frac{\sigma^2+Rb}{n}x_M\right) \tag{4.6}$$

The penalty is the sum of two terms: the first one is proportional to $\frac{|M|}{n}$ and the second one depends on the complexity of the family $\mathcal{M}_n$ via the weights $(x_M)_{M\in\mathcal{M}_n}$. For $\theta \in (0,1)$ and $K > 2-\theta$, the PLSE $\hat{\mu}_{\hat{M}}$ satisfies an oracle type inequality

$$\mathbb{E}\left(\|\mu - \hat{\mu}_{\hat{M}}\|_n^2\right) \quad \leq \quad C_1 \inf_M \left\{\|\mu - \mu_M\|_n^2 + \text{pen}(M)\right\} + \frac{C_2}{n}$$

where the constant C_1 only depends on θ, whereas C_2 depends on μ (via R), on p the number of stress variables, on the family $\mathcal{M}_n$ (via Σ) and on the integrability condition of $(Z_i)_{1\leq i\leq n}$ (via σ^2 and b).
In order to get an adequate form of penalty, we have to find weights $(x_M)_{M\in\mathcal{M}_n}$ such that inequality (4.4) holds, and replace the obtained values in expression (4.6) (or equivalently in the right term of (4.5)).

A natural way to get a collection $\mathcal{M}_n$ of partitions is first to draw a grid on $\mathcal{S} = [0,1]^p$ such that there are at least N_{min} points $\mathbf{x}_i$ in each cell of the grid. Then we take M_0 as the partition of $\mathcal{S}$ associated with this grid, and $\mathcal{M}_n$ as the collection of all partitions obtained by removing some axis of the original grid.
We choose weights $(x_M)_{M\in\mathcal{M}_n}$ which depend on M only via $K_M = |M|$, in order to penalize the same way partitions having the same number of elements. Since

$$|\{M \in \mathcal{M}_n;\ |M| = K\}| \leq \binom{N_n}{K},\ \text{where } N_n = |M_0|\,,$$

the weights $x_M = |M|\left(L + \log\left(\frac{N_n}{|M|}\right)\right)$, with $L > 1$ an absolute constant, satisfy (4.4) with $\Sigma = \left(e^{L-1}-1\right)^{-1}$.
Applying theorem 4.1, we finally get that there exists two constants $c_1^0(p)$ and $c_2^0(p)$ which only depend on p such that for any $c_1 \geq c_1^0(p)$ and any $c_2 \geq c_2^0(p)$:
taking

$$\text{pen}(M) \geq (\sigma^2 + Rb)\frac{|M|}{n}\left(c_1 \log\left(\frac{N_n}{|M|}\right) + c_2\right)$$

we have

$$\begin{aligned}\mathbb{E}\left(\|\mu - \hat{\mu}_{\hat{M}}\|_n^2\right) \quad \leq \quad & C_1(c_1,c_2)\inf_M \left\{\|\mu - \mu_M\|_n^2 + \text{pen}(M)\right\} \\ & + \frac{C_2(c_1,c_2,b,\sigma^2,R,p)}{n}.\end{aligned}$$

In the following, we focus on the case where the temperature T is the single stress variable and the considered ALT model is the Weibull Arrhenius one (see table 4.1). In this case, the transformed stress variable is $x = \frac{1}{T}$ and the transformed stress set is $\mathcal{S} = [\frac{1}{T_{max}}, \frac{1}{T_{min}}]$. We consider the grid $(v_j)_{1\leq j\leq N_n}$ with $v_j = x_{jN_{min}}$ in order that there are N_{min} points x_i between two consecutive grid points. We define $\mathcal{M}_n$ as the collection of all partitions of $\mathcal{S}$

with endpoints belonging to the grid. According to the preceding result, we can take the following form of penalty:

$$\text{pen}(M) = (\sigma^2 + Rb)\frac{|M|}{n}\left(c_1 \log\left(\frac{N_n}{|M|}\right) + c_2\right). \tag{4.7}$$

Whatever the underlying distribution, the recommended penalty $\text{pen}(M)$ is the sum of two terms: the first one is proportional to $\frac{|M|}{n}\log\left(\frac{N_n}{|M|}\right)$ and the second one is proportional to $\frac{|M|}{n}$. The right multiplicative factors are unknown. They depend on the parameters of the ditribution and on unknown absolute constants. In the next section, we propose two methods to calibrate the multiplicative factors according to the data. Before that, let us explain how to compute the partition $\hat{M}$ associated with a given penalty pen of the form (4.7).

Since $\text{pen}(M)$ depends on M only via $K_M = |M|$, in order to compute

$$\hat{M} = \arg\min_{M\in\mathcal{M}_n}\left\{\gamma_n(\hat{\mu}_M) + \text{pen}(M)\right\},$$

we proceed in two steps:

- First, we calculate for each $1 \le K \le N_n$ the "best" partition $\hat{M}_K$ of size K:

$$\hat{M}_K = \underset{M\in\mathcal{M}_n;\ |M|=K}{\arg\min}\ \gamma_n(\hat{\mu}_M)$$

  The corresponding estimator is

$$\hat{\mu}_K = \hat{\mu}_{\hat{M}_K}$$

  Its contrast is $\gamma_n(\hat{\mu}_K) = \min_{|M|=K}\gamma_n(\hat{\mu}_M) = \min_{u\in S_K}\gamma_n(u)$, where $S_K = \bigcup_{|M|=K} S_M$.

- Then we determine the "best" size

$$\hat{K} = \arg\min_{1\le K\le N_n}\left\{\gamma_n(\hat{\mu}_K) + \text{pen}(K)\right\}$$

We finally get $\hat{M} = \hat{M}_{\hat{K}}$.

Remark 4.1 *For a given K, there are $\binom{N_n-1}{K-1}$ partitions of size K. Thus the algorithmic cost to determine $\hat{M}_K$ is $\mathcal{O}(N_n^K)$. In order to reduce this cost, we compute the $(\hat{M}_K)_{1\le K\le N_n}$ thanks to the dynamic algorithm described in [16, section 3.3.2]. The total cost of the first step is then $\mathcal{O}(N_n^2)$.*

Remark 4.2 *Since the penalty function pen is only used to select the size $\hat{K}$, we measure the performance of the penalized criterion by comparing the risk of the PLSE to the oracle of the models $(S_K)_{K\ge1}$:*

$$\inf_K \mathbb{E}\left(\|\mu - \hat{\mu}_K\|_n^2\right).$$

In the simulation studies, we approximate this benchmark by Monte Carlo.

4.4 First method

We consider the following form of penalty:

$$\mathrm{pen}(M) = \frac{|M|}{n}\left(\alpha \log\left(\frac{N_n}{|M|}\right) + \beta\right)$$

and we calibrate α and β simultaneously according to the data. The method used in this section is based on Massart's heuristic [19, section 8.5.2, paragraph "Some heuristics"]. This method consists in estimating α and β by fitting $-\gamma_n(\hat{\mu}_K)$ on $\frac{1}{2}\frac{K}{n}\left(\alpha \log\left(\frac{N_n}{K}\right) + \beta\right) + \gamma$ for a series of large K.

