Université Paris Descartes
U.F.R. de Mathématiques et Informatique

Estimation par sélection de modèles à partir de données partiellement observées.

Sandra PLANCADE

Table des matières

Introduction

Chapitre 1

Introduction

1.1 Statistiques non paramétriques

Considérons l'estimation d'une application f à partir d'un échantillon $(V_1, \ldots, V_n)$ de variables ou de vecteurs aléatoires. Un estimateur de f est une application $\widehat{f}$ entièrement déterminée par la donnée de l'échantillon $(V_1, \ldots, V_n)$. Dans le cadre des statistiques non paramétriques, on ne suppose aucune forme a priori sur la fonction f à estimer. Néanmoins, des hypothèses générales sur f peuvent être nécessaires (f bornée, à support compact, dérivable...).

La performance d'un estimateur $\widehat{f}$ est mesurée par une fonction de perte, aussi appelée risque, de la forme $\mathbb{E}[d(\widehat{f}, f)]$ ou $d(.,.)$ est une distance ou une semi-distance sur l'ensemble des fonctions. Les risques les plus classiques sont le risque L^p où $d(s,t) = \|s-t\|_p^p$ est la distance associée à la norme L^p, et le risque ponctuel d'ordre p où $d(s,t) = |(s-t)(x_0)|^p$ et x_0 est un point fixé. On peut également considérer une distance d qui dépend des observations. Ainsi, la performance d'un estimateur dépend du risque considéré, et les méthodes d'estimation développées sont différentes selon le risque auquel on s'intéresse.

Plus précisément, deux types de risque sont considérés dans ce manuscrit.

(i) Le risque quadratique intégré : $\mathbb{E}\left[\|\widehat{f} - f\|_\nu^2\right]$ où ν est une fonction à valeurs positives et $\|t\|_\nu^2 = \int t^2(x)\nu(x)dx$.

(ii) Le risque quadratique ponctuel : $\mathbb{E}\left[(\widehat{f} - f)^2(x_0)\right]$ où x_0 est un point de I fixé.

On notera

$$\|t\|_0 = \begin{cases} \|t\|_\nu \text{ dans le cas (i)} \\ \\ |t(x_0)| \text{ dans le cas (ii)} \end{cases} \tag{1.1}$$

Remark 1 *En alternative au risque (i), on peut considérer son équivalent empirique,*

$$\mathbb{E}\left[\|\widehat{f}-f\|_n^2\right] \quad avec \quad \|t\|_n^2=\frac{1}{n}\sum_{i=1}^{n}t^2(X_i)$$

où X_i est un vecteur composé de coordonnées de V_i. En effet, $\|t\|_n$ est la moyenne empirique associée à $\|t\|_{f_X}$ où f_X est la densité de X_i et sous certaines conditions ces deux normes sont équivalentes sur un ensemble de forte probabilité.

On distingue deux catégories principales d'estimateurs non paramétriques: les estimateurs construits à partir d'un noyau, brièvement présentés en Section 1.1.1, et les estimateurs construits par minimisation d'un contraste, qui est la méthode utilisée dans ce manuscrit.

1.1.1 Estimateurs à noyaux

Considérons l'exemple de l'estimation de densité. Soit $(X_1,\dots,X_n)$ des variables aléatoires i.i.d. (indépendantes identiquement distribuées) de densité f à support dans $I\subset\mathbb{R}$. Considérons l'estimateur intuitif de f suivant. Soit $x\in I$, $h>0$ et $[x-h,x+h]$ un petit intervalle autour de x,

$$\widehat{f}_h^0(x)=\frac{1}{2nh}Card\{i,X_i\in[x-h,x+h]\}=\frac{1}{nh}\sum_{i=1}^{n}K_0\left(\frac{X_i-x}{h}\right) \tag{1.2}$$

où $K_0(x)=(1/2)1_{[-1,1]}$. K_0 est appelé le noyau et h la fenêtre de l'estimateur $\widehat{f}_h^0$ (cf Figure 1.1)

Par ailleurs, on remarque que

$$f(x)\simeq\frac{1}{2h}\int K_0\left(\frac{x-t}{h}\right)f(t)dt=\mathbb{E}\left[\frac{1}{2h}K_0\left(\frac{X_1-t}{h}\right)\right] \tag{1.3}$$

si h est petit. Ainsi $\widehat{f}_h^0$ est la moyenne empirique associée à $(1/h)K_0\left((X_1-t)/h\right)$.

Le choix de la fenêtre h dans le calcul de l'estimateur $\widehat{f}_h$ est déterminant.

$\star$ Si h est trop grand, l'erreur d'approximation dans (1.3) est trop importante.

$\star$ D'après l'expression (1.2), si h est trop petit, il n'y a pas suffisamment de valeurs $\{X_i\}$ dans l'intervalle $[x-h,x+h]$ pour obtenir une bonne estimation.

La fenêtre optimale h_n qui réalise un compromis entre ces deux erreurs d'estimation dépend de n et tend vers zéro quand n tend vers l'infini. En effet, si h est fixé, l'erreur d'approximation dans (1.3) est fixée et le nombre de valeurs $\{X_i\}$ dans l'intervalle $[x-h,x+h]$ augmente, donc l'erreur d'approximation dans (1.2) diminue. Ainsi, plus n est grand, plus il sera intéressant de considerer une fenêtre petite.

La définition de l'estimateur (1.2) se généralise en considérant d'autres formes de noyaux. Plus précisément, on appelle noyau une application $K:\mathbb{R}\to\mathbb{R}$ intégrable, qui vérifie $\int_{\mathbb{R}}K(u)du=1$. La Figure 1.2 présente quelques exemples de noyaux classiques.

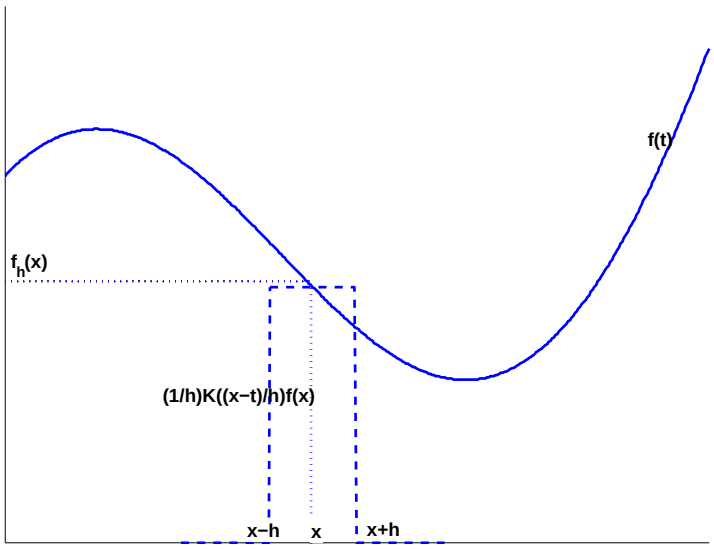
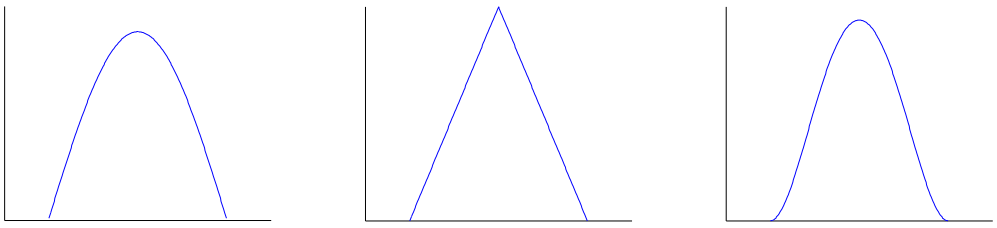


Figure 1.1: Estimateur à noyau



$\frac{15}{16}(1-x^2)^2 1(|x| \leq 1)$ $\quad$ $(1-|x|)1(|x| \leq 1)$ $\quad$ $\frac{1}{\sqrt{2\pi}} \exp(-x^2/2) 1(|x| \leq 1)$

Figure 1.2: Exemples de noyaux classiques

L'estimateur associé au noyau K et de fenêtre h est

$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right).$$

La méthode d'estimation par noyaux, présentée ici dans le cadre de l'estimation de densité, est très générale et s'applique dans de nombreux problèmes d'estimation. Elle est partic-

ulièrement adaptée à l'étude de risques ponctuels. Par ailleurs, la fenêtre peut-être choisie en fonction des données, notamment à l'aide d'une procédure developpée par Lepski and Spokoiny (1997). Nous reviendrons sur ce point dans la Section 1.2.7.

1.1.2 Estimation par minimisation d'un contraste

a) Construction d'un contraste.

Soit f une fonction à estimer à partir d'un échantillon $(V_1, \ldots, V_n)$ de variables ou de vecteurs aléatoires, et $\mathcal{F}$ un sous-ensemble des fonctions de I dans $\mathbb{R}$ tel que $f \in \mathcal{F}$. Un constraste empirique est une application

$$\gamma_n : t \in \mathcal{F} \to \gamma_n(t) \in \mathbb{R}$$

entièrement déterminée par les observations. L'estimateur considéré est la fonction $\widehat{f}$ qui minimise $\gamma_n(t)$ sur un ensemble de fonctions à déterminer. On considère essentiellement deux types de contraste empirique γ_n : les constrastes construits par maximum de vraisemblance, et les contrastes de type projection (les deux pouvant coïncider dans certains cas).

Les estimateurs présentés dans ce manuscrit sont construits à partir de contrastes de type projection. On considère une application $\gamma : t \in \mathcal{F} \to \gamma(t) \in \mathbb{R}$ telle que

$$f = \arg\min_{t \in \mathcal{F}} \gamma(t).$$

Plus précisément, les contrastes considérés dans ce manuscrit sont de la forme suivante :

$$\gamma(t) = \|t - f\|_\nu^2 := \int_I (t-f)^2(x)\nu(x)dx$$

où ν est une fonction à valeurs positives, définie sur $\mathcal{F} = L^2(I)$. On constate que pour tout $t \in \mathcal{F}$, $\gamma(t) \geq 0$ et $\gamma(f) = 0$, donc f minimise bien γ sur $\mathcal{F}$. L'application γ, inconnue (car elle dépend de f) est estimée par un contraste empirique γ_n. Plus précisément, γ_n est une application de $\mathcal{F}$ dans $\mathbb{R}$ uniquement déterminée par les observations et de la forme suivante.

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mu(V_i, t)$$

avec $\mathbb{E}[\mu(V_i, t)] = \gamma(t) + c_0$, et c_0 est une constante indépendante de V_i et t (et généralement dépendante de f). On remarque que

$$f = \arg\min_{t \in \mathcal{F}} \gamma(t) = \arg\min_{t \in \mathcal{F}} \gamma(t) + c_0.$$

Soit S un sous-ensemble de $\mathcal{F}$, on définit l'estimateur $\widehat{f}_S$ suivant.

$$\widehat{f}_S = \arg\min_{t \in S} \gamma_n(t). \tag{1.4}$$

$\widehat{f}_S$ n'est calculable que sous certaines conditions, la plus usuelle étant de supposer que l'espace S est de dimension finie. On se restreint donc à un ensemble $S = Vect\{\phi_1, \ldots, \phi_D\}$ où $(\phi_1, \ldots, \phi_D)$ est une base orthonormée de S pour la norme L^2.

b) Décomposition biais-variance

Tout comme le choix de la fenêtre h pour les estimateurs à noyaux, le choix de la dimension D est déterminant.

$\star$ Si D est trop petit, le modèle S n'est pas assez riche et $\widehat{f}_S$ ne pourra pas approcher correctement f.

$\star$ Si D est trop grand, le nombre de paramètres à estimer (c'est à dire le nombre de coefficients de $\widehat{f}_D$ dans la base $\{\phi_1, \ldots, \phi_D\}$) est trop élevé.

La Figure 1.3 illustre ce phénomène dans le cadre de l'estimation de densité, en considérant des espaces $S_D = Vect\{\cos(\pi k x/6), k = 0, \ldots, D\}$. Si D est petit, l'ensemble S_D n'approche pas suffisamment f, et si D est trop grand, l'estimateur s'adapte trop aux fluctuations des données.

Cette heuristique est confirmée par l'étude du risque de l'estimateur $\widehat{f}_S$,

$$\mathbb{E}\left[\|[\widehat{f}_S - f\|_0^2\right]$$

où $\|.\|_0^2$ est la semi-norme sur $\mathcal{F}$ définie en (1.1). Soit

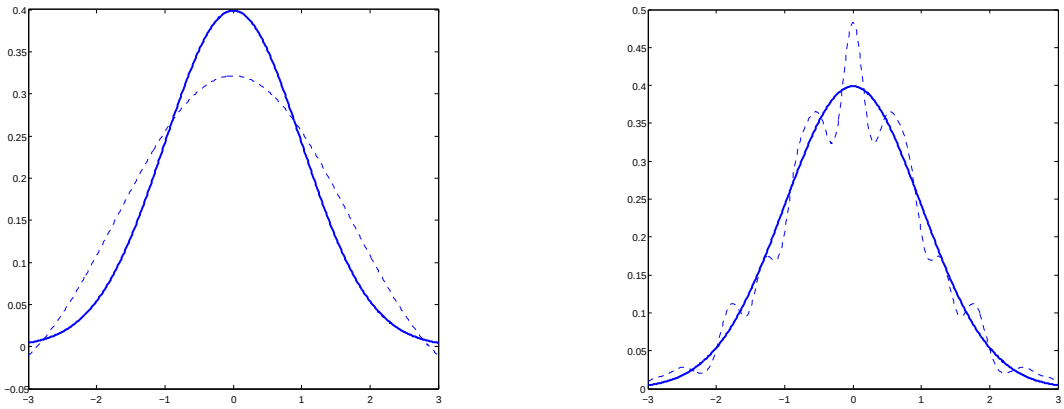
$$f_S(x) = \arg\min_{t \in S} \gamma(t) = \arg\min_{t \in S} \|f - t\|_\nu^2.$$

Alors, le risque de l'estimateur $\widehat{f}_S$ se décompose en deux termes appelés biais et variance.

$$\mathbb{E}\left[\|\widehat{f}_S - f\|_0^2\right] \leq \underbrace{2\|f - f_S\|_0^2}_{\text{biais}} + \underbrace{2\mathbb{E}\left[\|\widehat{f}_S - f_S\|_0^2\right]}_{\text{variance}}.$$

Cette décomposition fournit une majoration du risque de l'estimateur $\widehat{f}_S$, et il sera pertinent par la suite de s'interroger sur l'optimalité de cette majoration. La notion de vitesse minimax présentée en Section 1.2.2 permet une étude précise de cette question, mais on peut déjà constater que dans certains cas cette inégalité devient une égalité, au facteur 2 près. En effet, considérons le risque $\|.\|_\nu^2$. D'après le Théorème de Pythagore,

$$\mathbb{E}\left[\|\widehat{f}_S - f\|_\nu^2\right] = \underbrace{\|f_S - f\|_\nu^2}_{\text{biais}} + \underbrace{\mathbb{E}\left[\|\widehat{f}_S - f_S\|_\nu^2\right]}_{\text{variance}}. \tag{1.5}$$



S trop petit S trop grand

Figure 1.3: Fonction f à estimer (en trait plein) et estimateur $\widehat{f}_S$ (en pointillés).

L'expression "variance" fait référence à la situation où $\mathbb{E}\left[\widehat{f}_S(x)\right] = f_S(x)$ pour tout x. Soit $S = Vect(\phi_1, \ldots, \phi_D)$ où les (ϕ_k) sont orthonormés pour la norme $\|.\|_\nu$, alors

$$\widehat{f}_S = \sum_{k=1}^{D} \widehat{a}_k \phi_k \quad \text{et} \quad f_S = \sum_{k=1}^{D} a_k \phi_k \quad \text{avec} \quad \mathbb{E}[\widehat{a}_k] = a_k.$$

D'où

$$\mathbb{E}\left[\|\widehat{f}_S - f_S\|_\nu^2\right] = \mathbb{E}\left[\sum_{k=1}^{D} (\widehat{a}_k - a_k)^2\right] = Var\left(\sum_{k=1}^{D} \widehat{a}_k\right).$$

Si f_S n'est pas toujours égal à l'espérance de $\widehat{f}_S$, il est néanmoins assez proche car

$$\begin{cases} \gamma(t) + c_0 = \mathbb{E}\left[\gamma_n(t)\right], \quad \forall t \in S \\ \widehat{f}_S = \arg\min_{t \in S} \gamma_n(t) \\ f_S = \arg\min_{t \in S} \gamma(t) + c_0. \end{cases}$$

Plus l'espace S est grand, plus le terme de biais, égal à la distance de f à S, est petit. Par ailleurs, on verra dans la section suivante que le terme de variance augmente lorsque la

dimension de S augmente. Un "bon" espace S doit donc réaliser un compromis entre ces deux éléments et la procédure de sélection de modèle permet de choisir cet espace S en fonction des observations.

1.2 Sélection de modèles

1.2.1 Principe général

Considérons une collection de sous-espaces vectoriels de $\mathcal{F}$, appelés modèles

$$\mathcal{M}_n = \{S_m, m \in J_n\}.$$

Pour tout modèle $S_m \in \mathcal{M}_n$, on définit $\widehat{f}_m = \arg\min_{t \in S_m} \gamma_n(t)$ et $f_m = \arg\min_{t \in S_m} \gamma(t)$. On dispose donc d'une collection d'estimateurs $\{\widehat{f}_m, m \in J_n\}$. Le but de la sélection de modèle est de construire un critère basé sur les données permettant de choisir un estimateur parmi la collection. Plusieurs critères de sélection de modèles existent (AIC, BIC...), basés sur des heuristiques différentes, mais dans ce manuscrit nous ne considérons que la méthode développée par Birgé and Massart (1998), particulièrement adaptée aux estimateurs par projection. Le meilleur estimateur $\widehat{f}_m$ pour le risque $\|.\|_0^2$ est celui pour lequel l'erreur $\mathbb{E}\left[\|\widehat{f}_m - f\|_0^2\right]$ est minimale, mais cette erreur n'est pas observable. On cherche donc à l'estimer, ou plus préciément à estimer la somme biais-variance

$$\mathbb{E}\left[\|\widehat{f}_m - f\|_0^2\right] \leq 2\|f - f_m\|_0^2 + 2\mathbb{E}\left[\|\widehat{f}_m - f_m\|_0^2\right], \quad \forall m \in J_n. \tag{1.6}$$

On construit une quantité qui estime la somme biais-variance, à une constante indépendante de m près, et le modèle $m \in J_n$ sélectionné est celui qui minimise cette quantité. Si le principe général de la sélection de modèle demeure le même pour l'étude des risques ponctuel et intégré, les techniques mises en oeuvre sont sensiblement différentes. Nous nous contenterons donc, dans cette introduction, de donner un aperçu global de la démarche, qui sera détaillée dans les différents chapitres.

$\star$ Le terme de biais est estimé par la quantité $\tilde{\gamma}_n(\widehat{f}_m)$ où $\tilde{\gamma}_n$ est une application de $\cup_{m \in J_n} S_m$ dans $\mathbb{R}$ déterminée par les observations et telle que $\mathbb{E}[\tilde{\gamma}_n(t)]$ est de l'ordre de $\|f - t\|_0^2 + c_0$ où c_0 est une constante indépendante de t. La construction de $\tilde{\gamma}_n$ constitue la principale différence entre la sélection de modèle pour les risques ponctuel et intégré.

$\star$ Sous certaines hypothèses concernant la collection de modèles $\mathcal{M}_n$, le terme de variance est majoré par une quantité non aléatoire :

$$\mathbb{E}\left[\|\widehat{f}_m - f_m\|_0^2\right] \leq C\frac{D_m}{n}$$

où C est une constante et D_m est la dimension de l'espace S_m. Par ailleurs, cette majoration est optimale en un sens défini en Section 1.2.2.

Finalement, on sélectionne le modèle $\widehat{m}$ qui vérifie

$$\widehat{m} = \arg\min_{m \in J_n} \left[\tilde{\gamma}_n(\widehat{f}_m) + pen(m) \right]$$

où la fonction $pen(m)$, appelée pénalité, est égale à $C' D_m / n$. L'estimateur de sélection de modèle est $\widehat{f}_{\widehat{m}}$. Dans le cadre du risque (i) (cf (1.1)), on parlera de sélection de modèle globale, et dans le cadre (ii) de sélection de modèle ponctuelle.

1.2.2 Inégalité oracle et risque minimax

Dans le cas le plus favorable, l'estimateur de sélection de modèle vérifie le résultat suivant, appelé inégalité oracle.

$$\mathbb{E}\left[\|\widehat{f}_{\widehat{m}} - f\|_0^2 \right] \leq A \inf_{m \in J_n} \left\{ \|f_m - f\|_0^2 + pen(m) \right\} + r_n \tag{1.7}$$

où le terme de reste r_n est négligeable devant le terme principal
$A \inf_{m \in J_n} \{\|f_m - f\|_0^2 + pen(m)\}$. Ceci signifie que le risque de l'estimateur de sélection de modèle $\widehat{f}_{\widehat{m}}$ est inférieur ou égal, à constante multiplicative près, à la somme biais-variance du meilleur estimateur de la collection. Dans ce manuscrit, nous obtenons de telles inégalités dans le cas du risque intégré.