4.4.1 Massart's heuristic

We first recall Mallows' heuristic. The quality of a model S_M is given by the risk of the estimator $\hat{\mu}_M$:

$$\mathbb{E}\left(\|\mu - \hat{\mu}_M\|_n^2\right) = \|\mu - \mu_M\|_n^2 + \mathbb{E}\left(\|\mu_M - \hat{\mu}_M\|_n^2\right) = \|\mu - \mu_M\|_n^2 + \sigma^2\frac{D_M}{n}$$

where D_M is the dimension of the linear model S_M. Here $D_M = 2|M| = 2K_M$. The best model corresponds to the partition M^* which minimizes

$$-\|\mu_M\|_n^2 + \sigma^2\frac{D_M}{n} \tag{4.8}$$

Mallow's idea consists in replacing $\|\mu_M\|_n^2$ in (4.8) with an unbiased estimator. As

$$\mathbb{E}\left(\|\hat{\mu}_M\|_n^2\right) = \|\mu_M\|_n^2 + \mathbb{E}\left(\|\mu_M - \hat{\mu}_M\|_n^2\right) = \|\mu_M\|_n^2 + \sigma^2\frac{D_M}{n},$$

$\|\hat{\mu}_M\|_n^2 - \sigma^2\frac{D_M}{n}$ is an unbiased estimator of $\|\mu_M\|_n^2$. Mallows' heuristic is that the minimizer of

$$-\|\hat{\mu}_M\|_n^2 + 2\sigma^2\frac{D_M}{n}$$

mimics M^*. The corresponding C_p criterion is defined by

$$C_p(M) = \gamma_n(\hat{\mu}_M) + 2\sigma^2\frac{D_M}{n} = \gamma_n(\hat{\mu}_M) + 2\mathbb{E}\left(\|\mu_M - \hat{\mu}_M\|_n^2\right)$$

The weakness of Mallows' heuristic is that $\|\hat{\mu}_M\|_n^2$ will not necessarly stay of the same order of magnitude as its expectation for all M simultaneously. This analysis is true only if the number of partitions $M \in \mathcal{M}_n$ with a given dimension is not too large. Thanks to theorem 4.1, if $|\{M \in \mathcal{M}_n; K_M = K\}| \leq \Gamma K^a$, then the PLSE associated with $\mathrm{pen}(M) = 2\sigma^2\frac{D_M}{n} = 4\sigma^2\frac{|M|}{n}$ satisfies an oracle type inequality. For the collection $\mathcal{M}_n$ of all partitions of $\mathcal{S}$ with endpoints belonging to a grid of N_n distinct points, theorem 4.1 recommends a stronger form of penalty. In a general Gaussian regression framework, Birgé and Massart [4] prove that for such a big collection, Mallows' C_p criterion can lead to terrible results.

Since Mallows' C_p criterion works for small families of models and since we choose penalties which depend on M only via K_M, in practice for a given collection of models one can build a new smaller list of models $(S_K)_{K\geq 1}$ with $S_K = \bigcup_{K_M=K} S_M$. Like in section 4.3, we consider $\hat{\mu}_K = \hat{\mu}_{\hat{M}_K}$ and $\mu_K = \mu_{\hat{M}_K}$ with $\hat{M}_K = \underset{M\in\mathcal{M}_n;\ K_M=K}{\arg\min} \gamma_n(\hat{\mu}_M)$. Applying the preceding heuristic, we get that the right penalty is

$$\text{pen}(K) = 2\mathbb{E}\left(\|\mu_K - \hat{\mu}_K\|_n^2\right)$$

Since

- $\mathbb{E}\left[\gamma_n(\hat{\mu}_K)\right] = \mathbb{E}\left[\gamma_n(\mu_K)\right] + \mathbb{E}\left[\gamma_n(\hat{\mu}_K) - \gamma_n(\mu_K)\right] = \mathbb{E}\left[\gamma_n(\mu_K)\right] - \mathbb{E}\left[\|\hat{\mu}_K - \mu_K\|_n^2\right]$,
- $\mathbb{E}\left[\gamma_n(\mu_K)\right] \simeq \mathbb{E}\left[\gamma_n(\mu)\right] + \mathbb{E}\left[\|\mu - \mu_K\|_n^2\right]$ is nearly constant for large values of K,

if $\gamma_n(\hat{\mu}_K)$ is close to its expectation, then for large K

$$\begin{aligned} -\gamma_n(\hat{\mu}_K) &\simeq \frac{1}{2}\text{pen}(K) + \gamma \\ &\simeq \frac{1}{2}\frac{K}{n}\left(\alpha \log\left(\frac{N_n}{K}\right) + \beta\right) + \gamma \end{aligned}$$

We estimate α and β thanks to the least squares fit of $-\gamma_n(\hat{\mu}_K)$ on $\frac{1}{2}\frac{K}{n}\left(\alpha \log\left(\frac{N_n}{K}\right) + \beta\right)+\gamma$, with $K_{min} \leq K \leq K_{max}$.

4.4.2 Results obtained on simulated data

We simulate $Nsim = 100$ realisations of

$$Y_i = \mu(\mathbf{x}_i) + Z_i, \quad 1 \leq i \leq n \tag{4.9}$$

with variance $\tau^2 = 1$ fixed and the function μ defined as follows:

$$\mu(T) = \begin{cases} -4 + \frac{3300}{T} & \text{if } 300 < T \leq 330 \\ -\frac{384}{11} + \frac{13500}{T} & \text{if } 330 < T \leq 356 \\ -35.2 + \frac{13500}{T} & \text{if } 356 < T \leq 370 \\ -40 + \frac{13500}{T} & \text{if } 370 < T \leq 400 \end{cases} \tag{4.10}$$

The plot of μ is given in figure 4.1.
The random variables $Z_i = \sqrt{\frac{6}{\pi^2}}(S_i + \gamma)$ where $S_i \sim SEV(0,1)$.

We give here the results obtained by fitting $-\gamma_n(\hat{\mu}_K)$ on $\frac{1}{2}\frac{K}{n}\left(\alpha \log\left(\frac{N_n}{K}\right) + \beta\right) + \gamma$ and by using the penalty

$$\text{pen}(M) = \frac{|M|}{n}\left(\hat{\alpha} \log\left(\frac{N_n}{|M|}\right) + \hat{\beta}\right).$$

Since we do not know which values of K should be considered for the least squares fit used to estimate α and β, we apply the method for six different series of values of K with the restriction that the maximal value K_{max} has to be smaller than $\frac{n}{2}$:

- series 1: $5 \leq K \leq 25$

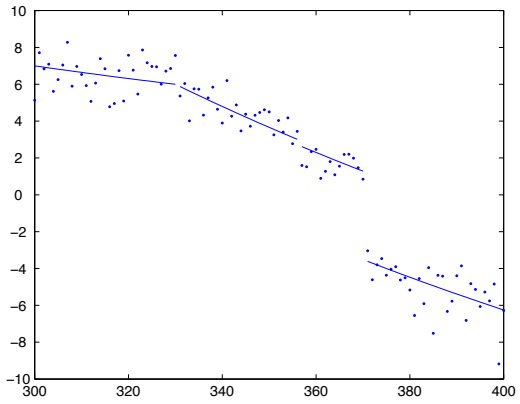

Figure 4.1: plot of μ

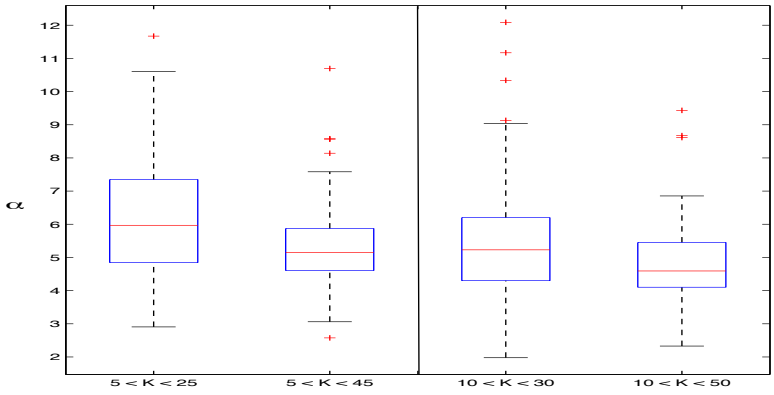

Figure 4.2: boxplot of $\hat{\alpha}$ for samples with size $n = 100$.

- series 2: $5 \leq K \leq 45$
- series 3: $5 \leq K \leq 95$
- series 4: $10 \leq K \leq 30$
- series 5: $10 \leq K \leq 50$
- series 6: $10 \leq K \leq 100$

The figures 4.2, 4.3 and 4.4 give the values of $\hat{\alpha}$, $\hat{\beta}$ and $\frac{\hat{\beta}}{\hat{\alpha}}$ obtained for $N_{sim} = 100$ simulated samples of size $n = 100$ with the different series of values of K. The values of $\hat{\alpha}$ and $\hat{\beta}$ vary a lot, whereas $\frac{\hat{\beta}}{\hat{\alpha}}$ is more stable.

The table 4.2 shows how many times the selected partition has 2, 3 or 4 elements, and the table 4.3 gives the evaluation of the ratio between the risk of the PLSE and the oracle of the

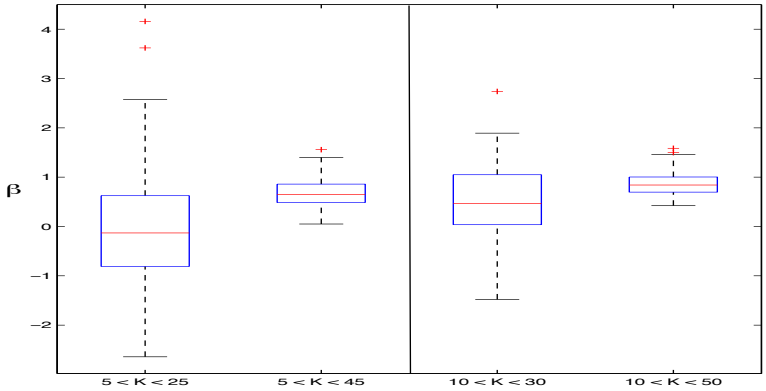

Figure 4.3: boxplot of $\hat{\beta}$ for samples with size $n = 100$.

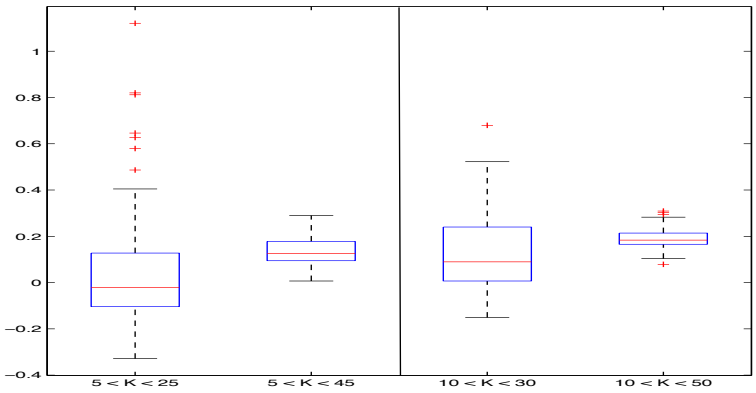

Figure 4.4: boxplot of $\frac{\hat{\beta}}{\hat{\alpha}}$ for samples with size $n = 100$.

n	50	100	200	400
$5 \leq K \leq 25$	23-24-4	25-61-12	4-87-7	0-87-12
$5 \leq K \leq 45$		18-65-13	2-86-8	0-87-10
$5 \leq K \leq 95$			0-77-14	0-81-12
$10 \leq K \leq 30$	13-30-18	19-62-13	2-80-10	0-86-10
$10 \leq K \leq 50$		16-64-12	0-80-11	0-81-9
$10 \leq K \leq 100$			0-71-15	0-78-12

Table 4.2: Performances of the first method: How many times the selected partition has 2, 3 or 4 elements

n	50	100	200	400
$5 \leq K \leq 25$	2.29	1.30	1.28	1.31
$5 \leq K \leq 45$		1.28	1.29	1.34
$5 \leq K \leq 95$			1.42	1.51
$10 \leq K \leq 30$	1.85	1.35	1.37	1.37
$10 \leq K \leq 50$		1.34	1.40	1.58
$10 \leq K \leq 100$			1.53	1.62

Table 4.3: Performances of the first method: Estimated ratio between the risk of the PLSE and the oracle of the collection $(S_K)_{K \geq 1}$

collection $(S_K)_{K \geq 1}$. For samples with size $n = 50$, we do not get a good result. For samples with size $n = 100$, the length of the series of values of K seems not to be very important, but it seems that we should rather begin with $K = 5$ rather than $K = 10$. For samples with size $n = 200$ or 400, the series of 21 values of K give better results than the series of 41 and 91 values. Like for $n = 100$, the minimal value of K should be close to the size of the ideal partition M^* (here $K^* = |M^*| = 3$).

The best results are obtained with $5 \leq K \leq 25$. This means that the series of values of K used for the fit should not be too large and that the minimal value K_{min} should be close to the size of the ideal partition M^*. The problem is that the number $|M^*|$ is unknown in practice. The bad result obtained with $n = 50$ may be explained by the fact that the series $5 \leq K \leq 25$ is large as compared to $n = 50$. But we can not take smaller series of K because we have to estimate 3 parameters α, β and γ.

The main weakness of this method is that we do not know which series of K should be taken for the fit. Series of 21 values should be preferred to larger series, but in practice we do not know which minimal value K_{min} should be taken.

4.5 Second method

We consider the following form of penalty:

$$\text{pen}(M) = \lambda \frac{|M|}{n} \left(\log \left(\frac{N_n}{|M|} \right) + c \right).$$

We determine the right constant c thanks to a study involving simulated extreme value variables $(Z_i)_{1\leq i\leq n}$ with mean 0 and variance 1, and then the multiplicative constant λ is chosen according to the data using the practical rule proposed by Birgé and Massart [4, section 4].

Remark 4.3 *The form of the penalty does not depend on the underlying distribution of the life-time, but the constant c is calibrated using simulated extreme value variables* $(Z_i)_{1\leq i\leq n}$. *Therefore the obtained c only suits for the Weibull ALT model.*

Recall that the theorem 4.1 recommends a penalty

$$\text{pen}(M) = (\sigma^2 + Rb)\frac{|M|}{n}\left(c_1 \log\left(\frac{N_n}{|M|}\right) + c_2\right)$$

with c_1 and c_2 absolute contants. But the theoretical result is not sharp enough to determine the right constants c_1 and c_2. The multiplicative constant $\sigma^2 + Rb$ depends on the variance τ^2 and on $R = \|\mu\|_\infty$.