Néanmoins, d'après (1.6) la somme biais-variance $\{\|f_m - f\|_0^2 + pen(m)\}$ n'est qu'un majorant du risque de l'estimateur $\widehat{f}_m$. Une minoration est donc nécessaire pour prouver que cette somme converge à la même vitesse que le risque $\mathbb{E}\left[\|\widehat{f}_m - f\|_0^2 \right]$ sur des classes de régularité classiques. Pour cela, on définit le risque de convergence minimax.

Definition 1.2.1 *Soit $(V_1, \ldots, V_n)$ un échantillon à partir duquel on veut estimer une fonction f et soit $\mathcal{F}$ une classe de fonctions qui contient f. Soit $d(.,.)$ une distance ou semi-distance. On définit le risque minimax sur $\mathcal{F}$, associé à la distance d comme le risque du meilleur estimateur construit à partir de $(V_1, \ldots, V_n)$, pour la fonction $f \in \mathcal{F}$ la moins favorable. Plus précisément,*

$$\mathcal{R}_n = \inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}[d(\widehat{f}_n, f)]$$

où l'infimum est considéré sur l'ensemble de tous les estimateurs de f. Le risque minimax dépend de la taille de l'échantillon $(V_1, \ldots, V_n)$. Soit $(\psi_n)_{n \in \mathbb{N}^*}$ une suite de réels strictement positifs. On dit que ψ_n est la vitesse optimale de convergence ou vitesse minimax sur $\mathcal{F}$ si il existe deux constantes positives c et C telles que

$$c \leq \lim \inf_{n \to +\infty} \psi_n^{-1} \mathcal{R}_n \leq \lim \sup_{n \to +\infty} \psi_n^{-1} \mathcal{R}_n \leq C.$$

On remarque que la vitesse de convergence minimax est définie à une constante multiplicative près.

Ainsi, soit $\mathcal{F}(\beta)$ une classe de régularité indexée par un paramètre de régularité β, et $\psi_{n,\beta}$ la vitesse minimax sur $\mathcal{F}(\beta)$. Soit $\mathfrak{B}$ un ensemble de paramètres (généralement un intervalle de $\mathbb{R}$) tel que pour tout $\beta \in \mathfrak{B}$, il existe un modèle $m \in J_n$ et une constante positive C tels que

$$\|t_m - t\|_0^2 + pen(m) \leq C\psi_{n,\beta}$$

pour tout $t \in \mathcal{F}_\beta$. Alors, d'après l'inégalité oracle (1.7), si $f \in \mathcal{F}(\beta)$ avec $\beta \in \mathfrak{B}$, l'estimateur de sélection de modèle $\widehat{f}_{\widehat{m}}$ converge à la vitesse minimax sur $\mathcal{F}(\beta)$. L'estimateur s'adapte donc à la régularité β de f sans que celle-ci soit connue : on dit que $\widehat{f}_{\widehat{m}}$ est adaptatif.

1.2.3 Risque minimax adaptatif

Dans certains cas, notamment en sélection de modèle ponctuelle, le résulat obtenu n'est pas tout à fait une inégalité oracle : l'estimateur $\widehat{f}_{\widehat{m}}$ ne converge pas exactement à la vitesse minimax sur les espaces de régularité classiques.

Considérons l'exemple de l'estimation ponctuelle de densité sur un intervalle I, developpé au Chapitre 3. On considère le risque ponctuel $\mathbb{E}[(\widehat{f} - f)^2(x_0)]$ où x_0 est un point fixé. Soit $\mathcal{H}(\beta, L)$ les classes de régularité suivantes : pour tout $(\beta, L) \in \mathbb{R}_+^*$

$$\mathcal{H}(\beta, L) = \{f : I \to \mathbb{R} \text{ r fois dérivable, } |f^{(r)}(x) - f^{(r)}(y)| \leq L|x - y|^{\beta - r}, \ \forall x, y \in I\}.$$

où r est le plus grand entier strictement inférieur à β (ces classes seront étudiées plus précisément dans la Section 1.2.6). On prouve que la vitesse minimax sur $\mathcal{H}(\beta, L)$ est $n^{-2\beta/(2\beta+1)}$ mais l'estimateur de sélection de modèle ponctuelle vérifie

$$\mathbb{E}[(\widehat{f}_{\widehat{m}} - f)^2(x_0)] \leq C \left(\frac{n}{\ln n}\right)^{-2\beta/(2\beta+1)}$$

si $f \in \mathcal{H}(\beta, L)$. $\widehat{f}_{\widehat{m}}$ ne converge pas à la vitesse minimax, son risque comporte une perte logarithmique. Néanmoins, on prouve que la perte logarithmique est inévitable pour un estimateur adaptatif dans ce contexte et que cette vitesse est optimale. Pour cela, Lepski (1991) définit la notion de vitesse de convergence minimax adaptative.

Soit $\{\mathcal{F}(\beta), \beta \in \mathfrak{B}\}$ un ensemble de classes de régularité et d une semi-distance. Pour tout $\beta \in \mathfrak{B}$, soit

$$\mathcal{R}_{n,\beta} = \inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}(\beta)} \mathbb{E}[d(\widehat{f}_n, f)].$$

Soit $\{\psi_{n,\beta}, n \in \mathbb{N}^*, \beta \in \mathfrak{B}\}$ une suite de nombres strictement positifs. $\psi_{n,\beta}$ est le taux de convergence minimax adaptatif sur les classes $\{\mathcal{F}(\beta), \beta \in \mathfrak{B}\}$ si il existe deux constantes strictement positives c et C telles que

$$c \leq \lim \inf_{n \to +\infty} \sup_{\beta \in \mathfrak{B}} \psi_{n,\beta}^{-1} \mathcal{R}_{n,\beta} \leq \lim \sup_{n \to +\infty} \sup_{\beta \in \mathfrak{B}} \psi_{n,\beta}^{-1} \mathcal{R}_{n,\beta} \leq C.$$

Ainsi, dans le cadre de l'estimation ponctuelle de densité, Butucea (2001) démontre que la vitesse de convergence minimax adaptative est $(n/\ln n)^{-2\beta/(2\beta+1)}$.

1.2.4 Inégalités de déviation

Les méthodes développées pour prouver l'adaptativité en sélection de modèle ponctuelle et globale sont très différentes. Néanmoins, elles sont essentiellement basées sur des inégalités de déviation de processus empiriques. Plus précisément, les preuves s'appuient sur la majoration de termes de la forme suivante. Pour la sélection de modèle globale

$$\sum_{m \in J_n} \int_0^\infty \mathbb{P}\left[\sup_{t \in S_m, \|t\|_0 = 1} (\tilde{\gamma}_n(t) - \|t\|_0^2) \geq pen(m) + x \right] dx \tag{1.8}$$

et pour la sélection de modèle ponctuelle

$$\sum_{m \in J_n} \int_0^\infty \mathbb{P}\left[\tilde{\gamma}_n(\widehat{f}_m) - \|f - f_m\|_0^2 \geq pen(m) + x \right] dx. \tag{1.9}$$

En effet, f minimise $\gamma(t) + c_0$ et $\widehat{f}_{\widehat{m}}$ minimise $\tilde{\gamma}_n(t) + pen(m)$ où $\tilde{\gamma}_n(t)$ est un processus empirique qui estime $\gamma(t) + c_0$, donc la distance entre $\widehat{f}_{\widehat{m}}$ et f est majorée par la déviation entre $\tilde{\gamma}_n$ et γ.

a) Inégalités de Talagrand

Les inégalités de Talagrand majorent des déviations de suprema de processus empiriques par rapport à leur espérance, c'est-à-dire des variables aléatoires de la forme : $\sup_{t \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (t(X_i) - \mathbb{E}[t(X_i)])$, où les (X_i) sont des variables ou des vecteurs aléatoires indépendants et $\mathcal{F}$ est un ensemble de fonctions. Ces résultats peuvent être exprimés sous forme de probabilité de déviation ou d'espérance, pour des variables indépendantes ou i.i.d.

Le résultat suivant, issu de Klein and Rio (2005) s'appuie sur les travaux de Talagrand et Ledoux.

Theorem 1.2.1 *Soit $(X_1, .., X_n)$ des variables aléatoires indépendantes. Soit $\mathcal{F}$ un ensemble dénombrable de fonctions de $\mathbb{R}$ dans $[-1, 1]^n$. Soit*

$$Z := \sup_{s \in \mathcal{F}} \sum_{i=1}^n s^{(i)}(X_i).$$

On suppose que $\mathbb{E}[s^{(i)}(X_i)] = 0$ *pour tout* i. *Alors pour tout* $x > 0$,

$$P[Z \geq \mathbb{E}Z + x] \leq \exp\left(-\frac{x^2}{2(V + 2\mathbb{E}Z) + 3x}\right)$$

où $V = \sup_{s\in\mathcal{F}} Var(\sum_{i=1}^n s^{(i)}(X_i))$.

Par des arguments de densité, ce résultat peut s'étendre à un ensemble de fonctions $\mathcal{F}$ non dénombrable qui possède une partie dénombrable dense pour la norme infinie. C'est le cas en particulier, si $\mathcal{F}$ est une boule pour la norme L^2 d'un sous-espace vectoriel de dimension finie de $L^2 \cap L^\infty$.

Cette inégalité est majoritairement utilisée pour des fonctions s dont toutes les composantes $(s^{(i)})_{i=1,\ldots,n}$ sont identiques et bornées en norme infinie par une constante b. Dans ce cas, en considérant les fonctions s/b à valeurs dans $[-1,1]^n$, le Théorème 1.2.1 peut être réécrit ainsi.

Theorem 1.2.2 *Soit* $\mathcal{F}$ *un ensemble de fonctions de* $\mathbb{R}$ *dans* $\mathbb{R}$, *uniformément bornées en norme infinie, qui possède une partie dénombrable dense pour la norme infinie. Soit* $(X_1, .., X_n)$ *des variables aléatoires indépendantes. Soit*

$$Z = \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n (f(X_i) - \mathbb{E}[f(X_i)]) \;\; ou \; Z = \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n (f(X_i) - \mathbb{E}[f(X_i)])\right|$$

et

$$b \geq \sup_{f\in\mathcal{F}} \|f\|_\infty, \quad v \geq \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n Var(f(X_i)), \quad \mathbb{H} \geq \mathbb{E}(Z).$$

Pour tout $\lambda > 0$,

$$P[Z \geq \mathbb{H} + \lambda] \leq \exp\left(-\frac{n\lambda^2}{2(v + 4b\mathbb{H}) + 6b\lambda}\right). \tag{1.10}$$

Par ailleurs,

$$\mathbb{E}\left[(Z^2 - (1+2\theta)\mathbb{H}^2)_+\right] \leq \int_0^{+\infty} P\left[Z^2 - (1+2\theta)\mathbb{H}^2 \geq x\right] dx$$

En effet, pour toute variable Z de densité f_Z et de fonction de survie $\overline{F}_Z(x) = P[Z \geq x]$.

$$\mathbb{E}[Z_+] = \int_0^\infty x f_Z(x)dx = -\int_0^\infty x\overline{F}'_Z(x)dx = -[x\overline{F}_Z(x)]_0^{+\infty} + \int_0^\infty \overline{F}_Z(x)dx \leq \int_0^\infty \overline{F}_Z(x)dx$$

Le Théorème 1.2.2 entraine alors le résultat suivant.

Theorem 1.2.3 *Soit $\mathcal{F}$ un ensemble de fonctions uniformément bornées en norme infinie, qui possède une partie dénombrable dense pour la norme infinie. Soit $(X_1, \ldots, X_n)$ des variables aléatoires indépendantes. Soit*

$$Z = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - \mathbb{E}[f(X_i)]) \right|.$$

Pour tout

$$b \geq \sup_{f \in \mathcal{F}} \|f\|_\infty, \quad v \geq \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} Var(f(X_i)), \quad \mathbb{H} \geq \mathbb{E}(Z)$$

et pour tout $\theta > 0$, il existe des constantes numériques $\overline{C}$, $\overline{C}'$, $\overline{\kappa}$, $\overline{\kappa}'$ (qui ne dépendent que de θ) telles que

$$\mathbb{E}\left[(Z^2 - (1+2\theta)\mathbb{H}_+^2\right] \leq \overline{C} \frac{v}{n} \exp\left(-\overline{\kappa} \frac{n\mathbb{H}^2}{v}\right) + \overline{C}' \frac{b^2}{n^2} \exp\left(-\overline{\kappa}' \frac{n\mathbb{H}}{b}\right).$$

Une preuve détaillée de ces deux Théorèmes est fournie au Chapitre 6, dans le cas plus général de fonctions $f^{(i)}$ non identiques.

b) Inégalité de Bernstein

L'inégalité de Bernstein permet de majorer la probabilité de déviation de processus empiriques.

Theorem 1.2.4 *Soit $(X_1, ..X_n)$ des variables aléatoires indépendantes. Soit*

$$S = \frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E}[X_i].$$

Supposons que

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i^2] \leq v \quad et \quad \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[(X_i)_+^k] \leq \frac{k!}{2} \times v \times c^{k-2} \quad \forall k \geq 2. \tag{1.11}$$

1) Pour tout $x > 0$,

$$P[S \geq \sqrt{2vx} + cx] \leq \exp(-nx),$$
$$P[|S| \geq \sqrt{2vx} + cx] \leq 2\exp(-nx).$$

2) De manière équivalente, pour tout $\epsilon > 0$,

$$P[S \geq \epsilon] \leq \exp\left(-\frac{n\epsilon^2}{2(v^2 + c\epsilon)}\right)$$
$$P[|S| \geq \epsilon] \leq 2\exp\left(-\frac{n\epsilon^2}{2(v^2 + c\epsilon)}\right).$$

Remark 2 *En particulier, la condition* (1.11) *est satisfaite si les* $\{X_i\}$ *sont i.i.d. et*

$$\mathbb{E}[X_i^2] \leq v \quad \textit{and} \quad \|X_i\|_\infty \leq c.$$

L'inégalité de Bernstein est énoncée en particulier dans Birgé and Massart (1998). Cette inégalité est moins forte que l'inégalité de Talagrand au sens où elle majore la probabilité de déviation de processus empirique, et non de supremum de processus.

Ainsi, le choix de la pénalité $pen(m)$ est guidé par deux exigences.

$\star$ D'après les résultats des Sections 1.2.2 et 1.2.3, la pénalité doit être choisie de l'ordre de la variance pour que $\|f_m - f\|_0^2 + pen(m)$ soit de l'ordre de $\mathbb{E}\left[\|\widehat{f}_m - f\|_0^2\right]$ (dans le cas d'une inégalité oracle), ou bien de telle sorte que $\inf_{m \in J_n}\{\|f_m - f\|_0^2 + pen(m)\}$ converge à la vitesse minimax adaptative. Ceci impose une limite supérieure à la pénalité.

$\star$ De manière évidente, les probabilités de déviation (1.8) et (1.9) sont des fonctions décroissantes de $pen(m)$. La pénalité doit donc être suffisamment grande pour que ces probabilités soient négligeables devant le risque de l'estimateur $\widehat{f}_{\widehat{m}}$.

1.2.5 Bases de modèles

La construction d'un estimateur par sélection de modèles nécessite le choix préalable d'une collection de modèles, et les théorèmes énoncés dans les chapitres suivants requièrent certaines propriétés concernant cette collection, regroupées dans cette section. Nous présenterons ensuite des exemples classiques de collections de modèles qui vérifient ces propriétés.

a) Propriétés des collections de modèles

Soit I un intervalle de $\mathbb{R}$. Considérons une collection de modèles

$$\mathcal{M}_n = \{S_m, m \in J_n\}$$

où S_m est un sous-espace vectoriel de $\mathbb{L}^2(I) \cap \mathbb{L}^\infty(I)$ pour tout $m \in J_n$. J_n dépend généralement de la taille n de l'échantillon. Pour tout $m \in J_n$, soit $\{\phi_k, k \in I_m\}$ une base de S_m orthonormée

pour la norme L^2. La plupart des collections classiques sont constituées de modèles de dimension finie $D_m = Dim(S_m)$, mais nous présenterons également un exemple de modèle de dimension infinie. Par ailleurs, pour les modèles de dimension finie, on se restreint à des dimensions $D_m \leq N_n$ avec $N_n \leq n$. En effet, dans ce cas le terme de variance est d'ordre D_m/n, donc le risque d'un estimateur $\widehat{f}_m$ ne peut tendre vers 0 quand n tend vers l'infini que si $D_m < n$.

• **Complexité de la collection**

Supposons que les modèles S_m sont de dimension finie, et notons $D_m = Dim(S_m)$. La collection $\mathcal{M}_n$ peut comporter plusieurs modèles de même dimension et la complexité désigne le nombre de modèles de dimension $D_m = D$, pour une valeur de $D \in \mathbb{N}^*$ donnée. On impose généralement des restrictions sous la forme suivante. Pour tout $A > 0$, il existe une constante A' telle que

$$\sum_{m \in J_n} \exp(-AD_m) \leq A' \quad \Leftrightarrow \quad \sum_{D \leq N_n} Card\left(\{m \in J_n, D_m = D\}\right) \exp(-AD) \leq A' \qquad (1.12)$$

où $N_n = \max_{m \in J_n} D_m$. Des variantes de cette condition consistent à remplacer $\exp(-AD_m)$ par $\exp(-A\sqrt{D_m})$, ou $D_m \exp(-AD_m)$.

• **Connexion de normes**

La connexion de normes est une relation entre les normes $\|t\|^2 = \int_I t^2(x)dx$ et $\|t\|_\infty = \sup_{x \in I} |t(x)|$ sur les espaces $\{S_m\}$. Plus précisément, on suppose qu'il existe une constante K indépendante de n telle que, pour tout $m \in J_n$,

$$\|t\|_\infty \leq K\sqrt{D_m}\|t\|, \quad \forall t \in S_m.$$

La proposition suivante fournit une caractérisation de la connexion de normes à l'aide des fonctions de base des espaces S_m.

Proposition 1.2.1 *Soit V un sev de $\mathbb{L}^2(I) \cap \mathbb{L}^\infty(I)$ de dimension d, ν une densité définie sur V et $(\psi_1, .., \psi_d)$ une base ν-orthonormée de V, alors il y a équivalence entre ces deux affirmations.*

(i) Pour tout $t \in V$,

$$\|t\|_\infty \leq K\sqrt{d}\|t\|_\nu.$$

(ii)

$$\|\sum_{i=1}^{d} \psi_i^2\|_\infty \leq K^2 d$$

où K est une constante indépendante de d.

Preuve de la Proposition 1.2.1

$\star$ Supposons (i) vérifiée. Soit $x \in I$. Notons $|\ .\ |$ la norme euclidienne de $\mathbb{R}^d$, $\langle\ .\ \rangle$ le produit scalaire euclidien de $\mathbb{R}^d$ et $\psi(x) = (\psi_1(x), .., \psi_d(x))$.

$$\begin{aligned}
\sum_{i=1}^{d} \psi_i(x)^2 &= |\psi(x)|^2 \\
&= \sup_{a \in \mathbb{R}^d, |a|=1} < a, \psi(x) >^2 \\
&= \sup_{a \in \mathbb{R}^d, |a|=1} (\sum_{i=1}^{d} a_i \psi_i(x))^2 \\
&= \sup_{\|t\|_\nu = 1} (t(x))^2 \\
&\leq K^2 d
\end{aligned}$$

ce qui prouve (ii).

$\star$ Supposons (ii) vérifiée. Soit $x \in I$ et $t = \sum_{i=1}^{d} a_i \psi_i \in V$. D'après l'inégalité de Cauchy Schwartz,

$$(t(x))^2 \leq \sum_{i=1}^{d} a_i^2 \times \sum_{i=1}^{d} (\psi_i(x))^2 = \|t\|_\nu^2 \times K^2 d$$

ce qui prouve (i). □

Remark 3 *En particulier, les bases bornées en norme infinie par une constante indépendante de n vérifient la propriété de connexion de normes car elles satisfont la condition* (ii).

Remark 4 *La Proposition* 1.2.1 *est également vérifiée si* $I \subset \mathbb{R}^2$.

- **Localisation**

La collection $\mathcal{M}_n$ est dite localisée si pour tout $m \in J_n$, l'espace S_m possède une base orthonormée $\{\phi_k^m, k = 1, \ldots, D_m\}$ telle que les supports des $\{\phi_k^m\}$ ne sont pas "trop superposés". Plus précisément, on suppose qu'il existe une constante K indépendante de m et n telle que pour tout $k \in \{1, \ldots, D_m\}$, et pour tout $m \in J_n$,

$$\begin{cases} (i) & Card(\{k' \in \{1, \ldots, D_m\}, \phi_k^m \phi_{k'}^m \neq 0\}) \leq K \\ (ii) & \|\phi_k^m\|_\infty^2 \leq K D_m. \end{cases}$$

Contrairement à la connexion de normes, la propriété de localisation dépend de la base de S_m considérée.