In a similar case where the multiplicative constant is the variance τ^2 (instead of $\sigma^2 + Rb$), Lebarbier [16] proposes a method to calibrate the constants c_1 and c_2 by using a class of functions μ and simulated data with fixed and known variance τ^2. Here we can not apply the same method since the multiplicative constant depends on $\|\mu\|_\infty$. Moreover, we do not trust in the factor $\sigma^2 + Rb$ in the sense that it may rather be $\sigma^2 + \kappa Rb$ with $\kappa \neq 1$.

We make a simulation study with $\mu = 0$ and $\tau^2 = 1$ both fixed to determine the right constant $c^* = \frac{c_2^*}{c_1^*}$. The simulation study described in the next subsection involves simulated data corresponding to $\mu = 0$ and $\tau^2 = 1$. Since $\mu = 0$, we do not need to know whether the right multiplicative factor is $\sigma^2 + Rb$ or $\sigma^2 + 0.5Rb$ or something else.
Then, we consider a penalty of the form $\text{pen}(M) = \lambda\frac{|M|}{n}\left(\log\left(\frac{N_n}{|M|}\right) + c^*\right)$, and we use the practical rule proposed by Birgé and Massart [4, section 4] to determine the right constant λ according to the data.

4.5.1 A simulation study to determine the constant c.

In the first method described in section 4.4, we use Massart's heuristic to estimate the two constants α and β of the penalty according to the data. The problem to use this heuristic in practice is that we do not know from which value of K, $\mathbb{E}[\gamma_n(\mu_K)]$ is approximately constant. Here, we use Massart's heuristic only to determinate the constant c from simulated data corresponding to $\mu = 0$ and $\tau = 1$. For such data, $\mathbb{E}[\gamma_n(\mu_K)] = 1$ for all $K \geq 1$. Thus, we fit $-\gamma_n(\hat{\mu}_K)$ on $\frac{1}{2}\frac{K}{n}\left(\alpha \log\left(\frac{N_n}{K}\right) + \beta\right) + \gamma$ with $1 \leq K \leq K_{max}$. We only have to choose the maximal value K_{max}. Then we take c^* as the median of $\hat{c} = \frac{\hat{\beta}}{\hat{\alpha}}$ obtained via $Nsim = 100$ simulated samples with size n.

The figure 4.5 gives the boxplot of the values of $\hat{c}$ obtained via $N_{sim} = 100$ simulated samples with $n = 50, 100, 200, 400$ and $K_{max} = 15, 25, 50, 100, 200$ ($K_{max} \leq \frac{n}{2}$). For some fixed sample size n, the amplitude of the boxplot of $\hat{c}$ decreases with K_{max}.
The table 4.4 gives the median of $\hat{c}$ as a function of n and K_{max}. For some fixed n, the median of $\hat{c}$ increases with K_{max}. For $\frac{K_{max}}{n} \leq \frac{1}{4}$, whatever the sample size n, the median of $\hat{c}$ is close

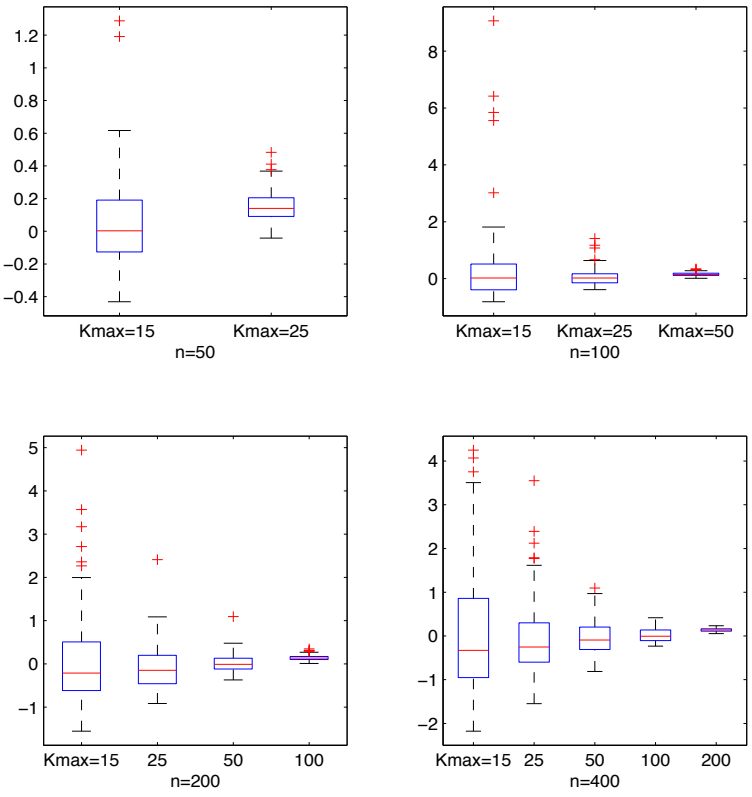

Figure 4.5: boxplot of $\hat{c}$ for $n = 50, 100, 200, 400$ and $K_{max} = 15, 25, 50, 100, 200$.

to 0. For $\frac{K_{max}}{n} = \frac{1}{2}$, the median of $\hat{c}$ is close to 0.14.
In figure 4.6, we compare the boxplots of $\hat{c}$ for fixed values of the ratio $\frac{K_{max}}{n}$ and various n. They have the same order of amplitude. For $\frac{K_{max}}{n} = \frac{1}{4}$, 0 belongs to the boxplot of $\hat{c}$ whatever n. For $\frac{K_{max}}{n} = \frac{1}{2}$, 0.14 belongs to the boxplot of $\hat{c}$ whatever n.

It remains to know whether we should take $K_{max} = \frac{n}{2}$ and therefore $c^* \simeq 0.14$, or $K_{max} \leq \frac{n}{4}$ and therefore $c^* \simeq 0$.

n	50	100	200	400
$1 \leq K \leq 15$	0.0032	0.0204	-0.2098	-0.3302
$1 \leq K \leq 25$	0.1400	0.0229	-0.1479	-0.2515
$1 \leq K \leq 50$		0.1440	-0.0088	-0.0920
$1 \leq K \leq 100$			0.1365	-0.0062
$1 \leq K \leq 200$				0.1382

Table 4.4: median of $\hat{c}$ obtained via 100 simulations

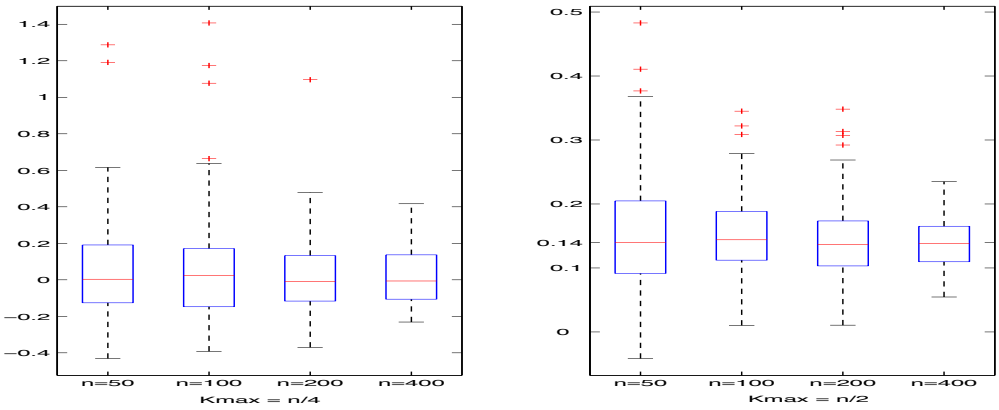

Figure 4.6: boxplot of $\hat{c}$ for $K_{max} = \frac{n}{4}$ and $\frac{n}{2}$.

4.5.2 A practical rule to determine λ according to the data.

In a Gaussian framework, Birgé and Massart [4] show that there exists a minimal penalty of the form:

$$\text{pen}_{min}(M) = \frac{\sigma^2}{n} D_M A(D_M)$$

with $A(D)$ which depends on $|\{M \in \mathcal{M}_n;\ D_M = D\}|$. More precisely, they prove that

- penalties of the form $\text{pen}(M) = (1+\eta)\frac{\sigma^2}{n} D_M A(D_M)$ with $\eta > 0$ lead to oracle type inequalities,
- if $\text{pen}(M) \leq (1-\eta)\frac{\sigma^2}{n} D_M A(D_M)$ for some $\eta > 0$ and D_M sufficiently large, then the risk of the PLSE can be arbitrarly large.

They also show that $\text{pen}(M) = 2\text{pen}_{min}(M)$ is always a reasonable penalty (non asymptotically) and sometimes an optimal one (asymptotically).
In practice, σ^2 is unknown and they consider penalties of the form

$$\text{pen}(M) = \lambda \frac{D_M}{n} A(D_M)$$

$\lambda = 2\sigma^2$ provides a good and sometimes nearly optimal penalty, and all $\lambda < \sigma^2$ lead to procedures which tend to choose a model of much too large dimension. Thus they suggest to estimate λ from the data by multiplying by 2 the first value for which the dimension of the selected model jumps to a smaller dimension.

Here the shape of the penalty $\text{pen}(M) = \lambda \frac{|M|}{n}\left(\log\left(\frac{N_n}{|M|}\right) + c*\right)$ is given by the theory and completed by a simulation study. c^* has been computed via a simulation study in subsection 4.5.1. The values of c^* are given in table 4.4. It remains to decide whether the right c^* is obtained with $K_{max} = \frac{n}{4}$ or $\frac{n}{2}$.
We consider

$$\hat{K}_\lambda = \arg \min_{1 \leq K \leq N_n} \left\{ \gamma_n(\hat{\mu}_K) + \lambda \frac{K}{n}\left(\log\left(\frac{N_n}{K}\right) + c*\right)\right\}$$

and we draw it as a function of λ. Then, we detect the value λ_{min} which corresponds to the largest jump. And we take $\hat{\lambda} = 2\lambda_{min}$.

Let $(Y_i)_{1 \leq i \leq n}$ a simulated sample with size $n = 100$ defined by (4.9), (4.10) and $Z_i = \sqrt{\frac{6}{\pi^2}}(S_i + \gamma)$ where $S_i \sim SEV(0,1)$. We consider penalties of the form $\text{pen}(M) = \lambda \frac{|M|}{n}\left(\log\left(\frac{N_n}{|M|}\right) + c^*\right)$ with c^* the median of the values of $\hat{c}$ obtained via the simulation study of section 4.5.1 with $n = 100$ and $1 \leq K \leq 50$, i.e. $c^* = 0.1440$. We give in figure 4.7 the graph of $\hat{K}_\lambda$ as a function of λ obtained with the simulated sample and $c^* = 0.1440$. In figure 4.7, we see that $\lambda_{min} = 2.51$, $\hat{\lambda} = 5.02$ and the selected number of subregions is 3 (which is the number K^* of subregions of the ideal partition M^*).

In order to evaluate the performances of this second method and compare it to the first one, we use $Nsim = 100$ simulated samples and we count how many times the ideal number

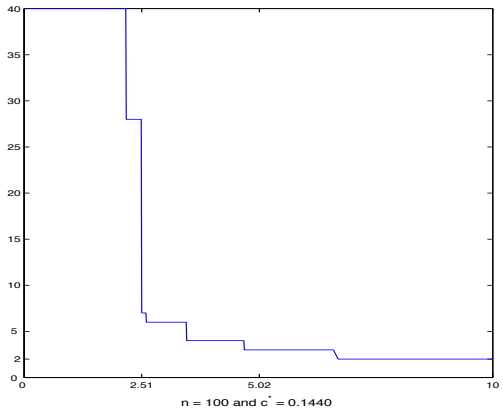

Figure 4.7: $\hat{K}_\lambda$ as a function of λ.

n	50	100	200	400
$c^* = 0$	0-0-0	28-61-9	0-78-13	0-77-14
$c^* = \text{median}(\hat{c}_{n,K_{max}=n/2})$	96-3-0	15-62-15	0-76-13	0-77-14

Table 4.5: Performances of the second method: How many times the selected partition has 2, 3 or 4 elements

of pieces is chosen. The ideal number is $K^* = 3$ when $n \geq 100$ and $K^* = 2$ when $n = 50$. The table 4.5 shows how many times the selected partition has 2, 3 or 4 elements. Thanks to the $Nsim = 100$ repetitions and a Monte Carlo method, we evaluate the ratio between the risk of the PLSE and the oracle (see table 4.6).

For $n \geq 100$, we get similar good results with $c^* = 0$ and $c^* \simeq 0.14$. For $n = 50$, we get a good result with $c^* \simeq 0.14$, but in this case $c^* = 0$ is not sufficient.

Figure 4.8 gives the values of $\hat{\lambda}$ obtained for the $Nsim = 100$ simulated samples of size $n = 100$ with $c^* = 0$ and 0.1440. This figure allows to understand the closeness of the performances of the first and second method for $n = 100$. With the first method and $5 \leq K \leq 25$, 0 belongs to the boxplot of $\frac{\hat{\beta}}{\hat{\alpha}}$ (see figure 4.4) and $\hat{\alpha}$ takes values close to those of $\hat{\lambda}$ when $c^* = 0$ (see figures 4.2 and 4.8). The method 1 with $5 \leq K \leq 25$ and the method 2 with $c^* = 0$ give similar penalty constants, and therefore similar results.

n	50	100	200	400
$c^* = 0$	3.60	1.32	1.42	1.64
$c^* = \text{median}(\hat{c}_{n,K_{max}=n/2})$	1.21	1.37	1.46	1.64

Table 4.6: Performances of the second method: Estimated ratio between the risk of the PLSE and the oracle of the collection $(S_K)_{K\geq 1}$

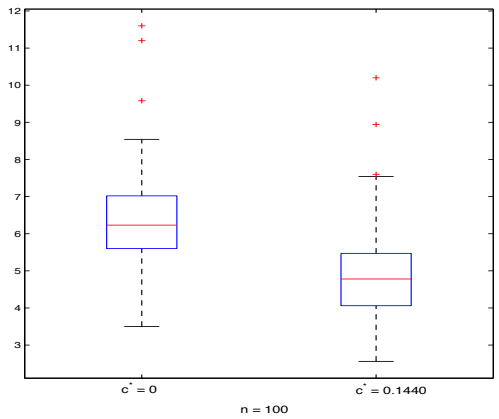

Figure 4.8: boxplot of $\hat{\lambda}$ obtained with $c^* = 0$ and 0.1440.

To sum up,

- For a small sample with size $n = 50$, the best result is obtained with the second method and $c^* = 0.14$ (which corresponds to $K_{max} = \frac{n}{2}$). The ideal size $K^* = 2$ is selected 96 times among 100 attempts. And the estimated ratio between the risk of the PLSE and the oracle is 1.21.