La propriété de localisation est plus forte que la propriété de connexion de normes. En effet, supposons que l'espace S_m possède une base localisée $\{\phi_k^m, k = 1, \ldots, D_m\}$. Soit $x \in I$, supposons qu'il existe $k_0 \in \{1, \ldots, D_m\}$ tel que $\phi_{k_0}^m(x) \neq 0$, alors

$$\begin{aligned}\sum_{k=1}^{D_m} \left(\phi_k^m\right)^2(x) &= \sum_{k=1,\ldots,D_m, \phi_k^m \phi_{k_0}^m \neq 0} \left(\phi_k^m\right)^2(x) \\ &\leq Card\left(\{k \in I_m, \phi_k^m \phi_{k_0}^m \neq 0\}\right) \sup_{k=1,\ldots,D_m} \|\phi_k^m\|_\infty^2 \\ &\leq K^2 D_m.\end{aligned}$$

Si $\phi_k(x) = 0$ pour tout $k \in \{1, \ldots, D_m\}$, cette même inégalité est évidemment vérifiée, ce qui prouve la propriété de connexion de normes.

Dans ce manuscrit, nous aurons parfois besoin d'une condition supplémentaire : pour tout $m \in J_n$, il existe une partition $\{I_1, \ldots, I_s\}$ de $\{1, \ldots, D_m\}$ telle que

$$Card(I_i) \leq K, \quad \forall i \in \{1, \ldots, s\} \tag{1.13}$$

et pour tout $i, j \in \{1, \ldots, s\}$, $i \neq j$ et pour tout $h \in I_i$, $k \in I_j$,

$$\phi_k^m \phi_h^m = 0 \tag{1.14}$$

$$\text{et} \quad \|\phi_k^m\|_\infty^2 \leq K^2 D_m. \tag{1.15}$$

De façon immédiate, cette propriété de localisation forte entraine la localisation simple. En effet, supposons (1.13), (1.14) et (1.15) vérifiées. Soit $k \in \{1, \ldots, D_m\}$, et soit $i \in \{1, \ldots, s\}$ tel que $k \in I_i$, alors

$$\{k' \in \{1, \ldots, D_m\}, \phi_k^m \phi_{k'}^m \neq 0\} \subset I_i$$

d'où

$$Card\left(\{k' \in \{1, \ldots, D_m\}, \phi_k^m \phi_{k'}^m \neq 0\}\right) \leq Card(I_i) \leq K.$$

Mais il n'y a pas équivalence entre ces deux notions, comme l'illustre l'exemple suivant. Soit $\phi(x) = (1/\sqrt{2})\left(1\!\!1_{\{[0,1/2[\cup[1,3/2[\}}(x) - 1\!\!1_{\{[1/2,1[\cup[3/2,2[\}}(x)\right)$. Pour tout $m \in \mathbb{N}^*$ et pour tout $k \in \{-2, \ldots, 2m\}$ soit

$$\phi_k^m(x) = \sqrt{m}\phi(mx - k).$$

Les fonctions ϕ_k^m, obtenues par translation et homothétie de la fonction ϕ, sont à support dans $[k/m, (k+2)/m]$. On considère la collection $\mathcal{M}_n = \{S_m, m = 1, \ldots, N\}$ avec

$$S_m = Vect\{\phi_k^m, k = -2, \ldots, 2m\}.$$

On constate que la famille $\{\phi_k^m, k = -2, \ldots, 2m\}$ constitue une base orthonormée de S_m, pour tout m.

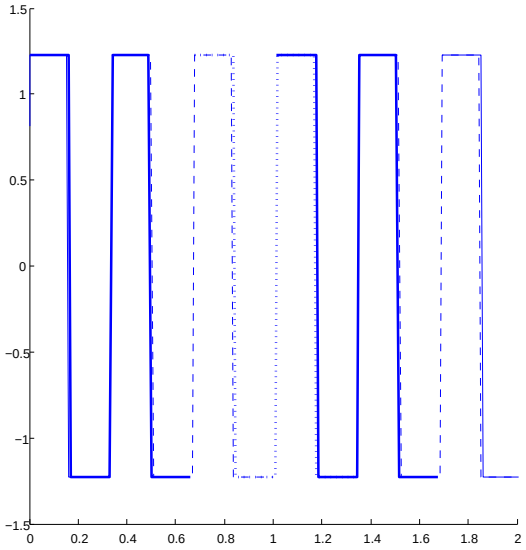


Figure 1.4: Base localisée mais ne vérifiant pas la propriété de localisation forte.

La Figure 1.4 représente les courbes des fonctions $\{\phi_k^3, k = -2, \ldots, 6\}$, restreintes à l'intervalle $[0, 2]$. On constate de manière évidente que la base $\{\phi_k^3, k =, -2 \ldots, 6\}$ est localisée, avec la constante de localisation $K = 2$. Néanmoins il n'y a pas de partition stricte de $\{-2, \ldots, 6\}$ telle que (1.13) et (1.14) soient vérifiées. Cette constatation se généralise à S_m, pour tout $m \in \mathbb{N}^*$. Ainsi, la collection $\mathcal{M}_n$ est localisée mais ne vérifie pas la propriété de localisation forte.

De plus, soit $\nu : I \to \mathbb{R}^+$ une application telle que $\nu(x) \geq m_0 > 0$ pour tout $x \in I$, et $\|t\|_\nu$ la norme L^2 associée. Si l'espace S_m possède une base $\|.\|$-orthonormée $\{\phi_k^m, k = 1, \ldots, D_m\}$ qui vérifie la propriété de localisation forte, alors il existe une base $\{\psi_k^m, k = 1, \ldots, D_m\}$ de S_m $\|.\|_\nu$-orthonormée qui vérifie également la propriété de localisation forte. En effet, pour tout I_i, on effectue une orthogonalisation de Gram-Schmidt de la famille $\{\phi_k^m, k \in I_i\}$ pour la norme $\|.\|_\nu$. La famille $\{\psi_k^m, k \in I_i\}$ vérifie immédiatement les propriétés (1.13) et (1.14). De plus, elle vérifie la connexion de normes donc

$$\|\psi_k^m\|_\infty \leq K\sqrt{D_m}\|\psi_k^m\| \leq \frac{K}{m_0}\sqrt{D_m}\|\psi_k^m\|_\nu = \frac{K}{m_0}\sqrt{D_m}.$$

Ainsi, la base $\{\psi_k^m, k = 1, \ldots, D_m\}$ est bien localisée.

- **Espace englobant**

La collection $\mathcal{M}_n$ possède un espace englobant si il existe un modèle $m_n \in J_n$ tel que

$$S_m \subset S_{m_n}, \quad \forall m \in J_n.$$

Cette propriété garantit que pour tous modèles S_m et $S_{m'}$ de $\mathcal{M}_n$, la somme $S_m + S_{m'}$ est incluse dans le modèle S_{m_n}. Par ailleurs, dans certains cas, les propriétés telles que la connexion de normes ne sont nécessaires que sur l'espace englobant, ce qui allège les hypothèses.

b) Collections de modèles classiques

Comme l'illustre l'exemple de la section précédente, les collections de modèles $\mathcal{M}_n$ sont généralement construites à partir de fonctions de base de même nature : histogrammes, fonctions trigonométriques... Des travaux récents (le Pennec and Rivoirard (To appear)) s'intéressent à des collections pouvant mêler des modèles de nature différente, mais les problèmes soulevés sont beaucoup plus complexes et nous ne les aborderons pas dans ce manuscrit.

• Histogrammes

Les bases d'histogrammes constituent l'exemple le plus simple de bases localisées, définies sur un compact que l'on supposera ici égal à $[0, 1]$. Une base d'histogrammes est constituée d'une famille d'indicatrices associée à une partition de $[0, 1]$. Soit $\mathcal{P}$ une partition de $[0, 1]$, notons $S_{\mathcal{P}}$ le modèle suivant

$$S_{\mathcal{P}} = Vect\left\{1_I, I \in \mathcal{P}\right\}.$$

Une idée naturelle est de considérer la partition régulière suivante, de pas $1/N_n$.

$$\mathcal{P}_n = \left\{ [\frac{k-1}{N_n}, \frac{k}{N_n}[, k = 1, \ldots, N_n \right\}$$

et la collection $\mathcal{M}_n = \{S_{\mathcal{P}}, \mathcal{P}$ partition plus grossière que $\mathcal{P}_n\}$. Mais la complexité de cette collection est trop importante. En effet, pour tout $D \in \mathbb{N}^*$, le nombre de modèles de dimension D et égal à $\binom{N_n-1}{D-1}$ et

$$\sum_{D=1}^{N_n} \binom{N_n-1}{D-1} \exp(-AD) = \frac{1}{e} \sum_{D=0}^{N_n-1} \binom{N_n-1}{D} (e^{-A})^D = \frac{1}{e} \left(1 + e^{-A}\right)^{N_n-1}$$

et cette quantité n'est pas bornée quand n tend vers l'infini. Par ailleurs, la majoration de termes de déviation de la forme (1.8) ou (1.9) induit une limitation du cardinal de la collection, qui doit généralement être polynomial en n. Or $Card(\mathcal{M}_n) = 2^{N_n}$ ce qui contraint N_n, et donc tous les D_m, à être plus petits qu'un terme d'ordre $\ln n$. Ceci exclut l'adaptativité sur les espaces de régularité classiques, comme nous le verrons dans les chapitres suivants.

Néanmoins, Akakpo and Durot (2010) proposent des estimateurs de sélection de modèles sur des bases d'histogrammes où la propriété de complexité (1.12) n'est pas vérifiée, ce qui permet de considérer des histogrammes très irréguliers.

On peut réduire le nombre de modèles dans la collection en se limitant à des histogrammes de pas régulier, c'est à dire associés à une partition de $[0,1]$ en intervalles de même longueur. Soit $\mathcal{M}_n = \{S_{\mathcal{P}_D}, D = 1, \ldots, N_n\}$ où

$$\mathcal{P}_D = \left\{ \left[\frac{k-1}{D}, \frac{k}{D} \right[, k = 1, \ldots, D \right\}.$$

Cette collection est beaucoup moins complexe car elle ne possède qu'un modèle de dimension D pour tout $D = 1, \ldots, N_n$. Néanmoins, elle ne possède pas de modèle englobant.

Enfin, les collections d'histogrammes le plus souvent considérées sont constituées d'histogrammes réguliers de pas 2^d, $d \in \mathbb{N}$, appelés histogrammes diadiques. Soit $N_n \leq n$ et

$$\mathcal{M}_n = \{S_{\mathcal{P}_{2^d}}, d \in \mathbb{N}, 2^d \leq N_n\}.$$

Comme la collection d'histogrammes réguliers précédente, cette collection est peu complexe (au plus un modèle par dimension), et les modèles sont emboités:

$$S_{\mathcal{P}_1} \subset S_{\mathcal{P}_2} \subset S_{\mathcal{P}_4} \subset \cdots \subset S_{\mathcal{P}_{2^{d_n}}}$$

où d_n est le plus grand entier d tel que $2^d \leq N_n$. En particulier, $S_{\mathcal{P}_{2^{d_n}}}$ est un modèle englobant.

Ces trois collections vérifient la propriété de localisation forte.

• **Polynômes par morceaux**

Les polynômes par morceaux (notés PPM) constituent une généralisation des histogrammes : soit $\mathcal{P} = \{I_1, \ldots, I_s\}$ une partition de $[0,1]$, et $\mathcal{R} = \{r_1, \ldots, r_s\}$ un ensemble d'entiers naturels, notons $S_{\mathcal{P},\mathcal{R}}$ l'ensemble des fonctions t telles que, pour tout $i = 1, \ldots, s$, la restriction de t à I_i est un polynôme de degré inférieur ou égal à r_i. Comme dans le cas des histogrammes, on peut considérer des PPM construits sur des partitions générales, régulières ou diadiques mais nous ne présentons ici que les collections de PPM diadiques.

Soit N_n et r_{max} deux entiers naturels tels que $(r_{max}+1)N_n \leq n$ et

$$\mathcal{M}_n = \left\{ S_{\mathcal{P}_D,\mathcal{R}}, D = 2^d \leq N_n \text{ et } d \in \mathbb{N}, \mathcal{R} = \{r_1, \ldots, r_D\} \text{ et } 0 \leq r_i \leq r_{max}, \forall i = 1, \ldots, D \right\}.$$

N_n dépend de n mais r_{max} est fixé. Cette collection possède un espace englobant: $S_{\mathcal{P}_{2^{d_{max}}},\mathcal{R}_{max}}$ où d_{max} est le plus grand entier d tel que $2^d \leq N_n$ et $\mathcal{R}_{max} = \{r_{max}, \ldots, r_{max}\}$, et vérifie la

propriété de localisation forte, mais sa complexité est trop élevée. En effet, pour toute partition $\mathcal{P}_D$ de $[0,1]$, le nombre de modèles de $\mathcal{M}_n$ basés sur cette partition est $(r_{max}+1)^D$. De plus, pour tout $\mathcal{R} = \{r_1, \ldots, r_D\}$,

$$Dim(S_{\mathcal{P}_D,\mathcal{R}}) = (r_1+1) + (r_2+1) + \cdots + (r_D+1) \leq D(r_{max}+1).$$

Ainsi,

$$\begin{aligned} \sum_{m\in\mathcal{M}_n} \exp(-AD_m) &\geq \sum_{d\in\mathbb{N}, D=2^d\leq N_n} (r_{max}+1)^D \exp(-AD(r_{max}+1)) \\ &= \sum_{d\in\mathbb{N}, D=2^d\leq N_n} [(r_{max}+1)\exp(-A(r_{max}+1)]^D \end{aligned}$$

et cette somme est divergente si $A < \ln(r_{max}+1)/(r_{max}+1)$.

On considère donc une collection moins complexe, mais plus restrictive, où les degrés sont les mêmes sur tous les intervalles, pour un modèle donné.

$$\mathcal{M}_n = \left\{ S_{\mathcal{P}_D,\mathcal{R}}, D = 2^d \leq N_n \text{ et } d \in \mathbb{N}, \mathcal{R} = \{r, \ldots, r\}, r \leq r_{max} \right\}.$$

Alors, la quantité

$$\begin{aligned} \sum_{m\in\mathcal{M}_n} \exp(-AD_m) &= \sum_{d\in\mathbb{N}, D=2^d\leq N_n} \sum_{r=0}^{r_{max}} \exp(-AD(r+1)) \\ &\leq (r_{max}+1) \sum_{d\in\mathbb{N}, D=2^d\leq N_n} \exp(-AD) \end{aligned}$$

est majorée par une constante indépendante de N_n. On peut également se restreindre à un degré égal à r_{max} pour tous les modèles:

$$\mathcal{M}_n = \left\{ S_{\mathcal{P}_D,\mathcal{R}_{max}}, D = 2^d \leq N_n \text{ et } d \in \mathbb{N} \right\}.$$

• **Fonctions trigonométriques**

Pour tout $m \in \mathbb{N}^*$, soit S_m le modèle suivant sur $[0,1]$:

$$S_m = Vect\left\{ \mathbb{1}_{[0,1]}, x \to \cos(2\pi kx), x \to \sin(2\pi kx), k = 1, \ldots, m \right\}$$

de dimension $2m+1$ et soit $\mathcal{M}_n = \{S_m, 2m+1 \leq N_n\}$.

De manière évidente, les modèles de la collection $\mathcal{M}_n$ sont emboités. La collection ne comporte donc qu'un modèle par dimension et possède un modèle englobant. On constate également

que les fonctions de base ne sont pas localisées, car leur support est égal à $[0,1]$, mais elles sont uniformément bornées par 1 donc la collection vérifie la propriété de connexion de normes.

• **Ondelettes**

La décomposition en base d'ondelettes est une technique originellement developpée en traitement du signal, mais qui constitue actuellement un outil important des statistiques non paramétriques. Les ondelettes permettent une décomposition de la fonction à plusieurs échelles : on parle d'analyse multi-résolution. On considère une fonction ψ appelée ondelette mère et une fonction φ appelée ondelette père à support dans un intervalle I, et un entier $r \in \mathbb{N}^*$ appelé régularité qui vérifient un certain nombre de propriétés.

1) $\psi, \ldots, \psi^{(r)} \in L^\infty(I)$.

2) Les dérivées successives de ψ sont à décroissance rapide i.e. pour tout $0 \leq k \leq r$, pour tout $n \geq 1$, il existe une constante C_n telle que

$$|\psi^{(k)}(x)| \leq C_n(1+|x|)^{-n} \ , \ \forall x \in I.$$

3) Pour tout $0 \leq k \leq r$, $\int_I x^k \psi(x)dx = 0$.

4) La famille $\{x \to 2^{j/2}\psi(2^{j/2}x-k), (j,k) \in \mathbb{Z}^2\}$ constitue une base orthonormée de $L^2(\mathbb{R})$.

On suppose que φ vérifie les propriétés 1) et 2) ci-dessus, ainsi que les conditions suivantes.

3') $\int_I \varphi(x)dx = 1$.

4') La famille $\{x \to \varphi(x-k), k \in \mathbb{Z}\} \cup \{\psi_{j,k}, j \in \mathbb{N}, k \in \mathbb{Z}\}$ constitue une base de $L^2(\mathbb{R})$.

Alors toute fonction $t \in L^2(\mathbb{R})$ admet une décomposition de la forme suivante :

$$t = \sum_{k\in\mathbb{Z}} \langle \varphi_k, t\rangle \varphi_k + \sum_{j\geq 0}\sum_{k\in\mathbb{Z}} \langle \psi_{j,k}, t\rangle \psi_{j,k}$$

où $\langle ., . \rangle$ désigne le produit scalaire de L^2 et

$$\varphi_k(x) = \varphi(x-k) \quad \text{et} \quad \psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k).$$

Le terme $\sum_{k\in\mathbb{Z}} \langle \varphi_k, t\rangle \varphi_k$ fournit une approximation grossière de t, et les termes $\sum_{k\in\mathbb{Z}} \langle \psi_{j,k}, t\rangle \psi_{j,k}$ pour $j \geq 0$ représentent les variations de t à une échelle de plus en plus petite quand j augmente.

Les modèles considérés sont, pour tout $m \in \mathbb{N}*$,

$$S_m = Vect\left(\{\varphi_k, k \in \mathbb{Z}\} \cup \{\psi_{j,k}, j = 1, \ldots, m-1, k \in \mathbb{Z}\}\right).$$

Ces modèles sont de dimension infinie. Pour se ramener à des modèles de dimension finie, on considère des ondelettes ψ et φ à support compact, noté $[-A, A]$, et on estime des fonctions à support dans $[0, 1]$. Ainsi, pour tout $j \geq 0$, et pour tout $k \notin [-2^j - A, 2^j + A]$, la restriction de $\psi_{j,k}$ à $[0, 1]$ est identiquement nulle d'où

$$S_m = Vect\left(\{\varphi_k, k \in \Gamma(0)\} \cup \{\psi_{j,k}, j = 0, \ldots, m-1, k \in \Gamma(j)\}\right)$$

où $\Gamma(j) = \{k \in \mathbb{Z}, k \in [-2^j - A, 2^j + A]\}$. Finalement, on considère la collection de modèles suivante:

$$\mathcal{M}_n = \{S_m, m \in \mathbb{N}^* | 2^m \leq N_n\}.$$

Ces modèles sont emboités et les fonctions de bases qui les engendrent vérifient la propriété de localisation, mais pas de localisation forte (cf exemple développé dans la Section "localisation").

• **Sinus cardinal**

Contrairement aux collections présentées précédemment, la collection suivante, construite par translations et homothéties de la fonction sinus-cardinal, permet d'estimer des fonctions à support dans $\mathbb{R}$. Soit

$$\phi(x) = \sin(\pi x)/(\pi x), \quad \forall x \in \mathbb{R}^*$$

et $\phi(0) = 1$. Pour tout $m > 0$, pour tout $k \in \mathbb{Z}$, soit

$$\phi_{m,k}(x) = \sqrt{m}\phi(mx - k), \quad \forall x \in \mathbb{R}$$

et

$$S_m = Vect\{\phi_{m,k}, k \in \mathbb{Z}\}.$$

La collection de modèles considérée est constituée de modèles S_m où m appartient à une grille de pas $1/M$, M étant un entier fixé. Plus précisément,

$$\mathcal{M}_n := \{S_m, m \in \frac{1}{M}\mathbb{N}, m \leq N_n\} \tag{1.16}$$

avec $N_n \leq n$.

Un simple calcul montre que pour tout m,

$$S_m = \left\{f \in L^2(\mathbb{R}), Supp(f^*) \subset [-\pi m, \pi m]\right\}$$

où f^* désigne la transformée de Fourier de f. Cette caractérisation permer de vérifier la propriété de connexion de normes avec $D_m = m$, bien que les modèles $\{S_m\}$ ne soient pas de

dimension finie. De plus, au Chapitre 3, l'étude biais-variance d'un estimateur de densité construit sur un modèle S_m met en évidence un terme de variance d'ordre m/n. D'une manière générale, l'entier m joue ici un rôle analogue à celui de la dimension de S_m pour les espaces de dimension finie.