- For larger sample sizes $n \geq 100$, the second method give similar good results with $c^* = 0$ and $c^* = 0.14$, but the first method with $5 \leq K \leq 25$ give better results. It seems that the minimal value of K (here 5) has to be close to the ideal number of subregions (here 3) and that the length of the series of K has to be moderate. 21 values of K are sufficient. We see on table 4.2 and 4.3 for $n \geq 200$, that the series $5 \leq K \leq 25$ and $10 \leq K \leq 30$ give better results than the longer ones. In order to use the first method, we should complete it by a procedure which automatically determines the minimal value of K. Without such a procedure, we have to content ourself with the first method, whose results are still quite good.

Appendix A

MARS

L'algorithme MARS (Multivariate Adaptive Regression Splines) proposé par Friedman [12] construit un estimateur d'une fonction de régression s définie sur un ensemble $\mathcal{X} \subset \mathbb{R}^p$. Friedman [12] présente MARS comme une extension des méthodes de régression par construction récursive d'une partition (cf CART [7]). Ces méthodes sont notées ici méthodes RPR (Recursive Partitioning Regression). Contrairement aux méthodes RPR, MARS construit un estimateur continu (et même de classe $\mathcal{C}^1$), et permet d'approcher les fonctions additives ou plus généralement les fonctions qui s'écrivent comme des sommes de termes ne faisant intervenir qu'un petit nombre de variables.

Commençons par décrire les méthodes RPR. Ces méthodes construisent des estimateurs $\hat{s}$ de la forme:

$$\text{si } \mathbf{x} \in R_k \text{ alors } \hat{s}(\mathbf{x}) = a_k$$

où les régions $(R_k)_{1 \leq k \leq K}$ forment une partition de $\mathcal{X}$ et les coefficients $(a_k)_{1 \leq k \leq K} \in \mathbb{R}^K$. Elles utilisent les données $(\mathbf{x}_i, Y_i)_{1 \leq i \leq n}$, où $\mathbf{x}_i = (x_i^1, \ldots, x_i^p) \in \mathcal{X}$ et $Y_i \in \mathbb{R}$, pour choisir la partition de $\mathcal{X}$ et estimer les coefficients.

La création de la partition se fait de manière récursive dyadique. La partition initiale est constituée d'une seule région: l'ensemble $\mathcal{X}$ tout entier. A chaque étape, on découpe en deux les régions de la partition existante. Les découpages possibles d'une région R en deux régions filles R_g et R_d sont les découpages définis comme suit:

$$\begin{aligned} R_g &= \left\{\mathbf{x} = (x^1, \ldots, x^p) \in R;\ x^j \leq t\right\} \\ R_d &= \left\{\mathbf{x} = (x^1, \ldots, x^p) \in R;\ x^j > t\right\} \end{aligned}$$

Un découpage est donc défini par une variable x^j et une valeur seuil t. Il est choisi de façon à optimiser l'adéquation aux données. On poursuit les découpages jusqu'à obtenir une partition fine constituée d'un grand nombre de régions. Dans CART, on poursuit les découpages jusqu'à ce que toutes les régions soient pures ou ne contiennent chacune qu'une seule observation. Cette construction est naturellement représentée par un arbre de profondeur maximale. Certaines régions sont ensuite recombinées en supprimant des découpages. Cela revient à élaguer l'arbre maximal pour obtenir un arbre moins profond.
La figure A.1 donne un exemple d'arbre ainsi construit.

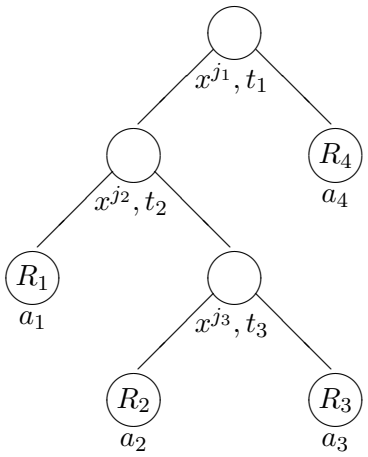

Figure A.1: Arbre binaire représentant un estimateur $\hat{s}$ obtenu par une méthode RPR

Sur l'exemple de la figure A.1, $R_2 = \left\{\mathbf{x} = (x^1, \ldots, x^p) \in \mathcal{X};\ x^{j_1} \leq t_1,\ x^{j_2} > t_2 \text{ et } x^{j_3} \leq t_3\right\}$, et si $\mathbf{x} \in R_2$ alors $\hat{s}(\mathbf{x}) = a_2$. Cette représentation géométrique de $\hat{s}$ facilite la lecture et l'interprétation.

Nous pouvons aussi donner une écriture analytique de $\hat{s}$ sous forme d'un développement par rapport à une famille de fonctions de base:

$$\hat{s}(\mathbf{x}) = \sum_{k=1}^{K} a_k B_k(\mathbf{x})$$

Les fonctions de base B_k sont les fonctions indicatrices des régions R_k: $B_k(\mathbf{x}) = \mathbb{1}_{R_k}(\mathbf{x})$. Les coefficients a_k sont déterminés de façon à optimiser l'adéquation aux données.
Notons γ_n le contraste des moindres carrés associé à l'échantillon d'observations $(\mathbf{x}_i, Y_i)_{1 \leq i \leq n}$:

$$\gamma_n(u) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - u(\mathbf{x}_i))^2 \quad \text{pour } u : \mathcal{X} \to \mathbb{R}.$$

Notons

$$\begin{aligned} H^+(\eta) &= \mathbb{1}_{\eta > 0} \\ H^-(\eta) &= \mathbb{1}_{\eta \leq 0} \end{aligned}$$

La construction de la partition fine $(R_k)_{1 \leq k \leq K_{max}}$ ou de la liste initiale de fonctions $(B_k)_{1 \leq k \leq K_{max}}$ se fait par l'algorithme suivant:

Algorithme 1 (1ère étape de la méthode RPR)

$B_1(\mathbf{x}) \leftarrow 1$
For $K = 2$ to K_{max} do: $lof^* \leftarrow \infty$
  For $k = 1$ to $K-1$ do:
    For $j = 1$ to p do:
      For $t \in \{x_i^j;\ B_k(\mathbf{x}_i) > 0\}$ do:

$$u \leftarrow \sum_{\substack{l=1 \\ l \neq k}}^{K-1} a_l B_l(\mathbf{x}) + a_k B_k(\mathbf{x}) H^+(x^j - t) + a_K B_k(\mathbf{x}) H^-(x^j - t)$$

$$lof \leftarrow \min_{a_1,\ldots,a_K} \gamma_n(u)$$

        if $lof < lof^*$ then $lof^* \leftarrow lof,\ k^* \leftarrow k,\ j^* \leftarrow j,\ t^* \leftarrow t$ endif
      endfor
    endfor
  endfor
  $B_K(\mathbf{x}) \leftarrow B_{k^*}(\mathbf{x}) H^-(x^{j^*} - t^*)$
  $B_{k^*}(\mathbf{x}) \leftarrow B_{k^*}(\mathbf{x}) H^+(x^{j^*} - t^*)$
endfor

La première ligne de l'algorithme 1 revient à définir la région initiale comme étant l'ensemble $\mathcal{X}$ tout entier. La première boucle itère les découpages jusqu'à ce que l'on ait une partition en K_{max} régions. (Rappelons que dans CART, les découpages se poursuivent jusqu'à ce que toutes les régions soient pures ou ne contiennent chacune qu'une seule observation.) Les 3 boucles suivantes déterminent une fonction B_{k^*}, une variable x^{j^*} et une valeur seuil t^* telles que le découpage de la région R_{k^*} par "$x^{j^*} \leq t^*$" et "$x^{j^*} > t^*$" permet la meilleure adéquation aux données. La fonction B_{k^*} est alors remplacée par son produit avec $H^+(x^{j^*} - t^*)$, et on ajoute la fonction $B_K(\mathbf{x}) = B_{k^*}(\mathbf{x}) H^-(x^{j^*} - t^*)$.
Cet algorithme correspond à la première étape d'une méthode RPR. Il faut ensuite faire une étape dite d'élagage que nous ne détaillons pas ici (voir CART [7]).