La structure de cette collection s'apparente à la construction des ondelettes, mais la fonction ϕ ne vérifie pas la condition 2) (décroissance rapide) de la définition des ondelettes. De plus, on peut montrer qu'aucune fonction engendrant les espaces S_m ne vérifie cette condition (cf Meyer (1990), Chapitre 2, Section 2). Ainsi, la base sinus-cardinal ne vérifie pas les propriétés propres aux bases d'ondelettes.

1.2.6 Espaces de régularités

Les définitions classiques de la régularité d'une fonction font intervenir ses dérivées successives, à travers un paramètre de régularité $\beta > 0$ et un rayon $L > 0$ qui majore une certaine norme de la fonction. Les classes de régularité les plus simples sont les classes C^k des fonctions k fois dérivables, pour $k \in \mathbb{N}$. On introduit

$$\tilde{W}(k, L) = \{f \in \mathcal{C}^k(I) \cap L^2(I), \|f^{(k)}\| \leq L\}$$

où $\|.\|$ désigne la norme L^2 et I est un intervalle de $\mathbb{R}$. Par ailleurs, soit $f \in \mathcal{C}^k(I) \cap L^2(I)$, en notant f^* la transformée de Fourier de la fonction f,

$$\left(f^{(k)}\right)^*(\lambda) = \lambda^k f^*(\lambda), \quad \forall \lambda \in \mathbb{R}.$$

D'après l'égalité de Parseval,

$$\|f^{(k)}\|^2 = \frac{1}{2\pi} \int_{\mathbb{R}} \lambda^{2k} |f^*(\lambda)|^2 d\lambda.$$

On généralise donc la définition des espaces $\tilde{W}$ ci-dessus en définissant, pour tout $\beta > 0$ et $L > 0$, l'espace de Sobolev :

$$W(\beta, L) = \left\{ f \in L^2(I), \frac{1}{2\pi} \int_{\mathbb{R}} \lambda^{2\beta} |f^*(\lambda)|^2 d\lambda \leq L^2 \right\}.$$

Par ailleurs, cette définition est invariante si on modifie f sur un ensemble de mesure de Lebesgue nulle, à l'inverse de $\tilde{W}(\beta, L)$.

Contrairement aux espaces de Sobolev dont la définition fait appel à des propriétés globales de la fonction, la définition des espaces de Hölder fait intervenir le comportement local de la fonction. Soit $\beta \in]0, 1]$, et $L > 0$, on pose

$$\mathcal{H}(\beta, L) = \{f : I \to \mathbb{R}, |f(x) - f(y)| \leq L|x - y|^\beta, \forall x, y \in I\}.$$

Cette définition n'est plus valable pour $\beta > 1$. En effet, soit $\beta > 1$ et f une fonction telle que $|f(x) - f(y)| \leq L|x - y|^\beta$ pour tout $x, y \in I$. Alors

$$|f'(y)| = \lim_{x \to y} \left| \frac{f(x) - f(y)}{x - y} \right| \leq L \lim_{x \to y} |x - y|^{\beta - 1} = 0$$

donc f est une fonction constante. Cette difficulté est contournée en définissant, pour tout $\beta > 1$,

$$\mathcal{H}(\beta, L) = \{f : I \to \mathbb{R}, |f^{(r)}(x) - f^{(r)}(y)| \leq L|x - y|^{\beta - r}, \forall x, y \in I\}$$

où r est le plus grand entier strictement inférieur à β.

Les espaces de Besov sont des espaces de régularité plus généraux, qui englobent les espaces de Sobolev et de Hölder. Soit f une fonction d'un intervalle I dans $\mathbb{R}$, on définit le module de continuité $\omega(f, .)$ qui mesure les accroissements de f :

$$\omega(f, t) = \sup_{x, y \in I, |x - y| < t} |f(x) - f(y)|, \quad \forall t > 0.$$

$\omega(f, t)$ s'exprime également à l'aide de la différence de premier ordre de f : pour tout $h > 0$, $\Delta_h(f, x) = f(x + h) - f(x)$.

$$\omega(f, t) = \sup_{0 \leq h \leq t} \sup_{x \in I} |\Delta_h(f, x)| = \sup_{0 \leq h \leq t} \|\Delta_h(f, .)\|_\infty.$$

La notion de module de continuité est ainsi généralisable aux espaces L^p, pour tout $p \in]0, +\infty]$.

$$\omega(f, t)_p = \sup_{0 \leq h \leq t} \|\Delta_h(f, .)\|_p.$$

Afin de mesurer des régularités supérieures, on définit le module de régularité de f. Pour tout $r \in \mathbb{N}^*$, soit Δ_h^r la différence d'ordre r,

$$\Delta_h^r(f, x) = \Delta_h(\Delta_h^{r-1}, x) = \sum_{k=0}^{r} (-1)^{r-k} f(x + kh)$$

et pour tout $p \in]0, +\infty]$,

$$\omega_r(f, t)_p = \sup_{0 \leq h \leq t} \|\Delta_h^r(f, .)\|_p.$$

Enfin, on définit la semi-norme de Besov qui dépend de 3 paramètres de régularité : $\beta > 0$, $p \in]0, +\infty]$ et $q \in]0, +\infty]$,

$$|f|_{\mathcal{B}^\beta_{p,q}} = \begin{cases} \left(\int_0^\infty \left[t^{-\beta} \omega_r(f, t)_p \right]^q \frac{dt}{t} \right)^{1/q} & \text{si} \quad 0 < q < \infty \\ \\ \sup_{t > 0} t^{-\beta} \omega_r(f, t)_p & \text{si} \quad q = +\infty \end{cases} \tag{1.17}$$

où r est le plus petit entier strictement supérieur à β. L'espace de Besov $\mathcal{B}^{\beta}_{p,q}$ est l'ensemble des fonctions f telles que $|f|_{\mathcal{B}^{\beta}_{p,q}} < \infty$. De plus, on définit la boule de Besov $\mathcal{B}^{\beta}_{p,q}(L)$ de rayon $L > 0$:

$$\mathcal{B}^{\beta}_{p,q}(L) = \{f \in L^p(I), |f|_{\mathcal{B}^{\alpha}_{p,q}} + \|f\|_p \leq L\}.$$

Proposition 1.2.2 *Pour tout $\beta > 0$, $L > 0$*

$$\begin{array}{ll} (i) & \mathcal{H}(\beta, L) \subset \mathcal{B}^{\beta}_{\infty,\infty} \quad et \quad |f|_{\mathcal{B}^{\beta}_{\infty,\infty}} \leq L, \quad \forall f \in \mathcal{H}(\beta, L) \\ (ii) & W(\beta, L) \subset \mathcal{B}^{\beta}_{2,\infty} \quad et \quad |f|_{\mathcal{B}^{\beta}_{2,\infty}} \leq L, \quad \forall f \in W(\beta, L) \end{array}$$

Preuve de la Proposition 1.2.2**.**

Ces résultats sont basés sur la propriété suivante, vérifiée par le module de régularité (cf DeVore and Lorentz (1993)): pour toute fonction f dérivable r fois et pour tous entiers $0 < k < r$,

$$\omega_r(f, t)_p \leq t^k \omega_{r-k}(f^{(k)}, t), \quad \forall t > 0.$$

(i) Soit $\beta, L > 0$, r le plus grand entier strictement inférieur à β et $f \in \mathcal{H}(\beta, L)$. Pour tout $t > 0$,

$$t^{-\beta}\omega_{r+1}(f, t) \leq t^{r-\beta}\omega_1(f^{(r)}, t) = t^{r-\beta} \sup_{x \in I} |f^{(r)}(x + t) - f^{(r)}(x)| \leq t^{r-\beta} L t^{\beta - r} = L$$

donc $f \in \mathcal{B}^{\beta}_{\infty,\infty}$ et $|f|_{\mathcal{B}^{\beta}_{\infty,\infty}} \leq L$.

(ii) Soit $f \in W(\beta, L)$ et soit r la partie entière de β.

$$\sup_{t>0} t^{-\beta}\omega_{r+1}(f, t)_2 \leq \sup_{t>0} t^{-\beta+r}\omega(f^{(r)}, t)_2$$

et

$$\omega(f^{(r)}, t)_2^2 = \sup_{0 \leq h \leq t} \|f_h^{(r)} - f^{(r)}\|^2$$

où $f_h^{(r)}(x) = f^{(r)}(x + h)$ pour tout x. Or d'après l'égalité de Parseval,

$$\begin{aligned} \|f_h^{(r)} - f^{(r)}\|^2 &= \frac{1}{2\pi} \int_{\mathbb{R}} \left| \left(f_h^{(r)}\right)^* (\lambda) - \left(f^{(r)}\right)^* (\lambda) \right|^2 d\lambda \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} \left| \left(f^{(r)}\right)^* (\lambda) \right|^2 \left| e^{i\lambda h} - 1 \right|^2 d\lambda \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} |f^*(\lambda)|^2 |\lambda|^{2r} (2\sin(\lambda h/2))^2 d\lambda. \end{aligned}$$

D'où

$$\begin{aligned}
&\sup_{t>0} t^{-\beta}\omega_{r+1}(f,t)_2 \\
&\leq \sup_{t>0}\sup_{0\leq h\leq t} h^{-\beta+r}\sqrt{\frac{1}{2\pi}\int_{\mathbb{R}} |f^*(\lambda)|^2 |\lambda|^{2r}(2\sin(\lambda h/2))^2 d\lambda} \\
&\leq \sup_{t>0}\sup_{0\leq h\leq t} \sqrt{\frac{1}{2\pi}\int_{\mathbb{R}} |f^*(\lambda)|^2 |\lambda|^{2r} \left|\frac{(2\sin(\lambda h/2))}{h}\right|^{2(\beta-r)} |2\sin(\lambda h/2)|^{2(1-\beta+r)} d\lambda}
\end{aligned}$$

Par définition de r, $0 < 1-\beta+r \leq 1$ donc

$$\begin{aligned}
\sup_{t>0} t^{-\beta}\omega_{r+1}(f,t)_2 &\leq \sup_{t>0}\sup_{0\leq h\leq t} \sqrt{\frac{1}{2\pi}\int_{\mathbb{R}} |f^*(\lambda)|^2 |\lambda|^{2r} \left|\frac{(2\sin(\lambda h/2))}{h}\right|^{2(\beta-r)} d\lambda} \\
&= \sup_{t>0}\sup_{0\leq h\leq t} \sqrt{\frac{1}{2\pi}\int_{\mathbb{R}} |f^*(\lambda)|^2 |\lambda|^{2\beta} \left|\frac{(\sin(\lambda h/2))}{\lambda h/2}\right|^{2(\beta-r)} d\lambda}
\end{aligned}$$

Or $|\sin u/u| \leq 1$ pour tout $u \in \mathbb{R}$ d'où

$$\sup_{t>0} t^{-\beta}\omega_{r+1}(f,t)_2 \leq \sqrt{\frac{1}{2\pi}\int_{\mathbb{R}} |f^*(\lambda)|^2 |\lambda|^{2\beta} d\lambda} \leq L.$$

1.2.7 Autres méthodes adaptatives

• Méthode de Lepski

Cette méthode, décrite pour la première fois par Lepski (1991) et reprise ensuite dans de nombreux articles, fournit des estimateurs adaptatifs pour le risque ponctuel à partir d'estimateurs à noyaux. On considère une collection $\{\widehat{f}_h, h \in [h_{\min}, h_{\max}]\}$ d'estimateurs à noyaux de type (1.2) et $\{\mathcal{F}(\beta), \beta > 0\}$ des espaces de régularité (Sobolev, Hölder,...) emboités :

$$\mathcal{F}(\beta) \subset \mathcal{F}(\beta'), \quad \forall \beta \leq \beta'.$$

On suppose connue la vitesse minimax adaptative $\psi_{n,\beta}$ sur l'espace $\mathcal{F}(\beta)$, pour tout $\beta > 0$, ainsi que la valeur h_β de la fenêtre telle que l'estimateur $\widehat{f}_{h_\beta}$ converge à la vitesse $\psi_{n,\beta}$. Un critère de sélection permet de déterminer une fenêtre $\widehat{h}$ dans un intervalle discret $[h_{\min}, h_{\max}]$ de telle sorte que l'estimateur $\widehat{f}_{\widehat{h}}$ converge à la vitesse minimax adaptative sur l'ensemble $\{\mathcal{F}(\beta), \beta \in [\beta_{\min}, \beta_{\max}]\}$ où $\beta_{\min}$ et $\beta_{\max}$ correspondent à $h_{\min}$ et $h_{\max}$.

Contrairement aux techniques de sélection de modèles, la méthode de Lepski estime explicitement la régularité β de la fonction. Par ailleurs, les résultats fournis par cette méthode sont asymptotiques, contrairement aux inégalités oracles, mais permettent d'atteindre asymptotiquement des constantes optimales.

• **Seuillage d'ondelettes**

La méthode de seuillage d'ondelettes consiste à estimer la fonction cible dans une base d'ondelettes, pour ne conserver ensuite que les coefficients suffisamment grands (cf Donoho et al. (1996)). Grâce à la définition des espaces de Besov, un niveau de seuillage dépendant des paramètres p et q fournit une méthode d'estimation adaptative sur les espaces $\mathcal{B}^{\beta}_{p,q}$ pour tout $\beta > 0$.

1.3 Données partiellement observées

Cette section présente les modèles étudiés, ainsi que les fonctions auxquelles on s'intéresse dans ce manuscrit.

1.3.1 Le modèle de régression homoscédastique

On considère un échantillon (X_i, Y_i) de couples de variables i.i.d. telle que pour tout i

$$Y_i = b(X_i) + \epsilon_i$$

où les variables $\{\epsilon_i\}$ sont i.i.d. et indépendantes des $\{X_i\}$. Les $\{X_i\}$ sont appelés les designs, les $\{\epsilon_i\}$ les erreurs ou le bruit de régression, et b la fonction de régression. Le but des Chapitres 2 et 3 est d'estimer la densité f de l'erreur de régression à partir de l'échantillon $\{(X_i, Y_i)\}$. Cette étude comporte de nombreuses applications.

Une des problematiques importantes dans l'étude du modèle de régression est la prédiction : l'échantillon observé est $\{(X_i, Y_i)\}$ mais on veut par la suite, en mesurant uniquement X, obtenir des informations sur le comportement de Y. Celui-ci est caractérisé par deux fonctions.

⋆ La fonction de régression $b(x) = \mathbb{E}[Y|X = x]$ détermine le comportement moyen de Y connaissant X.

⋆ la densité f de ϵ caractérise les fluctuations de Y autour de son espérance conditionnelle.

Ainsi, la construction d'un estimateur de f fournit par exemple des intervalles de confiance pour Y.

L'estimation de la densité des erreurs permet également de construire une procédure de test pour déterminer si la fonction de régression b appartient à une certaine classe de fonctions (linéaires,...).

Certaines méthodes d'estimation de la fonction de régression sont basées sur l'hypothèse d'une erreur gaussienne. L'estimation de la densité de l'erreur permet de valider cette hypothèse à posteriori.

De nombreux articles sont consacrés à l'estimation de densité, mais la difficulté ici réside dans le fait que les erreurs $\{\epsilon_i\}$ dont on veut estimer la densité ne sont jamais observées. La méthode générale consiste à construire des quantités qui approchent les $\{\epsilon_i\}$ et à leur appliquer une méthode d'estimation de densité comme s'il s'agissait des véritables $\{\epsilon_i\}$. Or $\epsilon = Y - b(X)$, il est donc naturel d'estimer ϵ_i par les résidus $\widehat{\epsilon}_i = Y_i - \widehat{b}(X_i)$ où $\widehat{b}$ est un estimateur de la fonction de régression b.

Malgré la diversité des applications qui en découlent, peu d'articles étudient l'erreur de régression d'un point de vue non paramétrique. Akritas and Keilegom (2001) proposent un estimateur de la fonction de distribution de l'erreur de régression par une méthode de noyaux. Efromovich (2005) s'intéresse à l'estimation de la densité de l'erreur et construit un estimateur très performant au sens où sa vitesse minimax de convergence est égale à celle que l'on obtiendrait en mesurant directement les $\{\epsilon_i\}$. Néanmoins, la procédure qu'il développe, basée sur des propriétés des fonctions trigonométriques est complexe, aussi bien pour la construction de l'estimateur que pour les preuves des résultats de convergence, et nécessite des hypothèses assez contraignantes (par exemple, une régularité 2 pour la densité des erreurs). La procédure d'estimation de l'erreur developpée aux Chapitres 2 et 3 s'appuie essentiellement sur des résulats d'estimation de densité et de fonction de régression.

Plus précisément, on considère un échantillon de taille $2n$:

$$\{(X_{-n}, Y_{-n}), \ldots, (X_{-1}, Y_{-1})\} \cup \{(X_1, Y_1), \ldots, (X_n, Y_n)\} = Z^- \cup Z^+.$$

- **Première étape.** Un estimateur $\widehat{b}$ de b est construit à partir de l'échantillon Z^-.

- **Deuxième étape.** On calcule les résidus de l'échantillon Z^+ :

$$\widehat{\epsilon}_i = Y_i - \widehat{b}(X_i), \quad \forall i = 1, \ldots, n.$$

$\widehat{b}$ étant indépendant de (X_i, Y_i) pour tout $i = 1, \ldots, n$, les résidus $\{\widehat{\epsilon}_i, i = 1, \ldots, n\}$ sont identiquement distribués selon une loi de densité f^- qui ne dépend que de Z^- et sont indépendants conditionnellement à Z^-. Cette propriété est particulièrement utile dans le calcul du risque de l'estimateur de $\widehat{f}$ et justifie la scission de l'échantillon en deux échantillons Z^- et Z^+ indépendants, même si cela divise par deux la taille de l'échantillon intervenant dans la majoration du risque.

• **Troisième étape.** Une méthode d'estimation de densité appliquée aux résidus $\{\widehat{\epsilon}_i, i = 1, \ldots, n\}$ fournit un estimateur $\widehat{f}$ de f.

Cette procédure nécessite donc le choix de deux méthodes d'estimation :

⋆ Une méthode d'estimation de la fonction de régression b,

⋆ Une méthode d'estimation de densité.

Celles-ci engendrent des erreurs de deux natures.

⋆ L'erreur due au remplacement des véritables ϵ_i par les résidus $\widehat{\epsilon}_i = \epsilon_i - (\widehat{b} - b)(X_i)$, qui dépend de l'erreur d'estimation de b,

⋆ L'erreur d'estimation de densité.

En effet,

$$\mathbb{E}\left[\|\widehat{f} - f\|_0^2\right] \leq \underbrace{2\mathbb{E}\left[\|\widehat{f} - f^-\|_0^2\right]}_{\text{erreur d'estimation de densité}} + \underbrace{2\mathbb{E}\left[\|f^- - f\|_0^2\right]}_{\text{erreur d'estimation des résidus}}$$

La majoration du terme

$$\mathbb{E}\left[\|\widehat{f} - f^-\|_0^2\right] = \mathbb{E}\left[\mathbb{E}\left[\|\widehat{f} - f^-\|_0^2 | Z^-\right]\right]$$

requiert une méthode d'estimation de densité adaptée au risque $\|.\|_0^2$. Par ailleurs, pour tout Z^- fixé, $\widehat{\epsilon}$ est la somme de deux variables indépendantes, ϵ et $(b - \widehat{b})(X)$, et par un simple calcul on montre que

$$f^-(x) = \int f(x - (b - \widehat{b})(t)) f_X(t) dt$$

où f_X est la densité de X. Si de plus f est lipschitzienne, pour tout $x \in \mathbb{R}$

$$\begin{aligned}
(f^- - f)^2(x) &= \left(\int \left[f(x - (b - \widehat{b})(t)) - f(x)\right] f_X(t) dt\right)^2 \\
&\leq \int \left[f(x - (b - \widehat{b})(t)) - f(x)\right]^2 f_X(t) dt \\
&\leq \mathrm{Lip}(f)^2 \int (b - \widehat{b})^2(t) f_X(t) dt \\
&= \mathrm{Lip}(f)^2 \|b - \widehat{b}\|_{f_X}^2
\end{aligned}$$

Ainsi nous avons besoin d'une méthode d'estimation de la fonction de régression qui fournisse un majorant du risque intégré $\mathbb{E}\left[\|\widehat{b} - b\|_{f_X}^2\right]$, y compris lorsque l'on considère le risque ponctuel pour l'estimateur $\widehat{f}$.

1.3.2 Données de survie

Les problèmatiques étudiées en analyse de survie sont issues du domaine bio-médical, tout comme le vocabulaire utilisé, mais les résultats obtenus sont applicables à d'autres domaines. Dans ce manuscrit, nous nous intéressons à l'étude du temps de survie, terme générique qui peut désigner le temps de survie d'un patient après une opération, la durée d'une maladie ou tout autre variable positive. Dans la plupart des études bio-médicales, les variables aléatoires auxquelles on s'interesse ne sont pas totalement observées : on parle alors de censure.

a) Censure à droite

Dans la majorité des études médicales auprès de patients, certains sujets quittent l'étude prématurément, et leur temps de survie n'est pas observé. Ainsi, soit Y_i le temps de survie du sujet i après une opération, deux situations sont possibles

-Le sujet i est suivi jusqu'à sa mort, Y_i est donc mesuré.

-Le sujet i quitte l'étude ou meurt d'une cause non imputable à l'opération (accident...) au bout d'un temps C_i, appelé temps de censure.

Ce phénomène est modélisé de la façon suivante : pour chaque patient i, il existe un temps de survie Y_i et un temps de censure C_i supposés indépendants. On observe le minimum de ces deux temps, et on sait également si ce minimum correspond au temps de censure ou au temps de survie. Ainsi, à partir d'un panel de n patients supposés indépendants, on observe l'échantillon

$$\left(T_i = \min(Y_i, C_i), \delta_i = 1_{\{Y_i \leq C_i\}}\right)_{i=1,\dots,n}. \tag{1.18}$$

On suppose généralement que les temps de survie et de censure sont indépendants (ou indépendants conditionnellement aux covariables s'il y en a). Cette supposition est discutable dans les applications pratiques, et d'autres modèles prenant en compte une dépendance entre ces deux variables existent. Néanmoins, comme dans de nombreux problèmes, le cas indépendant constitue un modèle d'étude intéressant.

L'idée naïve consistant à ne conserver que les temps de survie des patients non censurés conduit clairement à des résultats biaisés : en effet, plus un patient survit longtemps, plus il a de chance d'être censuré, donc réciproquement l'espérance de vie des patients non censurés est plus basse que celle des patients censurés. Il est donc nécessaire de considérer des méthodes d'estimation qui incluent l'ensemble des données, censurées et non censurées. D'un point de vue statistique, l'étude de données censurées à droite débute avec Kaplan and Meier (1958), qui proposent une méthode d'estimation de la fonction de survie à partir d'observations censurées à droite, encore très utilisée dans les travaux actuels.

Dans ce manuscrit, nous nous intéressons plus précisément à l'estimation du taux de risque instantané ou *hazard rate*, noté h. Le taux de risque instantané à l'instant x est la probabilité

que le patient meure juste après x, sachant qu'il était vivant juqu'au temps x. Plus précisément,

$$h(x) = \frac{P[Y \in [x, x+dx] | Y \geq x]}{dx} = \frac{P\left[Y \in [x, x+dx] \cap \{Y \geq x\}\right]}{P[Y \geq x]} = \frac{f_Y(x)}{\overline{F}_Y(x)}$$

où f_Y et $\overline{F}_Y$ désigne la densité et la fonction de survie de Y.

Certains estimateurs de h présents dans la litterature sont construits comme quotients d'estimateurs de la densité f_Y et de la fonction de distribution $\overline{F}_Y$, par exemple avec l'estimateur de Kaplan-Meier. h s'exprime également sous forme d'un quotient différent : $T = \min(Y, C)$ donc pour tout $x \in \mathbb{R}_+$,

$$\overline{F}_T(x) = P\left[\{Y \geq x\} \cap \{C \geq x\}\right] = P\left[Y \geq x\right] P\left[C \geq x\right] = \overline{F}_Y(x)\overline{F}_C(x),$$

d'où

$$h(x) = \frac{f_Y(x)\overline{F}_C(x)}{\overline{F}_T(x)} = \frac{\psi(x)}{\overline{F}_T(x)}. \tag{1.19}$$

La fonction $\overline{F}_T$ est plus facile à estimer que $\overline{F}_Y$ car les $\{T_i\}$ sont intégralement observées. Par ailleurs, la fonction ψ appelée sous-densité de Y correspond heuristiquement à la "densité" des variables $\{Y_i\}$ effectivement observées. En effet, pour toute fonction $t : \mathbb{R}^+ \to \mathbb{R}$ telle que $t(0) = 0$,

$$\begin{aligned}
\mathbb{E}[t(\delta T)] &= \mathbb{E}[\delta t(Y)] \\
&= \mathbb{E}\left[\mathbb{E}[1_{\{Y \leq C\}} t(Y) | Y]\right] \\
&= \mathbb{E}\left[t(Y)\overline{F}_C(Y)\right] \\
&= \int t(x)\psi(x)dx.
\end{aligned}$$

L'expression (1.19) permet de construire des estimateurs comme quotients d'estimateurs de ψ et $\overline{F}_T$. Enfin, on peut remarquer que $f_Y(x) = -\overline{F}'_Y(x)$,

$$h(x) = -\left(\ln \overline{F}_Y(x)\right)'.$$

Cette écriture donne lieu à un troisième type de procédure d'estimation : on calcule un estimateur de $\overline{F}_Y$ (généralement discontinu), puis on lui applique une méthode de dérivation discrète.

Une synthèse bibliographique sur ce sujet est présentée au Chapitre 4.

Les procédures présentées ci-dessus comportent deux étapes (auxquelles s'ajoute éventuellement l'étape de sélection de modèle, de fenêtre...). A l'inverse, l'estimateur présenté au Chapitre 4 est calculé directement en minimisant un contraste empirique de type projection. La construction

de ce contraste repose sur le choix de la norme $\|.\|_{\overline{F}_T}$ qui est la norme associée à la fonction de distribution de T. En effet, on montre que

$$\mathbb{E}[\delta_i t(Y_i)] = \int t(x)h(x)\overline{F}_T(x)dx \quad \forall t \in L^2$$

et de façon évidente

$$\mathbb{E}\left[\int t^2(x)\mathbb{1}_{\{T_i \geq x\}}dx\right] = \|t\|^2_{\overline{F}_T},$$

ce qui fournit un estimateur empirique de

$$\|t\|^2_{\overline{F}_T} - 2\int t(x)h(x)\overline{F}_T(x)dx = \|h-t\|^2_{\overline{F}_T} - \|h\|^2_{\overline{F}_T}.$$

Par ailleurs, la minimisation de ce contraste conduit à des estimateurs de la forme suivante: soit S_m un modèle, le vecteur $\widehat{A}_m$ des coordonnées de $\widehat{h}_m$ vérifie une relation de la forme

$$\widehat{G}_m\widehat{A}_m = \widehat{V}_m \tag{1.20}$$

où $\widehat{G}_m$ est la matrice de Grahm de la base de S_m considérée, pour la norme empirique :

$$\frac{1}{n}\sum_{i=1}^{n}\int t^2(x)\mathbb{1}_{\{T_i \leq x\}}dx.$$

Ce type d'estimateur est semblable aux estimateurs de la fonction de régression calculés à l'aide d'un contraste des moindres carrés. On parle souvent d'estimateurs "de type régression", dans un cadre différent, l'estimateur construit au Chapitre 6 est également de cette forme. Nous nous intéressons, dans le Chapitre 5, à une généralisation de la méthode de sélection de modèles ponctuelle pour ce type d'estimateurs, en mettant en exergue les points fondamentaux de la preuve developpée au Chapitre 3.

Si l'on considère un modèle S_m d'histogrammes, l'équation (1.20) qui détermine $\widehat{A}_m$ possède une solution explicite car la matrice de Gram $\widehat{G}_m$ est diagonale. Mais comme nous le verrons dans les chapitres suivants, les bases d'histogrammes sont peu performantes pour estimer des fonctions régulières (plus précisément, elles ne permettent d'obtenir une vitesse minimax que sur des espaces de régularité inférieure à 1). Nous pallions cet inconvenient en considèrant une collection de modèles vérifiant la propriété de localisation forte et les matrices de Gram $\widehat{G}_m$ s'écrivent alors sous forme de matrices diagonales par blocs avec une dimension de blocs borné indépendamment de n.

Par ailleurs, l'étude de données censurées à droite s'inscrit dans un cadre plus général d'étude de processus de comptage. Plus précisément, soit N et Z des processus de comptage sur $\mathbb{R}^+$, et $(\mathcal{F}_t)_{t\geq 0}$ une filtration. On définit le compensateur Γ de N par rapport à $(\mathcal{F}_t)_{t\geq 0}$ comme le

processus tel que $N - \Gamma$ soit une $(\mathcal{F}_t)$-martingale. On s'intéresse généralement à l'estimation de l'intensité, c'est à dire la fonction α telle que

$$\Gamma(t) = \int_0^t \alpha(z) Z(z) dz, \quad \forall t \geq 0.$$

Ce modèle est étudié de façon particulièrement détaillée dans Andersen et al. (1993). Il s'applique à notre cadre d'étude en considérant

$$\begin{aligned} N(z) &= \mathbb{1}_{\{T \leq z, \delta = 1\}} \\ Z(z) &= \mathbb{1}_{\{T \geq z\}} \\ \mathcal{F}_t &= \sigma(N_u, u \leq t) \end{aligned}$$

où $T = \min(Y, C)$ et $\delta = \mathbb{1}_{\{Y \leq C\}}$ sont les variables définies précédemment. Alors, $\alpha(z) = h(z)$. En effet, on vérifie que

$$\Gamma(t) = \int_0^t h(z) Z(z) dz$$

est bien le compensateur de N par rapport à $(\mathcal{F}_t)$, c'est à dire

$$\mathbb{E}[N(t) - \Gamma(t) | \mathcal{F}_s] = N(s) - \Gamma(s), \quad \forall s < t. \tag{1.21}$$

Soit $s < t$. Si $N(s) = 1$, alors $T \leq s \Rightarrow T \leq t$ p.s. donc $N(t) = 1$, d'où $\mathbb{E}[N(t)\mathbb{1}_{\{N(s)=1\}} | \mathcal{F}_S] = 1 = N(s)$. De plus, $Z(z) = 0$ pour tout $z > s$ donc

$$\mathbb{E}[(\Gamma(t) - \Gamma(s))\mathbb{1}_{\{N(s)=1\}} | \mathcal{F}_s] = \int_s^t h(z) Z(z) dz = 0$$

$$\begin{aligned} \text{et} \quad \mathbb{E}[(N(t) - N(s))\mathbb{1}_{\{N(s)=0\}} | \mathcal{F}_s] &= P\left[\{Y \in [s,t]\} \cap \{\delta = 1\} | \mathcal{F}_s\right] \mathbb{1}_{\{N(s)=0\}} \\ &= \int_s^t f_Y(z) P[C \geq z | \mathcal{F}_s] \mathbb{1}_{\{N(s)=0\}} dz \\ &= \int_s^t h(z) P[\min(C, Y) \geq z | \mathcal{F}_s] \mathbb{1}_{\{N(s)=0\}} dz \\ &= \int_s^t h(z) \mathbb{E}[Z(z) | \mathcal{F}_s] \mathbb{1}_{\{N(s)=0\}} dz \\ &= \mathbb{E}\left[\Gamma(t)\mathbb{1}_{\{N(s)=0\}} | \mathcal{F}_s\right] - \Gamma(s)\mathbb{1}_{\{N(s)=0\}} \end{aligned}$$

ce qui prouve (1.21).

L'étude de processus de comptage de cette forme fournit donc également des méthodes d'estimation du risque instantané en présence de censure à droite.

b) Censure par intervalle, cas I

Considérons le temps de contamination Y d'un patient par un virus. Ce temps n'est en général jamais directement observé. En effet, un test de dépistage effectué au temps T indique si le patient a été contaminé avant l'instant T ou non, c'est à dire si Y appartient à l'intervalle $[0, T]$ ou $]T, +\infty[$. Ce type d'observation est appelé censure par intervalle, cas I ou *current status data* car l'information dont on dispose est le statut du sujet (contaminé ou non) à l'instant T. Ce type de données est largement étudié depuis une vingtaine d'année, on peut en particulier citer Groeneboom and Wellner (1992) qui consacrent plusieurs articles à ce sujet. La plupart des estimateurs non paramétriques présents dans la littérature sont basés sur le maximum de vraisemblance, auquel sont appliquées des méthodes de régularisation prenant en compte la régularité de la fonction (cf Chapitre 4 pour une bibliographie plus détaillée). Néanmoins, très peu d'articles s'intéressent à l'adaptativité sur des espaces de régularité.

Plus précisément, dans ce manuscrit, nous nous intéressons à un temps de survie Y dépendant d'une covariable $X \in \mathbb{R}$ avec un temps d'observation T dépendant également de X, et l'on suppose Y et T indépendants conditionnellement à X. On observe alors le triplet

$$(X, T, \delta = 1_{\{Y \leq T\}}).$$

Au Chapitre 6, nous présentons un estimateur adaptatif de la fonction de répartition de Y conditionnellement à X :

$$F(x, y) = P[Y \leq y | X = x]$$

construit à partir d'un échantillon i.i.d. $(X_i, T_i, \delta_i)_{i=1,\dots,n}$. On remarque que

$$\mathbb{E}[\delta | X, T] = \mathbb{E}[1_{\{Y \leq T\}} | X, T] = \mathbb{E}[F(X, T) | X, T] = F(X, T)$$

Ainsi, F est la fonction de régression de δ sur (X, T). Grâce à cette observation, nous proposons une méthode d'estimation par minimisation d'un contraste de type régression.

Une des particularités du contexte de censure par intervalle est la vitesse de convergence de l'estimateur de la fonction de survie.En effet, dans un contexte non censuré, ou censuré à droite, en l'absence de covariable la fonction de survie est estimée à vitesse paramétrique $(1/n)$, et la fonction de distribution conditionnelle converge à une vitesse qui ne dépend que de la régularité de F par rapport à x. Par contre, dans le cadre de la censure par intervalle, la vitesse minimax d'estimation dépend de la régularité de F par rapport à x et y, comme nous le prouvons par une étude minimax au Chapitre 6.

Première partie

Estimation de la densité de l'erreur de régression

Chapitre 2

Estimation de l'erreur de régression pour le risque quadratique intégré

Ce chapitre présente un estimateur de la densité de l'erreur de régression ϵ dans le modèle de régression homoscedastique

$$Y = b(X) + \epsilon$$

par sélection de modèle globale. Seules les variables X et Y sont observées, on construit donc, à partir d'un échantillon (X_i, Y_i), des quantités qui estiment les variables (ϵ_i). La procédure est la suivante : l'échantillon observé (X_i, Y_i) est séparé en deux échantillons indépendants de même taille : Z^- et Z^+. Un estimateur $\tilde{b}$ de la fonction de regression est calculé à partir de l'échantillon Z^-, selon une méthode proposée par Baraud (2002). En remarquant que $\epsilon = Y - b(X)$, les erreurs (ϵ_i) peuvent être estimées par les résidus du deuxième échantillon : $(\widehat{\epsilon_i} = Y_i - \tilde{b}(X_i))$ pour tout (X_i, Y_i) de Z^+. Ces variables approchent d'autant mieux les véritables (ϵ_i) que $\tilde{b}$ est un bon estimateur de b. Finalement, nous appliquons aux résidus une méthode d'estimation de densité developpée par Massart (2007) comme s'il s'agissait des véritables (ϵ_i) et obtenons un estimateur $\widehat{f}^-_{\widehat{m}}$.

Le risque quadratique intégré de l'estimateur $\tilde{f}^-_{\widehat{m}}$ se décompose en deux termes : un terme d'erreur d'estimation de densité et un terme dû au remplacement des erreurs par les résidus $\widehat{\epsilon}_i$. En remarquant que les résidus $(\widehat{\epsilon_i})$ sont i.i.d. conditionnellement à l'échantillon Z^-, le premier terme est majoré grâce à l'inégalité oracle vérifiée par l'estimateur de densité dans Massart (2007). Le deuxième terme s'exprime en fonction du risque de l'estimateur $\tilde{b}$ de b, qui est majoré par un résultat de Baraud (2002).

Le principal résultat présenté dans ce chapitre fait l'objet d'une note publiée dans les Comptes Rendus de l'Académie des Sciences (Plancade (2008)).

2.1 Introduction

Consider an i.i.d. sample

$$Z = (X_i, Y_i)_{i\in\{-n,\dots,-1\}\cup\{1,\dots,n\}} \tag{2.1}$$

from the homoscedastic regression framework

$$Y_i = b(X_i) + \epsilon_i, \tag{2.2}$$

where the $\{\epsilon_i\}$'s are i.i.d. random variables of common density f, independent of the $\{X_i\}$'s, and such that $\mathbb{E}[\epsilon_i] = 0$ for every i. In this chapter, we want to build an estimator $\widehat{f}$ of the density f of the errors by a model selection procedure adapted to the integrated risk

$$\mathbb{E}\left[\int (\widehat{f} - f)^2(x)dx\right].$$

by a model selection procedure. The specificity of error density estimation is that the errors $\{\epsilon_i\}$'s are unobserved. The classical procedure (presented with more details in Section 1.3.1) is the following. The sample Z is split in two independent samples:

$$Z^- = (X_i, Y_i)_{i-n,\dots,-1} \quad \text{and} \quad Z^+ = (X_i, Y_i)_{i=1,\dots,n}$$

$\star$ An estimator $\tilde{b}$ of the regression function b is computed from the sample Z^-.

$\star$ As the errors $\{\epsilon_i\}$'s are unobserved, we build quantities which estimate them. According to (2.2), $\epsilon_i = Y_i - b(X_i)$ hence the residuals from the sample Z^+, $\{\widehat{\epsilon}_i = Y_i - \tilde{b}(X_i), i = 1,\dots,n\}$, are relevant estimators of the $(\epsilon_i)_{i=1,\dots,n}$. Besides, the samples Z_+ and Z^- are independent, so the $(\widehat{\epsilon}_i)_{i=1,\dots,n}$ are i.i.d. given Z^-, thus we can apply them a density estimation procedure developed for i.i.d. samples.

$\star$ We apply a density estimation procedure to the residuals as if they were the true $\{\epsilon_i\}$'s.

The litterature about regression noise estimation is not extensive, and the non parametric estimation of error density is especially studied by Efromovich (2005). He follows the same general outline as we do, but with different estimators of density and regression function. He obtains a more powerful result but under more restrictive assumption. Moreover, our estimator is easier to implement. A parallel between his estimator and the one presented in this chapter is developed at the end of Section 2.4. Other papers are devoted to the estimation of the cumulative distribution function of regression errors, like Akritas and Keilegom (2001) and Kiwitt et al. (2008).

The chapter is organised as follows. Section 2.2 presents the notations and assumptions. Section 2.3 is devoted to the preliminary estimators of density and regression function: we present the estimation procedures and the performance of the estimators. In Section 2.4, we apply these procedures to compute the error density estimator following the outline described above, and expose the main result. Section 2.5 presents numerical example and the proofs are gathered in Section 2.6.

2.2 Notations and Assumptions

2.2.1 Notations

Let t, $s \in L^2(\mathbb{R}) \cap L^\infty(\mathbb{R})$, and ν a density on $\mathbb{R}$, we consider the following norms:

$$\|t\| = \sqrt{\int t^2(x)dx}, \quad \|t\|_\nu = \sqrt{\int t(x)^2\nu(x)dx}, \quad \|t\|_\infty = \sup_{t\in\mathbb{R}} |t(x)|,$$

and the scalar products

$$\langle t, s\rangle = \int t(x)s(x)dx, \quad \langle t, s\rangle_\nu = \int t(x)s(x)\nu(x)dx.$$

We consider the empirical norm and scalar product associated to the sample $(X_{-n}, \ldots, X_{-1})$:

$$\|t\|_n = \sqrt{\frac{1}{n}\sum_{i=-n}^{-1} t(X_i)} \quad \text{and} \quad \langle t, s\rangle_n = \frac{1}{n}\sum_{i=-n}^{-1} t(X_i)s(X_i).$$

We denote by t^* the Fourier transform of t, namely

$$t^*(\lambda) = \int_{\mathbb{R}} t(x)e^{-i\lambda x}dx, \quad \forall\lambda \in \mathbb{R}.$$

For every $x \in \mathbb{R}$, we denote by $\lfloor x \rfloor$ the integer part of x i.e. the largest integer smaller than or equal to x.

In the whole chapter, C, C' and C" denote constants that may change from one line to the other.

2.2.2 General assumptions

We note that the regression function $b(x) = \mathbb{E}[Y|X = x]$ is defined on the set $\{x \in \mathbb{R}, f_X(x) > 0\}$, but in most regression studies, b is estimated on a more restricted set: $\{x \in \mathbb{R}, f_X(x) \geq m_0\}$ where m_0 is a positive number. Indeed b can not be well estimated on a set where the $\{X_i\}$'s are not dense enough.

More precisely, we assume that the following assumption holds.

($\mathbf{H_{gen}}$) : The density f_X of the $\{X_i\}$'s is supported on $[0, 1]$ and there exists $m_0 \in \mathbb{R}_+^*$ such that

$$m_0 \leq f_X(x) \quad \forall x \in [0, 1].$$

The regression function $b \in L^2([0, 1]) \cap L^\infty([0, 1])$.

Moreover, the error density f is supposed to satisfy one of the following alternative conditions.

$(\mathbf{H_{err(1)}})$: $f \in L^2([-1,1])$ and is Lipschitz on $[-1,1]$ with Lipschitz constant $\mathrm{Lip}(f)$.

$(\mathbf{H_{err(2)}})$: $f \in L^2(\mathbb{R})$, and $\lim_{x\to+\infty} f(x) = \lim_{x\to-\infty} f(x) = 0$. Moreover, f is differentiable and its derivative f' belongs to $L^2(\mathbb{R})$.