Les fonctions de base construites par les méthodes RPR sont de la forme:

$$B_k(\mathbf{x}) = \prod_{l=1}^{N_k} H^{s_{kl}} \left(x^{v(k,l)} - t_{kl} \right)$$

où N_k est le nombre de découpages qui ont permis de définir R_k et donc B_k. Pour le $l^{\text{ième}}$ découpage, $v(k,l)$ est le numéro de la variable sur laquelle porte le découpage, t_{kl} est la valeur seuil du découpage et $s_{kl} = \pm$ selon que $R_k \subset \{\mathbf{x};\ x^{v(k,l)} > t_{kl}\}$ ou $R_k \subset \{\mathbf{x};\ x^{v(k,l)} \leq t_{kl}\}$.

Les méthodes RPR ont deux inconvénients:

1. elles produisent des estimateurs fortement discontinus,
2. elles ne permettent pas d'approcher les fonctions additives ou les fonctions dont les interactions sont d'ordre petit.

La discontinuité vient de l'utilisation des fonctions $H^{\pm}(x^j - t)$ qui sont des splines d'ordre 0. En les remplaçant par les splines d'ordre 1: $[\pm(x^j - t)]_+$, on obtient des fonctions de base continues puis des estimateurs continus. Avec les splines d'ordre 0: $H^{\pm}(x^j - t)$, les fonctions de base B_k sont des produits tensoriels de splines d'ordre 0. Avec les splines d'ordre 1: $[\pm(x^j - t)]_+$, pour que les fonctions de base B_k soient des produits tensoriels de splines d'ordre 1, il faudrait que chaque variable x^j apparaisse au plus une fois dans le produit définissant B_k. On ne peut malheureusement pas interdire de découper plusieurs fois sur la même variable.
Le deuxième inconvénient cité ci-dessus vient du fait que l'on enlève les fonctions faisant intervenir un petit nombre de variables et qu'à la sortie de l'algorithme les fonctions B_k font intervenir beaucoup de variables. En effet, à chaque étape de l'algorithme 1, on remplace une fonction par deux nouvelles fonctions dont le niveau d'interaction est augmenté de 1 (sauf si le découpage se fait sur une variable qui apparaissait déjà). Lors d'un découpage, au lieu de remplacer la fonction "mère" par ses deux fonctions "filles", on peut garder la fonction "mère" et simplement ajouter les deux nouvelles fonctions. On perd alors la réprésentation sous forme d'arbre, mais on peut ainsi espérer approcher les fonctions additives et les fonctions dont les interactions sont d'ordre petit. En faisant cette deuxième modification, on peut maintenant imposer que les facteurs d'une même fonction de base fassent intervenir des variables distinctes.

Friedman propose donc 3 modifications:

1. remplacer $H^{\pm}(x - t)$ par $[\pm(x - t)]_+$,

2. lors d'une découpe, ne pas enlever la fonction choisie B_{k^*}, qui devient donc elligible comme ses filles pour les découpes suivantes,

3. imposer que tous les facteurs d'une même fonction de base fassent intervenir des variables distinctes.

Avec ces 3 modifications, on obtient la première partie de l'algorithme MARS, notée ci-dessous algorithme 2.

Algorithme 2 (1ère partie de l'algorithme MARS)

$B_1(\mathbf{x}) \leftarrow 1;\ K \leftarrow 2$
While $K < K_{max}$ do: $lof^* \leftarrow \infty$
  For $k = 1$ to $K-1$ do:
    For $j \notin \{v(l,k);\ 1 \le l \le N_k\}$ do:
      For $t \in \{x_i^j;\ B_k(\mathbf{x}_i) > 0\}$ do:
        $u \leftarrow \sum_{l=1}^{K-1} a_l B_l(\mathbf{x}) + a_K B_k(\mathbf{x})[(x^j - t)]_+ + a_{K+1} B_k(\mathbf{x})[-(x^j - t)]_+$
        $lof \leftarrow \min_{a_1,\ldots,a_{K+1}} \gamma_n(u)$
        if $lof < lof^*$ then $lof^* \leftarrow lof,\ k^* \leftarrow k,\ j^* \leftarrow j,\ t^* \leftarrow t$ endif
      endfor
    endfor
  endfor
  $B_K(\mathbf{x}) \leftarrow B_{k^*}(\mathbf{x})[(x^{j^*} - t^*)]_+$
  $B_{K+1}(\mathbf{x}) \leftarrow B_{k^*}(\mathbf{x})[-(x^{j^*} - t^*)]_+$
  $K \leftarrow K + 2$
endfor

La deuxième partie de l'algorithme MARS (notée ici algorithme 3) consiste à éliminer les fonctions de base qui n'ont pas apporté une amélioration suffisante. Friedman utilise pour cela un critère qui pénalise les modèles associés à un grand nombre de fonctions de base. A l'issue de la première partie de MARS (algorithme 2), on dispose d'une famille de fonctions de base: $\mathcal{F} = \{B_1, B_2, \ldots, B_{K_{max}}\}$. Nous notons $\mathcal{M}_n$ la collection de toutes les sous-familles de $\mathcal{F}$:

$$\mathcal{M}_n = \{m = (B_k)_{k\in\mathcal{N}};\ \mathcal{N} \subset \{1, 2, \ldots, K_{max}\}\},$$

et à chaque $m = (B_k)_{k\in\mathcal{N}} \in \mathcal{M}_n$, nous associons S_m l'espace vectoriel engendré par les fonctions $(B_k)_{k\in\mathcal{N}}$, et $\hat{s}_m$ l'estimateur des moindres carrés de s sur le modèle S_m. Friedman propose le critère suivant:

$$\mathrm{crit}_F(m) = \frac{\gamma_n(\hat{s}_m)}{\left(1 - \frac{D_m + c|m|}{n}\right)^2}$$

où D_m est la dimension du modèle S_m, $|m|$ est le cardinal de la famille m et c est une constante à choisir.
A l'aide de ce critère, l'algorithme 3 ci-dessous sélectionne une sous-famille $(B_j)_{j\in J^*}$ de $\mathcal{F}$.