2.2.3 Assumptions about the collections of models

In the following sections, we present several model selection estimators which involve several collections of models. Thus the assumptions about the collections of models, pooled in this section, are expressed for a general collection Σ_n. These properties are presented with more details in Section 1.2.5 and we only recall them for the reader's convenience. Let

$$\Sigma_n = \{S_m, m \in J_n\}$$

be a collection of linear subsets of $L^2(\mathbb{R})$ and $\{D_m, m \in J_n\}$ be a collection of positive integers. For every $m \in J_n$, let $\{\phi_k^m, k \in I_m\}$ be a L^2-orthogonal basis of S_m.

$(\mathbf{H_{glob}(\Sigma_n)})$: There exists a model $m_n \in J_n$ such that, for every $m \in J_n$,

$$S_m \subset S_{m_n}.$$

We denote by $S_n = S_{m_n}$, and $N_n = D_{m_n}$. Moreover, there exists positive constants Γ and R such that

$$Card\left(\{m \in J_n, D_m = D\}\right) \leq \Gamma D^R \quad \forall D \in \mathbb{N}^*. \tag{2.3}$$

$(\mathbf{H_{con}(\Sigma_n)})$: There exists a positive number K such that for every $m \in J_n$ and for every $t \in S_m$,

$$\|t\|_\infty \leq K\sqrt{D_m}\|t\|. \tag{2.4}$$

According to Proposition 1.2.1 in the Introduction, (2.4) is equivalent to

$$\sum_{k\in I_m} \left(\phi_k^m(x)\right)^2 \leq K^2 D_m, \quad \forall x \in I.$$

The following property is only defined for finite dimensional models.

$(\mathbf{H_{loc}(\Sigma_n)})$: For every $m \in J_n$, $I_m = \{1,\dots,D_m\}$ and there exists a partition $\{I_1,\dots,I_s\}$ of $\{1,\dots,D_m\}$ such that

1) $Card(I_k) \leq K, \quad \forall k = 1,\dots,s.$

2) For every $i \neq j$ in $\{1,\dots,s\}$, and for every $k \in I_i$, $l \in I_j$,

$$\phi_k^m \phi_l^m = 0 \quad \text{and} \quad \|\phi_k^m\|_\infty \leq K\sqrt{D_m}.$$

2.3 Preliminary results: density estimator and regression function estimator

2.3.1 Density estimator

In this section, we present the density estimation procedure developed by Massart (2007), which is applied to the residuals in Section 2.4. Let $(V_1, \ldots, V_n)$ be a sample of i.i.d. random variables with common density $g \in L^2(I)$ where I is an interval $\mathbb{R}$. It is usual in theoretical studies to consider that the interval which supports the target function is known but in practical examples this compact is often determined from the data. Let

$$\mathcal{M}_n = \{S_m, m \in J_n\}$$

be a collection of linear models and for every $m \in J_n$, let $\{\phi_k^m, k \in I_m\}$ be a $\|.\|$-orthonormal basis of S_m.

a) Projection estimators

We note that g is the minimiser of γ on $L^2(I)$ where

$$\gamma(t) = \|t - g\|^2 - \|g\|^2 = \|t\|^2 - 2\langle t, g\rangle, \quad \forall t \in S_m.$$

Besides, for every $t \in L^2(I)$, $\mathbb{E}[t(V_i)] = \langle t, g\rangle$ so $\langle t, g\rangle$ is estimated by the empirical mean $(1/n)\sum_{i=1}^n t(V_i)$ and $\gamma(t)$ is estimated by the following empirical contrast:

$$\gamma_n(t) = \|t\|^2 - \frac{2}{n}\sum_{i=1}^n t(V_i).$$

Then, for every $m \in S_m$, let

$$\widehat{g}_m = \arg\min_{t \in S_m} \gamma_n(t) = \sum_{k \in I_m} \widehat{a}_k^m \phi_k^m. \tag{2.5}$$

Writting that the partial derivative of $\gamma_n(t)$ with respect to each coordinate of t is equal to 0, we obtain: $\widehat{a}_k^m = (1/n)\sum_{i=1}^n \phi_k^m(V_i)$ for every $k \in I_m$. Besides, we denote by

$$g_m = \sum_{k \in I_m} a_k^m \phi_k^m \quad \text{with} \quad a_k^m = \langle g, \phi_k^m\rangle$$

the orthogonal projection of g on S_m.

b) Bias-variance decomposition and model selection

For every $m \in J_n$, by Pythagoras formula, the risk of $\widehat{g}_m$ decomposes as follows.

$$\mathbb{E}\left[\|\widehat{g}_m - g\|^2\right] = \|g - g_m\|^2 + \mathbb{E}\left[\|\widehat{g}_m - g_m\|^2\right] = \gamma(g_m) + \sum_{k \in I_m} \mathbb{E}\left[(\widehat{a}_k^m - a_k^m)^2\right] + \|f\|^2$$

and $\|f\|^2$ is independent of m. The best model in the collection is the one which minimises $\gamma(g_m) + \sum_{k \in I_m} \mathbb{E}\left[(\widehat{a}_k^m - a_k^m)^2\right]$. Thus, this unknown quantity is estimated and the model which minimises this estimator is selected. For every $k \in I_m$,

$$\begin{aligned}
\mathbb{E}\left[(\widehat{a}_k^m - a_k^m)^2\right] &= Var(\widehat{a}_k^m) \\
&= Var\left(\frac{1}{n}\sum_{i=1}^{n} \phi_k^m(V_i)\right) \\
&= \frac{1}{n} Var\left(\phi_k^m(V_1)\right) \\
&\leq \frac{1}{n}\mathbb{E}\left[(\phi_k^m(V_1))^2\right]. \qquad (2.6)
\end{aligned}$$

Hence under Assumption $(\mathbf{H_{con}}(\mathcal{M}_\mathbf{n}))$,

$$\mathbb{E}\left[\|\widehat{g}_m - g_m\|^2\right] \leq \frac{1}{n}\sum_{k \in I_m} \mathbb{E}\left[(\phi_k^m(V_1))^2\right] \leq K^2 \frac{D_m}{n}.$$

Moreover, $\gamma(g_m)$ is naturally estimated by $\gamma_n(\widehat{g}_m)$ and we select the model

$$\widehat{m} = \arg\min_{m \in J_n} \left\{\gamma_n(\widehat{g}_m) + pen(m)\right\}$$

where $pen(m) = \theta K^2 D_m / n$ for some constant $\theta > 1$.

c) Result

Theorem 2.3.1 *Assume that* $(\mathbf{H_{con}}(\mathcal{M}_\mathbf{n}))$ *and* $(\mathbf{H_{glob}}(\mathcal{M}_\mathbf{n}))$ *hold, then for every* $\theta > 1$

$$\mathbb{E}\left[\|\widehat{g}_{\widehat{m}} - g\|^2\right] \leq C_1 \inf_{m \in J_n} \left\{\|g - g_m\|^2 + pen(m)\right\} + \frac{C_2 \|g\|}{n}$$

where C_1 *is an absolute constant and* C_2'' *depends on* (K, R, Γ) *defined in* $(\mathbf{H_{con}}(\mathcal{M}_\mathbf{n}))$ *and* $(\mathbf{H_{glob}}(\mathcal{M}_\mathbf{n}))$. *More precisely,*

$$C_2 = C_2'\left(1 + \|g\|\Gamma \sum_{D \in \mathbb{N}^*} D^{R+1/2} \exp\left(-\frac{K\sqrt{D}}{\|g\|}\right)\right)$$

for some absolute constant C_2'.

Comments

1. Theorem 2.3.1 states that the risk of the estimator $\widehat{g}_{\widehat{m}}$ converges as well as the best estimator in the collection $\mathcal{M}_n$, up to a multiplicative constant.

2. Assume that the models $\{S_m\}$'s belong to one of the classical collections described in Introduction, Section 1.2.5 (piecewise polynomials, wavelets...), then if g is in a Besov ball $\mathcal{B}^{\beta}_{p,q}(L)$,
$$\inf_{t\in S_m} \|g-t\|^2 = \|g-g_m\|^2 \leq CD_m^{-2\beta}. \tag{2.7}$$
(See for example DeVore and Lorentz (1993) for trigonometric and piecewise polynomial models, and Meyer (1990) for wavelet basis.) Hence
$$\mathbb{E}\left[\|\widehat{g}_{\widehat{m}}-g\|^2\right] \leq C_1 \inf_{m\in J_n}\left\{CD_m^{-2\beta}+\frac{D_m}{n}\right\}+\frac{C_2\|g\|}{n}.$$
The model m^* which minimises $\left\{CD_m^{-2\beta}+D_m/n\right\}$, called the oracle, satisfies
$$D_{m^*} = C' n^{1/(2\beta+1)}.$$
Thus the better model depends on the unknown regularity of the function g. Moreover,
$$\mathbb{E}\left[\|\widehat{g}_{\widehat{m}}-g\|^2\right] \leq C'' n^{-2\beta/(2\beta+1)}.$$

3. Besides, $n^{-2\beta/(2\beta+1)}$ is the minimax rate of convergence for the quadratic integrated risk in density estimation (see for example Tsybakov (2004)). Thus our density estimator is minimax over every Besov ball.

2.3.2 Regression function estimator

In this section, we propose an estimator of the regression function b presented by Baraud (2002), computed from the sample
$$Z^- = \{X_i, Y_i), i = -n, \ldots, -1\}.$$

Consider a collection of finite dimensional models on $[0,1]$:
$$\mathcal{M}'_n = \{S'_m, m \in J'_n\}$$
and for every $m \in J'_n$, let $\{\psi_k^m, k = 1, \ldots, D'_m\}$ be a $\|.\|$-orthonormal basis of S'_m.

a) Non adaptive estimators

For every $t \in L^2([0,1])$, we consider the least-square empirical contrast

$$\gamma'_n(t) = \frac{1}{n} \sum_{i=-n}^{-1} (Y_i - t(X_i))^2.$$

$\gamma'_n(t)$ measures how the $\{t(X_i)\}$'s approach the $\{Y_i\}$'s. Moreover, for every function $t \in L^2([0,1])$,

$$\mathbb{E}[(Y_1 - t(X_1))^2] = \mathbb{E}[((b-t)^2(X_1) + \epsilon_1)^2] = \|b-t\|^2_{f_X} + \mathbb{E}[\epsilon_1^2] - 2\mathbb{E}[\epsilon_1(b-t)(X_1)].$$

X_1 and ϵ_1 are independent so

$$\mathbb{E}[\epsilon_1(b-t)(X_1)] = \mathbb{E}[\epsilon_1]\mathbb{E}[(b-t)(X_1)] = 0.$$

Therefore,

$$\mathbb{E}[(Y_1 - t(X_1))^2] = \|b-t\|^2_{f_X} + \sigma^2$$

and b minimises $\|b-t\|^2_{f_X}$ which justifies the definition of the empirical contrast γ'_n. Thus, for every $m \in J'_n$, let

$$\widehat{b}_m = \arg\min_{t \in S'_m} \gamma'_n(t)$$

and $b_m = \arg\min_{t \in S_m} \gamma'(t)$.

b) Model selection

For every $m \in J'_n$,

$$\mathbb{E}\left[\|\widehat{b}_m - b\|^2_{f_X}\right] \leq \|b - b_m\|^2_{f_X} + \mathbb{E}\left[\|\widehat{b}_m - b_m\|^2_{f_X}\right].$$

The variance term $\mathbb{E}\left[\|\widehat{b}_m - b_m\|^2_{f_X}\right]$ is more difficult to upper bound than in the density estimation context and we do not do it here. Thus the theoretical penalty function is chosen in order to fit the deviation inequalities involved in the proof of the result. The penalty term which appears is

$$pen_{th}(m) = \sigma^2 \frac{D_m}{n}.$$

where σ^2 is the variance of ϵ_1. But this penalty term can not be plugged in the definition of the estimator since σ^2 is unknown, so we replace it by an estimator $\widehat{\sigma}_n$, defined as follows.

Let W_n be a linear subspace of $L^2[0,1]$ of dimension $Dim(W_n) = \lfloor n/2 \rfloor - 1$ and

$$\widehat{\sigma}^2_n = \frac{n}{n - \lfloor n/2 \rfloor} \left(\inf_{t \in W_n} \frac{1}{n} \sum_{i=1}^{n} (Y_i - t(X_i))^2 \right).$$

Then let

$$\widehat{m} = \arg\min_{m \in J'_n} \left\{ \gamma'_n(\widehat{b}_m) + pen(m) \right\}$$

where

$$pen(m) = \theta' \widehat{\sigma}_n^2 \frac{D_m}{n}$$

and $\theta' > 1$.

Moreover, for computational reasons, the L^2-norm of our estimator has to be bounded, thus we consider the following estimator:

$$\tilde{b} = \begin{cases} \widehat{b}_{\widehat{m}} & \text{if} \quad \|\widehat{b}_{\widehat{m}}\| \leq n \\ \\ 0 & \text{otherwise.} \end{cases} \tag{2.8}$$

c) Result

We consider two alternative assumptions about the collection $\mathcal{M}'_n$.

$(\mathbf{H_1})$: The collection $\mathcal{M}'_n$ satisfies $(\mathbf{H_{loc}}(\mathcal{M}'_\mathbf{n}))$ and $N_n \leq n/\log^2 n$. Moreover, $\mathbb{E}[\epsilon_1^6] < +\infty$.

$(\mathbf{H_2})$: The collection $\mathcal{M}'_n$ satisfies $(\mathbf{H_{con}}(\mathcal{M}'_\mathbf{n}))$ and $N_n \leq \sqrt{n}/\log^2 n$. Moreover, $\mathbb{E}[\epsilon_1^4] < +\infty$.

Assumption $(\mathbf{H_1})$ is more restrictive than $(\mathbf{H_2})$ about the nature of the models since the property of localisation is stronger than norm connexion. Nevertheless, it allows models of larger dimension. Besides, if f is compactly supported, the conditions about the moments of ϵ_1 obviously hold.

The estimator $\widehat{b}_{\widehat{m}}$ satisfies the following result, which arises from Baraud (2000) and Baraud (2002).

Theorem 2.3.2 *Assume that* $(\mathbf{H_{gen}})$, $(\mathbf{H_{glob}}(\mathcal{M}'_\mathbf{n}))$ *and* $(\mathbf{H_1})$ *or* $(\mathbf{H_2})$ *are satisfied, then there exist a numerical constant* C_3 *and a constant* C_4 *depending on* $(\sigma^2, m_0, \|b\|_{f_X}, K)$ *and* $\mathbb{E}[\epsilon_1^4]$ *or* $\mathbb{E}[\epsilon_1^6]$ *such that*

$$\mathbb{E}\left[\|\tilde{b} - b\|_{f_X}^2\right] \leq C_3 \inf_{m \in J'_n} \left\{ \|b - b_m\|_{f_X}^2 + \sigma^2 \frac{D_m}{n} \right\} + \frac{C_4}{n}.$$

Comments

1. The estimator $\tilde{b}$ converges as the same rate as the best estimator in the collection (see Comment 1. after Theorem 2.3.1).

2. According to Comment 2. after Theorem 2.3.1, if $b \in \mathcal{B}^{\beta}_{p,q}(L)$, the model m^* which realises the bias-variance trade-off satisfies $D_{m^*} = Cn^{1/(2\beta+1)}$ and for this model,

$$\left\{ \|b - b_{m^*}\|^2_{f_X} + \sigma^2 \frac{D_{m^*}}{n} \right\} \leq C' n^{-2\beta/(2\beta+1)}$$

which is the minimax rate. Thus, the estimator $\tilde{b}$ reaches the minimax rate of convergence on $\mathcal{B}^{\beta}_{p,q}(L)$ iff $m^* \in J_n$, that is

$$n^{1/(2\beta+1)} \leq N_n.$$

Hence, under the assumptions $(\mathbf{H_2})$, $\tilde{b}$ is minimax over $\mathcal{B}^{\beta}_{p,q}(L)$ iff $\beta > 1/2$.

2.4 Estimator of error density and main result

This section presents the estimator of error density built from the sample (2.1).

a) Construction of the residuals

As the $\{\epsilon_i\}$'s are unobserved, we compute quantities, called residuals, which estimate them. Let $\tilde{b}$ be the estimator of b built from the sample Z^- in Section 2.3.2. For every $i \in \{1, \ldots, n\}$, let

$$\widehat{\epsilon}_i = Y_i - \tilde{b}(X_i) = \epsilon_i + (b - \tilde{b})(X_i).$$

Let the sample Z^- be fixed, then $\tilde{b}$ is fixed, so the $\{\widehat{\epsilon}_i\}_{i=1,\ldots,n}$ are i.i.d and let f^- be their common density. For every $t \in L^1(\mathbb{R})$,

$$\begin{aligned}
\mathbb{E}[t(\widehat{\epsilon}_1)] &= \mathbb{E}[t(\epsilon_1 + (b - \tilde{b})(X_1))] \\
&= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} t(u + (b - \tilde{b})(x)) f(u) du \right] f_X(x) dx \\
&= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} t(y) f(y - (b - \tilde{b})(x)) dy \right] f_X(x) dx \\
&= \int_{\mathbb{R}} t(y) \left[\int_{\mathbb{R}} f(y - (b - \tilde{b})(x)) f_X(x) dx \right] dy.
\end{aligned}$$

Hence, for every $y \in \mathbb{R}$,

$$f^-(y) = \int_{\mathbb{R}} f(y - (b - \tilde{b})(x)) f_X(x) dx. \tag{2.9}$$

b) Collection of estimators and model selection procedure

The non adaptive estimators are computed exactly like in Section 2.3.1, except that the $\{V_i\}$ are replaced by the residuals $\{\widehat{\epsilon}_i\}$. Let

$$\mathcal{M}_n = \{S_m, m \in J_n\}$$

be a collection of linear models and $\{\phi_k^m, k \in I_m\}$ be a $\|.\|$-orthonormal basis of S_m, for every $m \in J_n$.

For every $t \in L^2(\mathbb{R})$, let

$$\gamma_n^-(t) = \|t\|^2 - \frac{2}{n}\sum_{i=1}^{n} t(\widehat{\epsilon}_i)$$

and for every $m \in S_m$,

$$\widehat{f}_m^- = \arg\min_{t \in S_m} \gamma_n^-(t) = \sum_{k \in I_m} \widehat{a}_k^- \phi_k^m.$$

Then, let

$$\widehat{m} = \arg\min_{m \in J_n} \left\{ \gamma_n^-(\widehat{f}_m^-) + pen(m) \right\}$$

where $pen(m) = \theta K^2 D_m/n$ for some constant $\theta > 1$. Our estimator of f is $\widehat{f}_{\widehat{m}}^-$.

c) Main result

Theorem 2.4.1 *Assume that Assumptions* $(\mathbf{H_{gen}})$, $(\mathbf{H_{glob}}(\mathcal{M}_\mathbf{n}))$, $(\mathbf{H_{con}}(\mathcal{M}_\mathbf{n}))$, $(\mathbf{H_{err(1)}})$ *or* $(\mathbf{H_{err(2)}})$, *and* $(\mathbf{H_1})$ *or* $(\mathbf{H_2})$ *hold. There exist a numerical constant* C_5 *and a constant* C_6 *which depends on* $(\sigma^2, m_0, \|b\|_{f_X}, K)$ *and* $\mathbb{E}[\epsilon_1^4]$ *or* $\mathbb{E}[\epsilon_1^6]$ *such that*

$$\mathbb{E}\left[\|\widehat{f}_{\widehat{m}}^- - f\|^2\right] \leq C_5 \left(\inf_{m \in J_n} \left\{ \|f - f_m\|^2 + K^2 \frac{D_m}{n} \right\} + \inf_{m \in J_n'} \left\{ \|b - b_m\|_{f_X}^2 + \sigma^2 \frac{D_m}{n} \right\} \right) + \frac{C_6}{n}.$$

Comments

1. The term $\inf_{m \in J_n} \left\{ \|f - f_m\|^2 + K^2 \frac{D_m}{n} \right\}$ is the adaptive rate of convergence we would get if the $\{\epsilon_i\}$'s were observed. Thus, the rate of convergence of our estimator is upper bounded by the maximum of two rates:

   $\star$ the minimax rate of convergence of the regression function b.

   $\star$ the minimax rate of the density f if the $\{\epsilon_i\}$'s were observed.

   In particular, if b is more regular than f, our estimator converge as fast as if the $\{\epsilon_i\}$'s were observed, up to a multiplicative constant.

2. In the estimation procedure, we have considered a particular estimator $\tilde{b}$ of b which is minimax over classical regularity classes, but a more general result holds. If $\tilde{b}$ is replaced by any estimator $\widehat{b}$ of b built from the sample Z^-,

$$\mathbb{E}\left[\|\widehat{f}^-_{\widehat{m}} - f\|^2\right] \leq C_5 \left(\inf_{m \in J_n} \left\{ \|f - f_m\|^2 + K^2 \frac{D_m}{n} \right\} + \mathbb{E}\left[\|\widehat{b} - b\|^2_{f_X} \right] \right) + \frac{C_6}{n}.$$

d) Comparison with the estimator of Efromovich (2005)

In this section, we point out the outline of the error density estimation procedure developed by Efromovich (2005). As we do in this chapter, Efromovich splits the sample in two independent sequences. With the first one he estimates the regression function then he applies a density estimation procedure to the residuals of the second sequence and obtains an estimator $\widehat{f}^-$.