Algorithme 3 (2ème partie de l'algorithme MARS)

$J^* \leftarrow \{1, 2, \ldots, K_{max}\}$; $L^* \leftarrow J^*$;
$lof^* \leftarrow \mathrm{crit}_F\left((B_j)_{j\in J^*}\right)$
for $K = K_{max}$ to 2 do: $b \leftarrow \infty; L \leftarrow L^*$
  For $k = 2$ to K do: $I \leftarrow L - \{k\}$
    $lof \leftarrow \mathrm{crit}_F\left((B_i)_{i\in I}\right)$
    if $lof < b$, then $b \leftarrow lof; L^* \leftarrow I$ endif
    if $lof < lof^*$, then $lof^* \leftarrow lof; J^* \leftarrow I$ endif
  endfor
endfor

A l'issue de cette deuxième partie, on dispose d'une nouvelle liste (plus petite) de fonctions de base $(B_j)_{j\in J^*}$. Ces fonctions B_j sont des produits de fonctions de la forme $b(x|s,t) = [s(x-t)]_+$. Pour obtenir un estimateur de classe $\mathcal{C}^1$, Friedman remplace ces fonctions par:

$$C(x|s=+,t_-,t,t_+) = \begin{cases} 0 & \text{si } x \le t_- \\ p_+(x-t_-)^2 + r_+(x-t_-)^3 & \text{si } t_- < x < t_+ \\ x - t & \text{si } x \ge t_+ \end{cases}$$

$$C(x|s=-,t_-,t,t_+) = \begin{cases} -(x-t) & \text{si } x \le t_- \\ p_-(x-t_+)^2 + r_-(x-t_+)^3 & \text{si } t_- < x < t_+ \\ 0 & \text{si } x \ge t_+ \end{cases}$$

où $t_- < t < t_+$ et

$$\begin{aligned} p_+ &= (2t_+ + t_- - 3t)/(t_+ - t_-)^2, \\ r_+ &= (2t - t_+ - t_-)/(t_+ - t_-)^3, \\ p_- &= (3t - 2t_- - t_+)/(t_- - t_+)^2, \\ r_- &= (t_- + t_+ - 2t)/(t_- - t_+)^3. \end{aligned}$$

Les deux valeurs seuil supplémentaires t_- et t_+ sont placées de façon à réduire les discontinuités des dérivées d'ordre 2.
Finalement on obtient une famille de fonctions $(\tilde{B}_j)_{j\in J^*}$ et un estimateur $\hat{s} = \sum_{j\in J^*} a_j \tilde{B}_j$ de classe $\mathcal{C}^1$, où les coefficients a_j sont obtenus en minimisant le critère des moindres carrés.

Bibliography

[1] Y. Baraud. Model selection for regression on a fixed design. *Probability Theory and Related Fields*, 117:467–493, 2000.

[2] Y. Baraud, F. Comte, and G. Viennet. Model Selection for (auto-)regression with dependent data. *ESAIM Probability and Statistics*, 5:33–49, 2001.

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

[4] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. To be published in Probability Theory and Related Fields, 2005.

[5] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C.R. Math. Acad. Sci. Paris*, 334:495–500, 2002.

[6] O. Bousquet. Concentration Inequalities for Sub-Additive Functions Using the Entropy Method. *Stochastic Inequalities and Applications*, 56:213–247, 2003.

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

[8] G. Castellan. Modified Akaike's criterion for histogram density estimation. *C.R. Acad. Sci. Paris Sér. I Math.*, 330(8):729–732, 2000.

[9] G. Castellan. Density estimation via exponential model selection. *IEEE Transactions on information theory*, 49(8):2052–2060, 2003.

[10] F. Comte and Y. Rozenholc. A new algorithm for fixed design regression and denoising. *Ann. Inst. Statist. Math.*, 56(3), 2004.

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

[12] J. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19:1:141, 1991.

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

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

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

[16] E. Lebarbier. *Quelques approches pour la détection de ruptures à horizon fini*. PhD thesis, Université Paris XI Orsay, 2002.

[17] C.L. Mallows. Some comments on C_p. *Technometrics*, 15:661:675, 1973.

[18] P. Massart. Some Applications of Concentration Inequalities to Statistics. *Annales de la Faculté des Sciences de Toulouse*, 9(2):245–303, 2000.

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

[20] P. Massart and E. Nédélec. Risk bounds for statistical learning. *Annals of Statistics*, 34(5), 2006.

[21] J.M. Poggi and C. Tuleau. Classification supervisée en grande dimension. Application à l'agrément de conduite automobile. *Preprint Université Paris XI Orsay*, 2005.

[22] M. Sauvé. Histogram selection in non gaussian regression. Research Report 5911, INRIA, 2006.

[23] M. Sauvé and C. Tuleau. Variable selection through CART. Research Report 5912, INRIA, 2006.

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

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

Résumé. Cette thèse traite de la sélection de modèles en régression non gaussienne. Notre but est d'obtenir des informations sur une fonction $s : \mathcal{X} \to \mathbb{R}$ dont on ne connaît qu'un certain nombre de valeurs perturbées par des bruits non nécessairement gaussiens. Dans un premier temps, nous considérons des modèles de fonctions constantes par morceaux associés à une collection de partitions de $\mathcal{X}$. Nous déterminons un critère des moindres carrés pénalisés qui permet de sélectionner une partition dont l'estimateur associé (de type regressogramme) vérifie une inégalité de type oracle. La sélection d'un modèle de fonctions constantes par morceaux ne conduit pas en général à une bonne estimation de s, mais permet notamment de détecter les ruptures de s. Nous proposons aussi une méthode non linéaire de sélection de variables qui repose sur l'application de plusieurs procédures CART et sur la sélection d'un modèle de fonctions constantes par morceaux. Dans un deuxième temps, nous considérons des modèles de fonctions polynomiales par morceaux, dont les qualités d'approximation sont meilleures. L'objectif est d'estimer s par un polynôme par morceaux dont le degré peut varier d'un morceau à l'autre. Nous déterminons un critère pénalisé qui sélectionne une partition de $\mathcal{X} = [0,1]^p$ et une série de degrés dont l'estimateur polynomial par morceaux associé vérifie une inégalité de type oracle. Nous appliquons aussi ce résultat pour détecter les ruptures d'un signal affine par morceaux. Ce dernier travail est motivé par la détermination d'un intervalle de stress convenable pour les tests de survie accélérés.

Mots-clés. sélection de modèles, sélection de variables, détection de ruptures, estimation polynomiale, inégalités de concentration, regression, CART, tests de survie.

Abstract. This thesis deals with model selection in non Gaussian regression. Our aim is to get informations on a function $s : \mathcal{X} \to \mathbb{R}$ given only some values perturbed by noises non necessarily Gaussian. In a first part, we consider histogram models (i.e. classes of piecewise constant functions) associated with a collection of partitions of $\mathcal{X}$. We determine a penalized least squares criterion which selects a partition whose associated estimator satisfies an oracle inequality. Selecting a histogram model does not always lead to an accurate estimation of s, but allows for example to detect the change-points of s. In order to perform variable selection, we also propose a non linear method which relies on the use of CART and on histogram model selection. In a second part, we consider piecewise polynomial models, whose approximation properties are better. We aim at estimating s with a piecewise polynomial whose degree can vary from region to region. We determine a penalized criterion which selects a partition of $\mathcal{X} = [0,1]^p$ and a series of degrees whose associated piecewise polynomial estimator satisfies an oracle inequality. We also apply this result to detect the change points of a piewise affine signal. The aim of this last work is to provide an adequate stress interval for Accelerating Life Test.

Keywords. model selection, variable selection, change-points detection, polynomial estimation, concentration inequalities, regression, CART, Accelerating Life Test.