The density estimation procedure used by Efromovich is based on projection in a trigonometric basis with dimension $D = n^{1/3}$ and which only includes cosine coefficients, followed by a coefficients shrinkage which produces an adaptive estimator. Contrary to our estimator $\widehat{b}_m$ which estimates the projection of b on S_m for the norm $\|.\|_{f_X}$, Efromovich estimates the Fourier coefficients of b, namely the projection of b in a trigonometric model for the norm $\|.\|$. by plugging in an estimator of f_X. This enables to upper bound the bias term with Fourier analysis argument.

His error density estimator is theoretically very powerful since it reaches the rate of convergence that we would obtain if the $\{\epsilon_i\}$'s were observed. Nevertheless, this requires that f_X and b are differentiable and f is two times differentiable, whereas, for example, we make no assumption about the regularity of the designs' density f_X. Besides, our estimator is more computable and the results that we present are very simple to prove from existing result about density and regression function estimation.

We present here an heuristic of the upper bound of the risk of Efromovich's estimator, to underline the differences with the proof of our result. Efromovich splits the risk of $\widehat{f}^-$ in a different way than in Theorem 2.4.1. Let $\widehat{f}$ be the pseudo-estimator that would come from the density estimation procedure applied to the $(\epsilon_i)_{i=1,\dots,n}$ if they were observed.

$$\mathbb{E}[\|\widehat{f}^- - f\|^2] \leq 2\{\mathbb{E}[\|\widehat{f}^- - \widehat{f}\|^2] + \mathbb{E}[\|\widehat{f} - f\|^2]\}.$$

On the one hand, $\mathbb{E}[\|\widehat{f} - f\|^2]$ has the required order. On the other hand, Efromovich proves that $\mathbb{E}[\|\widehat{f}^- - \widehat{f}\|$ has order $1/n$. For sake of simplicity, we omit the shrinkage in the following development. For every $\lambda = 1, \dots, D$,

$$\begin{aligned} \widehat{a}^-_\lambda &= \tfrac{1}{n} \textstyle\sum_{i=1}^n \cos(\pi\lambda\widehat{\epsilon}_i) \\ \widehat{a}_\lambda &= \tfrac{1}{n} \textstyle\sum_{i=1}^n \cos(\pi\lambda\epsilon_i) \end{aligned}$$

be the coefficients of $\widehat{f^-}$ and $\widehat{f}$, then $\|\widehat{f^-} - \widehat{f}\|^2 = \sum_\lambda (\widehat{a}_\lambda^- - \widehat{a}_\lambda)^2$ and

$$\widehat{a}_\lambda^- - \widehat{a}_\lambda = \frac{1}{n}\sum_{i=1}^{n} \cos(\pi\lambda(\epsilon_i + (b - \widehat{b})(X_i)) - \cos(\pi\lambda\epsilon_i).$$

A Taylor expansion of cosine at order 2K for some integer K in the above expression provides the following upper bound:

$$\mathbb{E}[\|\widehat{f^-} - \widehat{f}\|^2] \leq C \sum_{\lambda=1}^{D}\sum_{k=1}^{K} \left\{ \left(\frac{1}{n}\sum_{i=1}^{n} \sin(\pi\lambda\epsilon_i)\lambda^{2k-1}(b-\widehat{b})^{2k-1}(X_i) \right)^2 \right.$$
$$\left. + \left(\frac{1}{n}\sum_{i=1}^{n} \cos(\pi\lambda\epsilon_i)\lambda^{2k}(b-\widehat{b})^{2k}(X_i) \right)^2 \right\} + \left(\frac{1}{n}\sum_{i=1}^{n} \lambda^{2K+1}|b-\widehat{b}|^{2k-1}(X_i) \right)^2$$

and these terms are upper bounded by $\log n/n$, by means of Fourier analysis result through very complicated calculus.

2.5 Numerical results

This section illustrates the error density estimation procedure on simulated data. We use programs from Yves Rozenholc (available on www.math-info.univ-paris5.fr/∼rozen/) which compute density and regression function estimators following model selection procedures very close to the ones developed in Sections 2.3.1 and 2.3.2. The collections of models considered mix trigonometric and piecewise polynomial functions. Moreover, following a remark in the beginning of Section 2.3.1, the compact support is actually chosen from the data : the density estimation program computes an estimator from a sample $(V_1, \ldots, V_n)$ on the interval $[\min(V_i), \max(V_i)]$.

Density of ϵ

The basis used in these programs are adapted to the estimation of a compactly supported density, thus the densities f of the $\{\epsilon_i\}$'s we consider are supported on a compact. Nevertheless for every variable, with an appropriate compact the most part of the density is compactly-supported. So we can choose any density f provided that $\mathbb{E}[\epsilon_i] = 0$. We consider the following densities for ϵ_i:

- ϵ_i is gaussian with mean 0 and variance 1 (denoted by $\epsilon \sim \mathcal{N}(0,1)$). The density of ϵ_i is symmetric.
- ϵ_i is a centered χ^2-variable of parameter 3 (denoted by $\epsilon \sim \chi^2(3)^*$). The density of ϵ_i is not symmetric.

Density of X

The density f_X of the $\{X_i\}$'s is supposed to be supported on a compact, but contrary to f, it has to be lower bounded on this compact. Thus the examples given above are inapropriate for f_X. We consider densities built from the uniform distribution: let U be a variable uniform on $[0,1]$, and let $G : [a,b] \to [0,1]$ be an increasing bijective function. Then, $G^{-1}(U)$ has density $g = G'$ on $[a,b]$. So for a given function G, we simulate a sample $(U_i)_{i\in\{-n,\dots,-1\}\cup\{1,\dots,n\}}$ of i.i.d. uniform variables on $[0,1]$ and compute the i.i.d. sample $(X_i = G^{-1}(U_i))_{i\in\{-n,\dots,-1\}\cup\{1,\dots,n\}}$ with density $g = G'$. More precisely, we consider the following densities for X_i:

- X_i is uniform on $[0,1]$ (denoted by $X \sim \mathcal{U}([0,1])$).
- $X_i = G_1^{-1}(U_i)$ where U_i is uniform on $[0,1]$ and

$$\begin{array}{rccc} G_1: & [0,\pi/6] & \to & [0,1] \\ & x & \to & 2\sin x \end{array}$$

  i.e. X_i has density $g_1(x) = 2\cos x$ on $[0,\pi/6]$ (denoted by $X \sim \mathcal{L}_1$).
- $X_i = G_2^{-1}(U_i)$ where U_i is uniform on $[0,1]$ and

$$\begin{array}{rccc} G_2: & [1/4,1] & \to & [0,1] \\ & x & \to & 2\sqrt{x}-1 \end{array}$$

  i.e. X_i has density $g_2(x) = 1/\sqrt{x}$ on $[1/4,1]$. (denoted by $X \sim \mathcal{L}_2$)

Examples

We consider the following examples:

Example 1: $X \sim \mathcal{L}_1$, $\epsilon \sim \mathcal{N}(0,1)$, $b(x) = x^3 - 5x$
Example 2: $X \sim \mathcal{L}_1$, $\epsilon \sim \chi^2(3)^*$, $b(x) = x^3 - 5x$
Example 3: $X \sim \mathcal{L}_1$, $\epsilon \sim \mathcal{N}(0,1$, $b(x) = x^3 - 5x$
Example 4: $X \sim \mathcal{L}_2$, $\epsilon \sim \mathcal{N}(0,1)$, $b(x) = x^3 - 5x$
Example 5: $X \sim \mathcal{L}_2$, $\epsilon \sim \mathcal{N}(0,1)$, $b(x) = \exp(-10x)$
Example 6: $X \sim \mathcal{U}([0,1])$, $\epsilon \sim \chi^2(3)^*$, $b(x) = \exp(-10x)$

For each example, we compute the estimator $\widehat{f}^-_{\widehat{m}}$ of f as follows.

(i) We simulate a sample $(X_i, \epsilon_i)_{i\in\{-n,\dots,-1\}\cup\{1,\dots,n\}}$.

(ii) We compute the sample $(Y_i = b(X_i))_{i\in\{-n,\dots,-1\}\cup\{1,\dots,n\}}$.

(iii) We apply the regression function estimation program to the sample $(X_i, Y_i)_{i\in\{-n,\dots,-1\}}$ to compute the estimator $\tilde{b}$ of b.

(iv) For $i = 1, \dots, n$, we compute the residuals $(\widehat{\epsilon}_i)_{i=1,\dots,n}$

(v) We apply the density estimation program to the $(\widehat{\epsilon}_i)_{i=1,\dots,n}$ and obtain the estimator $\widehat{f}^{-}_{\widehat{m}}$.

For examples 3, 4, 5 and 6, we also compute the theoretical estimator we would obtain if the $(\epsilon_i)_{i=1,\dots,n}$ were observed.

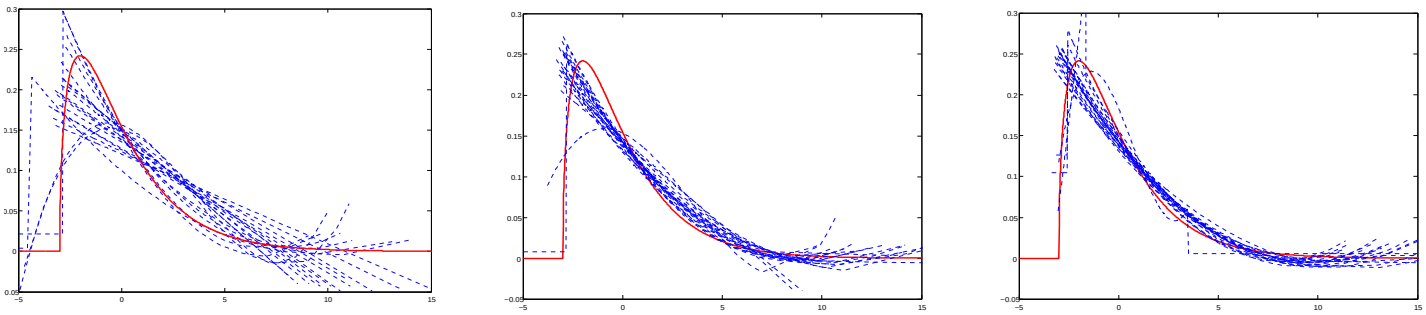
(vi) We apply the density estimation program to the sample $(\epsilon_i)_{i=1,\dots,n}$ estimated in (i). We obtain a theoretical estimator f^*.

Figures 2.1 and 2.2 present a beam of 20 error density estimators (dotted lines) and the true error density for examples 1 and 2, with $n = 200$, 500 and 1000. Figures 2.3, 2.4, 2.5 and 2.6 present

- the true density (solid line)
- the estimator $\widehat{f}^{-}_{\widehat{m}}$ (large dotted line)
- the theoretical estimator f^* (thin dotted line)

respectively for examples 3, 4, 5 and 6, for sample of size $2n$ with $n = 200$ and 1000.
Comments

1. For all the examples presented here, we see that, obviously, the estimation gets better as n increases.
2. According to Comment 2 after Theorem 2.3.1, the rate of convergence depends on the regularity: the more regular is the function, the faster is the convergence. We observe this phenomenon on numerical results: for example, the estimation is much better in Figure 2.1 where f is a gaussian density which is very regular, than in Figure 2.2 where f is a $\chi^2(3)^*$ which is much less regular, especially in 0.
3. On Figures 2.3, 2.4, 2.5 and 2.6, we note that the error density estimator $\widehat{f}^{-}_{\widehat{m}}$ estimates f nearly as well as the theoretical estimator f^* that we would obtain if the errors were observed. Thus, in the two errors terms appearing in Theorem 2.4.1, the error from the computing of the residuals seems to be negligible compared to the density estimation error.



$n = 200$ $n = 500$ $n = 1000$

Figure 2.1: $X \sim \mathcal{L}_1$, $\epsilon \sim \mathcal{N}(0,1)$, $b(x) = x^3 - 5x$

4. In examples 3 and 4, represented in Figures 2.3 and 2.4, b and f are the same but f_X is different. The estimation is not really better in an example than in the other, which illustrates the fact that the regularity of the designs does not have an influence on the risk of f.

2.6 Proofs

2.6.1 Proof of Theorem 2.3.1

We only give a sketch of the proof which is detailed in Massart (2007). Besides, for sake of simplicity, we only prove the result for a constant $\theta > 4$ in the penalty, but provided slight changes in the splitting of terms in the proof, the result holds for every $\theta > 1$.

Let $m \in J_n$, and $g_m \in S_m$. By definition of $\widehat{g}_m$ and $\widehat{m}$,

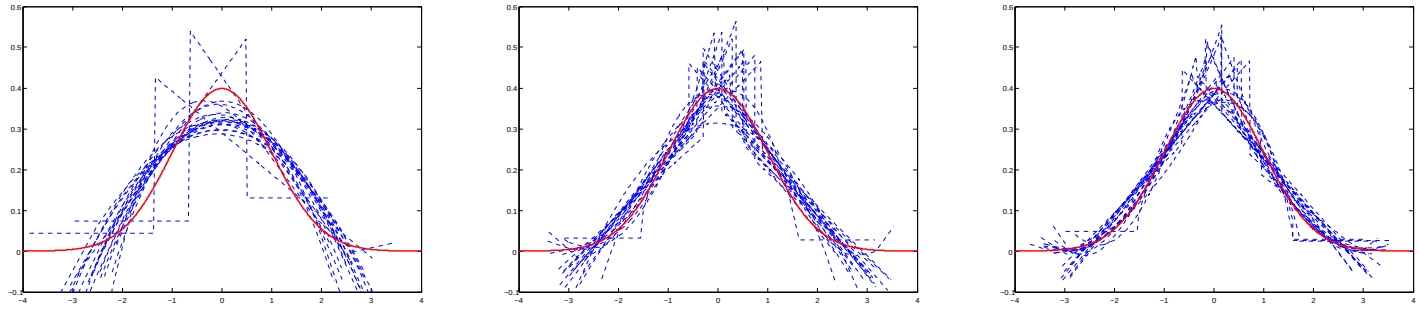
$$\gamma_n(\widehat{g}_{\widehat{m}}) + pen(\widehat{m}) \leq \gamma_n(\widehat{g}_m) + pen(m) \leq \gamma_n(g_m) + pen(m).$$

We replace γ_n by its expression and use the development $\|t-s\|^2 = \|t\|^2 + \|s\|^2 - 2\langle s,t\rangle$.

$$\|\widehat{g}_{\widehat{m}} - g\|^2 \leq \|g_m - g\|^2 + pen(m) - pen(\widehat{m}) + 2\nu_n(\widehat{g}_{\widehat{m}} - g_m)$$

where

$$\nu_n(t) = \frac{1}{n}\sum_{i=1}^{n}(t(V_i) - < t, g >) \quad \forall t \in L^2(\mathbb{R}).$$



$n = 200$ $n = 500$ $n = 1000$

Figure 2.2: $X \sim \mathcal{L}_1$, $\epsilon \sim \chi^2(3)^*$, $b(x) = x^3 - 5x$

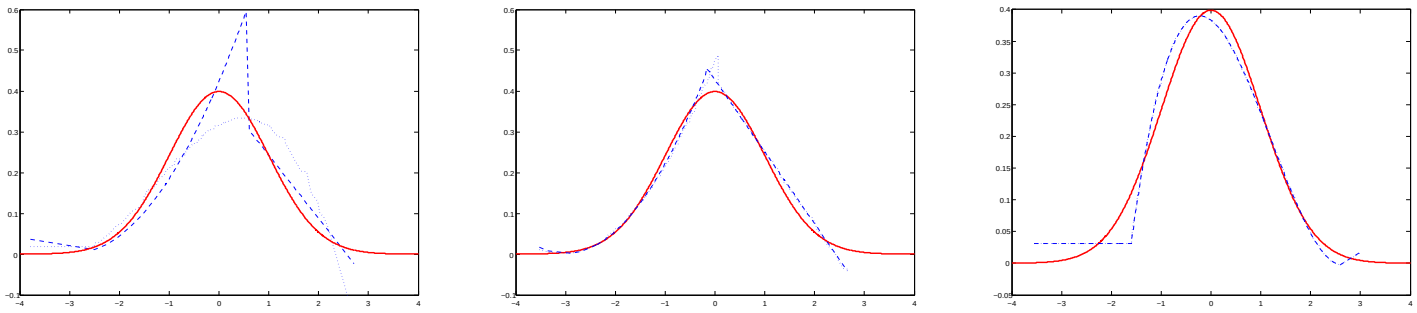
Thus,

$$\begin{aligned}
\|\widehat{g}_{\widehat{m}} - g\|^2 &\leq \|g_m - g\|^2 + pen(m) - pen(\widehat{m}) + 2\|\widehat{g}_{\widehat{m}} - g_m\| \sup_{t \in S_m + S_{\widehat{m}}, \|t\| \leq 1} |\nu_n(t)| \\
&\leq \|g_m - g\|^2 + pen(m) - pen(\widehat{m}) + \frac{1}{4}\|\widehat{g}_{\widehat{m}} - g_m\|^2 + 4 \sup_{t \in S_m + S_{\widehat{m}}, \|t\| \leq 1} (\nu_n(t))^2 \\
&\leq \|g_m - g\|^2 + pen(m) - pen(\widehat{m}) + \frac{1}{2}\|\widehat{g}_{\widehat{m}} - g\|^2 + \frac{1}{2}\|g_m - g\|^2 \\
&\quad + 4 \sup_{t \in S_m + S_{\widehat{m}}, \|t\| \leq 1} (\nu_n(t))^2.
\end{aligned}$$

Then, let $p(m, m') = \frac{1}{4}\left(pen(m) + pen(m')\right)$, and consider the expectation.

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}\left[\|\widehat{g}_{\widehat{m}} - g\|^2\right] \leq \frac{3}{2}\|g_m - g\|^2 + 2pen(m) + 4\mathbb{E}\left[\sup_{t \in S_m + S_{\widehat{m}}, \|t\| \leq 1} \left((\nu_n(t))^2 - p(m, \widehat{m})\right)\right] \\
&\leq \frac{3}{2}\|g_m - g\|^2 + 2pen(m) + 4\sum_{m' \in J_n} \mathbb{E}\left[\sup_{t \in S_m + S_{m'}, \|t\| \leq 1} \left((\nu_n(t))^2 - p(m, m')\right)_+\right]. \qquad (2.10)
\end{aligned}$$

Now $\sum_{m' \in J_n} \mathbb{E}\left[\sup_{t \in S_m + S_{m'}, \|t\| \leq 1} \left((\nu_n(t))^2 - p(m, m')\right)_+\right]$ is upper bounded with Talagrand Inequality (see Theorem 1.2.3). The penalty function has been chosen so that $4p(m, m') = pen(m) + pen(m')$ has the order of the term $\mathbb{H}^2$. More precisely, with Cauchy-Schwartz Inequality we prove that



$n = 200$ $n = 500$ $n = 1000$

Figure 2.3: $X \sim \mathcal{L}_1$, $\epsilon \sim \mathcal{N}(0, 1$, $b(x) = x^3 - 5x$

$$\mathbb{E}\left[\sup_{t \in S_m + S_{m'}, \|t\| \leq 1} (\nu_n(t))^2\right]_+ \leq \frac{K^2(D_m + D_{m'})}{n} = \mathbb{H}^2$$

$$\sup_{t \in S_m + S_{m'}, \|t\| \leq 1} Var\left(t(V_1)\right) \leq K\sqrt{D_m + D_{m'}}\|g\| = v$$

$$\sup_{t \in S_m + S_{m'}, \|t\| \leq 1} \|t(.)\|_\infty \leq K\sqrt{D_m + D_{m'}} = b$$

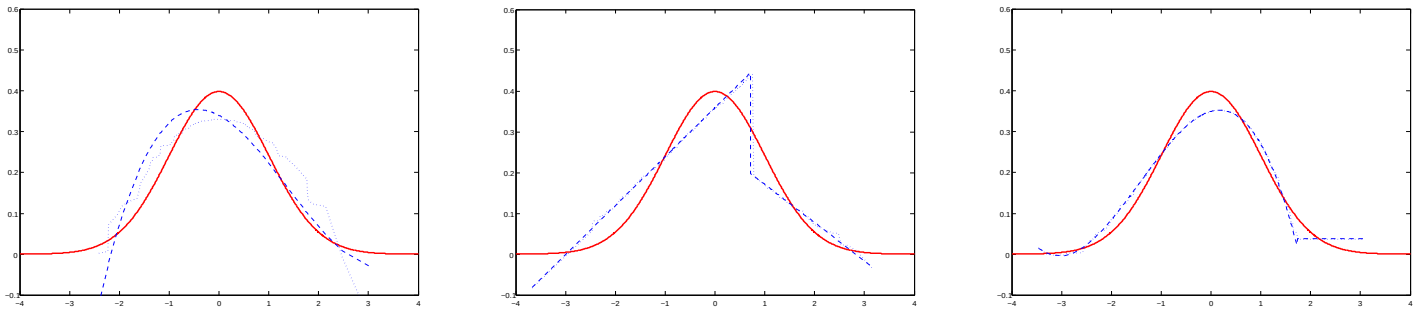
and Talagrand Inequality (Theorem 1.2.3) implies

$$\begin{aligned}
\sum_{m' \in J_n} \mathbb{E}\left[\sup_{t \in S_m + S_{m'}, \|t\| \leq 1} \left((\nu_n(t))^2 - p(m, m')\right)_+\right] &\leq \frac{C}{n} + \frac{C'}{n} \sum_{m' \in J_n} \sqrt{D_m + D_{m'}} \exp\left(-\frac{K\sqrt{D_m + D_{m'}}}{\|g\|}\right) \\
&\leq \frac{C}{n} + \frac{C'}{n} \sum_{D \in \mathbb{N}^*} \Gamma D^R \sqrt{D_m + D} \exp\left(-\frac{K\sqrt{D_m + D}}{\|g\|}\right) \\
&\leq \frac{C}{n} + \frac{C'}{n} \sum_{D \in \mathbb{N}^*} \Gamma D^{R+1/2} \exp\left(-\frac{K\sqrt{D}}{\|g\|}\right).
\end{aligned}$$

Then (2.10) concludes the proof of Theorem 2.3.1. □

2.6.2 Proof of Theorem 2.3.2

We only give a sketch of the proof which is a combination of results from Baraud (2002) which proves this result when σ^2 is known, and Baraud (2000) in which he introduces an estimator



$n = 200$ $n = 500$ $n = 1000$

Figure 2.4: $X \sim \mathcal{L}_2$, $\epsilon \sim \mathcal{N}(0,1)$, $b(x) = x^3 - 5x$

of σ but in a fixed-design context (i.e. with non random $\{X_i\}$). Thus we need the following lemma, which is proved at the end of this section.

Lemma 2.6.1

$$(1) \quad \mathbb{E}[\widehat{\sigma}_n^2] \leq 2\left(\sigma^2 + \inf_{t \in W_n} \|b - t\|_{f_X}^2\right).$$

$$(2) \quad P[\widehat{\sigma}_n^2 \leq \sigma^2/2] \leq \frac{C}{n}.$$

The term $\mathbb{E}\left[\|\tilde{b} - b\|_{f_X}^2\right]$ splits in three terms. Let

$$A = \left\{ \|t\|_{f_X}^2 - \|t\|_n^2 \leq \frac{1}{2}\|t\|_{f_X}^2, \ \forall t \in S_n \right\}$$

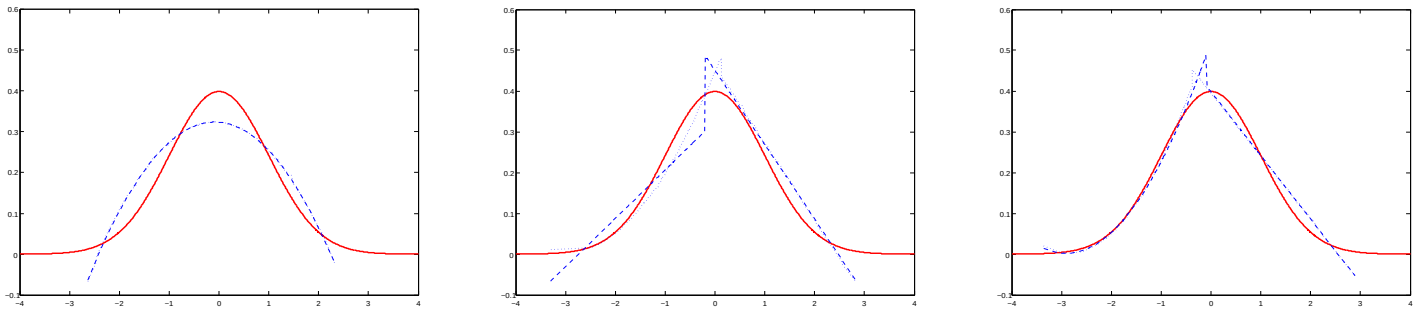
and

$$T_1 = \mathbb{E}\left[\|\tilde{b} - b\|_{f_X}^2 \mathbb{1}_{A \cap \{\|\widehat{b}_{\widehat{m}}\| \leq n\}}\right]$$

$$T_2 = \mathbb{E}\left[\|\tilde{b} - b\|_{f_X}^2 \mathbb{1}_{A \cap \{\|\widehat{b}_{\widehat{m}}\| > n\}}\right]$$

$$T_3 = \mathbb{E}\left[\|\tilde{b} - b\|_{f_X}^2 \mathbb{1}_{A^c}\right].$$

Claim 2.1 *There exist an absolute constant C and a constant C' which depends on the parameters of the problem such that*



$n = 200$ $n = 500$ $n = 1000$

Figure 2.5: $X \sim \mathcal{L}_2$, $\epsilon \sim \mathcal{N}(0,1)$, $b(x) = \exp(-10x)$

$$T_1 \leq C \inf_{m \in J_n'} \left\{ \|b - b_m\|_{f_X}^2 + \sigma^2 \frac{D_m}{n} \right\} + \frac{C'}{n}.$$

Let us prove Claim 2.1. Let $m \in J_n'$ and $b_m \in S_m$. Similarly to the proof of Theorem 2.3.1,

$$\begin{aligned} \|\widehat{b}_{\widehat{m}} - b\|_n^2 &\leq \|b_m - b\|_n^2 + pen(m) - pen(\widehat{m}) + 2\langle \epsilon, \widehat{b}_{\widehat{m}} - b_m \rangle_n \\ &\leq \|b_m - b\|_n^2 + pen(m) - pen(\widehat{m}) + \frac{1}{8}\|\widehat{b}_{\widehat{m}} - b_m\|_{f_X}^2 + 8 \sup_{t \in S_m + S_{\widehat{m}}, \|t\|_{f_X} \leq 1} \langle \epsilon, t \rangle_n^2. \end{aligned}$$

Let $\{k_n\}$ be a sequence of positive numbers which will be precised later. We define

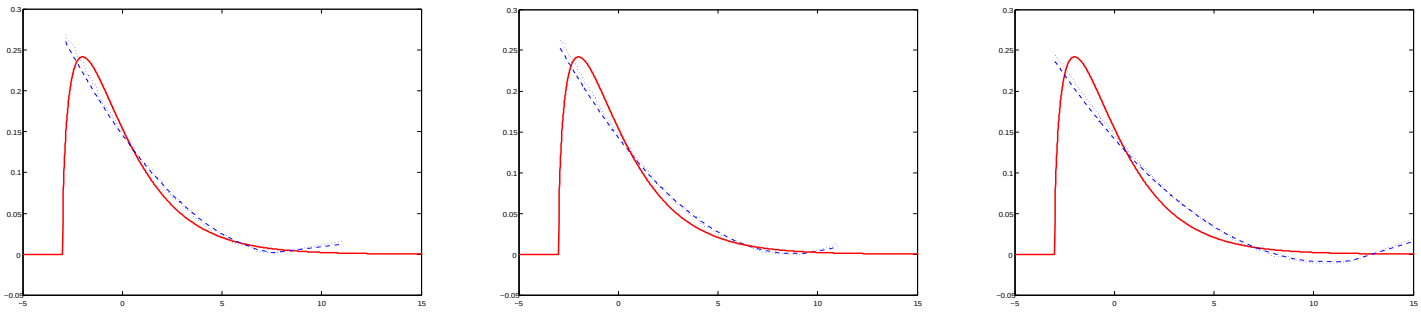
$$Z_1(m, m') = \frac{1}{n} \sup_{t \in S_m + S_{m'}, \|t\|_{f_X} \leq 1} \sum_{i=-n}^{-1} t(X_i)(\epsilon_i 1_{|\epsilon_i| \leq k_n} - \mathbb{E}(\epsilon_i 1_{|\epsilon_i| \leq k_n})) = \frac{1}{n} \sup_{t \in S_m + S_{m'}, \|t\|_{f_X} \leq 1} \sum_{i=-n}^{-1} v_i,$$

$$Z_2(m; m') = \frac{1}{n} \sup_{t \in S_m + S_{m'}, \|t\|_{f_X} \leq 1} \sum_{i=-n}^{-1} t(X_i)(\epsilon_i 1_{|\epsilon_i| > k_n} - \mathbb{E}(\epsilon_i 1_{|\epsilon_i| > k_n})) = \frac{1}{n} \sup_{t \in S_m + S_{m'}, \|t\|_{f_X} \leq 1} \sum_{i=-n}^{-1} u_i.$$

Then

$$\sup_{t \in S_m + S_{\widehat{m}}, \|t\|_{f_X} \leq 1} \langle \epsilon, t \rangle_n \leq Z_1(m, \widehat{m}) + Z_2(\widehat{m}, m).$$

Let



$n = 200$ $n = 500$ $n = 1000$

Figure 2.6: $X \sim \mathcal{U}([0,1])$, $\epsilon \sim \chi^2(3)^*$, $b(x) = \exp(-10x)$

$$p(m,m') = \frac{1}{16}\theta'\sigma^2\frac{D_m + D_{m'}}{n},$$

then

$$\begin{aligned}\|\widehat{b}_{\widehat{m}} - b\|_n^2 \;\leq\; & \|b_m - b\|_n^2 + \frac{1}{4}\|b_m - b\|_{f_X}^2 + (pen(m) - pen(\widehat{m}) + 8p(m,\widehat{m})) \\ & + \frac{1}{4}\|\widehat{b}_{\widehat{m}} - b\|_{f_X}^2 + 16\mathbb{E}\left[\left(Z_1(m,\widehat{m})^2 - p(m,\widehat{m})\right)_+\right] + 16\mathbb{E}\left[(Z_2(m,\widehat{m}))^2\right].\end{aligned}$$

Moreover, as $\mathbb{E}\left[\|b_m - b\|_n^2\right] = \|b_m - b\|_{f_X}^2$,

$$\begin{aligned}\mathbb{E}\left[\|\widehat{b}_{\widehat{m}} - b\|_n^2 \mathbb{1}_A\right] \;\leq\; & \frac{5}{4}\|b_m - b\|_{f_X}^2 + \mathbb{E}\left[pen(m) - pen(\widehat{m}) + 8p(m,\widehat{m})\right] + \frac{1}{4}\mathbb{E}\left[\|\widehat{b}_{\widehat{m}} - b\|_{f_X}^2 \mathbb{1}_A\right] \\ & + 16\mathbb{E}\left[\left(Z_1(m,\widehat{m})^2 - p(m,\widehat{m})\right)_+\right] + 16\mathbb{E}\left[(Z_2(m,\widehat{m}))^2\right]. \qquad (2.11)\end{aligned}$$

Besides, by Assumption $(\mathbf{H_{glob}}(\mathcal{M}'_{\mathbf{n}})$ there exists a model S_n which contains every model in the collection. In order to upper bound T_1 by $\mathbb{E}\left[\|\widehat{b}_{\widehat{m}} - b\|_{f_X}^2 \mathbb{1}_A\right]$, the empirical norm $\|.\|_n$ has to be replaced by the norm $\|.\|_{f_X}$ in the left side of (2.11), but the two norms are equivalent

only on S_n. Nevertheless,

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{b}_{\widehat{m}}-b\|_{f_X}^2 \mathbb{1}_A\right] &\leq \frac{5}{4}\mathbb{E}\left[\|\widehat{b}_{\widehat{m}}-b_m\|_{f_X}^2 \mathbb{1}_A\right] + 5\|b_m-b\|_{f_X}^2 \\
&\leq \frac{5}{2}\mathbb{E}\left[\|\widehat{b}_{\widehat{m}}-b_m\|_n^2 \mathbb{1}_A\right] + 5\|b_m-b\|_{f_X}^2 \\
&\leq 3\mathbb{E}\left[\|\widehat{b}_{\widehat{m}}-b\|_n^2 \mathbb{1}_A\right] + \frac{9}{2}\mathbb{E}\left[\|b_m-b\|_n^2\right] + 5\|b_m-b\|_{f_X}^2 \\
&= 3\mathbb{E}\left[\|\widehat{b}_{\widehat{m}}-b\|_n^2 \mathbb{1}_A\right] + \frac{19}{2}\|b_m-b\|_{f_X}^2.
\end{aligned}$$

This provides a lower bound for $\mathbb{E}\left[\|\widehat{b}_{\widehat{m}}-b\|_n^2 \mathbb{1}_A\right]$ that we plug in (2.11), which implies

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{b}_{\widehat{m}}-b\|_{f_X}^2 \mathbb{1}_A\right] \leq\ & C\Big\{\|b_m-b\|_{f_X}^2 + (pen(m)-pen(\widehat{m})+8p(m,\widehat{m})) \\
& + 16\sum_{m'\in J_n}\mathbb{E}\left[\left(Z_1(m,m')^2 - p(m,m')\right)_+\right] + 16\mathbb{E}\left[(Z_2(m,\widehat{m}))^2\right]\Big\} \quad (2.12)
\end{aligned}$$

for some numerical constant C.

Let us upper bound $\mathbb{E}\left[(Z_2(m,\widehat{m}))^2\right]$. Let $\{\psi_k^n, k=1,\ldots,N_n\}$ be a $\|.\|_{f_X}$-orthonormal basis of the global space S_n. The $\{u_i\}$'s are independent of the $\{X_i\}$'s so

$$\begin{aligned}
\mathbb{E}[Z_2^2(m,\widehat{m})] &= \frac{1}{n^2}\mathbb{E}\left[\Big(\sup_{t\in S_m+S_{\widehat{m}},\|t\|_{f_X}\leq 1}\sum_{i=-n}^{-1} t(X_i)u_i\Big)^2\right] \\
&\leq \frac{1}{n^2}\mathbb{E}\left[\Big(\sup_{|a|=1}\sum_{k=1}^{N_n} a_k\Big(\sum_{i=-n}^{-1}\psi_k^n(X_i)u_i\Big)\Big)^2\right] \\
&\leq \frac{1}{n^2}\mathbb{E}\left[\sup_{|a|=1}\sum_{k=1}^{N_n} a_k^2 \times \sum_{k=1}^{N_n}\Big(\sum_{i=-n}^{-1}\psi_k^n(X_i)u_i\Big)^2\right] \\
&= \frac{1}{n^2}\sum_{k=1}^{N_n}\mathbb{E}\left[\sum_{i,j=-n}^{-1}\psi_k^n(X_i)\psi_k^n(X_j)u_iu_j\right] \\
&= \frac{1}{n}\sum_{k=1}^{N_n}\mathbb{E}[(\psi_k^n)^2(X_1)u_1^2] \\
&= \frac{1}{n}\mathbb{E}[u_1^2]\sum_{k=1}^{N_n}\mathbb{E}[(\psi_k^n)^2(X_1)]
\end{aligned}$$

The $\{\psi_k^n\}$'s are $\|.\|_{f_X}$-orthonormal, hence

$$\begin{aligned}\mathbb{E}[Z_2^2(m,\widehat{m})] &= \frac{1}{n}Var(\epsilon_1 1_{|\epsilon_1|>k_n}) \times N_n \\ &\leq \frac{N_n}{n}\mathbb{E}[\epsilon_1^2 1_{|\epsilon_1|>k_n}].\end{aligned}$$

Thus, by Markov's Inequality, for every $u > 0$,

$$\mathbb{E}[Z_2^2(m,\widehat{m})] \leq \frac{N_n}{nk_n^u} \times \mathbb{E}[|\epsilon_1|^{2+u}]. \tag{2.13}$$

Now, let us upper bound $\mathbb{E}\left[(Z_1(m,m')^2 - p(m,m'))_+\right]$ for every $m, m' \in J_n$ with Talagrand Inequality (Theorem 1.2.3). Let $\{\psi_k^{m+m'}, k = 1, \ldots, D_{m+m'}\}$ a $\|.\|_{f_X}$-orthonormal basis of $S_m + S_{m'}$.

$$\begin{aligned}\mathbb{E}[Z_1^2(m,m')] &= \frac{1}{n^2}\mathbb{E}\left[\sup_{t\in S_m+S_{m'},\|t\|_{f_X}\leq 1}(\sum_{i=-n}^{-1} t(X_i)v_i)^2\right] \\ &\leq \frac{1}{n}\mathbb{E}\left[\epsilon_1^2 1_{|\epsilon_1|\leq k_n}\right]\sum_{k=1}^{D_{m+m'}}\mathbb{E}[(\psi_k^{m+m'})^2(X_1)] \\ &\leq \sigma^2\frac{D_{m+m'}}{n} \\ &\leq \sigma^2\frac{D_m + D_{m'}}{n} = \mathbb{H}^2.\end{aligned}$$

$$\begin{aligned}\sup_{t\in S_m+S_{m'},\|t\|_{f_X}\leq 1}\sup_{(x,y)\in\mathbb{R}^2}|t(x)y \times 1_{|y|\leq k_n}| &\leq k_n \sup_{t\in S_m+S_{m'},\|t\|_{f_X}\leq 1}\|t\|_\infty \\ &\leq \frac{K}{m_0}k_n\sqrt{D_m + D_{m'}} = b.\end{aligned}$$

The $\{v_i\}$'s and the $\{X_i\}$'s are independent, hence

$$\begin{aligned}\sup_{t\in S_m+S_{m'},\|t\|_{f_X}\leq 1} Var(t(X_1)v_1) &\leq \sup_{t\in S_m+S_{m'},\|t\|_{f_X}\leq 1}\mathbb{E}\left[t^2(X_1)v_1^2\right] \\ &= \sup_{t\in S_m+S_{m'},\|t\|_{f_X}\leq 1}\mathbb{E}\left[t^2(X_1)\right] \times \mathbb{E}\left[v_1^2\right] \\ &\leq \sigma^2 = v.\end{aligned}$$

By Talagrand Inequality (Theorem 1.2.3),

$$\mathbb{E}\left[\left(Z_1(m,m')^2 - p(m,m')\right)_+\right]$$
$$\leq \overline{C}\frac{\sigma^2}{n}\exp\left(-\frac{\overline{\kappa}}{\sigma^2}(D_m + D_{m'})\right) + \overline{C}'\frac{k_n^2 N_n}{n^2}\exp\left(-\frac{\overline{\kappa}' m_0}{K}\frac{\sqrt{n}}{k_n}\right)$$

Moreover, by assumption (2.3) in $\mathbf{H_{glob}}(\mathcal{M}'_{\mathbf{n}})$, for every positive constant a, there exists a constant A such that for every n,

$$\sum_{m' \in J_n} \exp\left(-a(D_m + D_{m'})\right) \leq A.$$

Thus, consider

$$k_n = n^{1/4},$$

then

$$\sum_{m' \in J_n} \mathbb{E}\left[\left(Z_1(m,m')^2 - p(m,m')\right)_+\right] \leq \frac{C}{n} + \overline{C}'\frac{N_n^2}{n^{3/2}}\exp\left(-\frac{\overline{\kappa}' m_0}{K} n^{1/4}\right) \leq \frac{C''}{n}.$$

Now, assume that $\mathbf{H_1}$ holds, then $N_n \leq n/\log^2 n \leq n$ and by (2.13) with $u = 4$,

$$\mathbb{E}[Z_2^2(m,\widehat{m})] \leq \frac{N_n}{n^2} \times \mathbb{E}[|\epsilon_1|^6] \leq \mathbb{E}[|\epsilon_1|^6] \times \frac{1}{n}.$$

Similarly, if $\mathbf{H_2}$ holds, $N_n \leq \sqrt{n}/\log^2 n \leq \sqrt{n}$ and by (2.13) with $u = 2$

$$\mathbb{E}[Z_2^2(m,\widehat{m})] \leq \frac{N_n}{n^{3/2}} \times \mathbb{E}[|\epsilon_1|^4] \leq \mathbb{E}[|\epsilon_1|^4] \times \frac{1}{n}.$$

Now, consider the term

$$\mathbb{E}\left[pen(m) - pen(\widehat{m}) + 8p(m,\widehat{m})\right] = \theta'\mathbb{E}\left[\widehat{\sigma}_n^2\frac{D_m - D_{\widehat{m}}}{n} + \frac{\sigma^2}{2}\frac{D_m + D_{\widehat{m}}}{n}\right].$$

Let

$$B = \left\{\widehat{\sigma}_n^2 \geq \frac{\sigma^2}{2}\right\}.$$

For every m, $m' \in J'_n$,

$$\left(pen(m) - pen(m') + 8p(m,m')\right)\mathbb{1}_B \leq 2\theta'\widehat{\sigma}_n^2\frac{D_m}{n}$$

and

$$(pen(m) - pen(m') + 8p(m,m'))\,\mathbb{1}_{B^c} \leq \theta' \widehat{\sigma}_n^2 \frac{D_m}{n} + \sigma^2.$$

Thus, with Lemma 2.6.1,

$$\mathbb{E}\left[(pen(m) - pen(m') + 8p(m,m'))\right] \leq C\left\{\sigma^2 \frac{D_m}{n} + \inf_{t \in W_n} \|b - t\|_{f_X}^2 + \frac{1}{n}\right\} \tag{2.14}$$

and (2.12), (2.13) and (2.14) conclude the proof of Theorem 2.3.2. □

Claim 2.2 *There exists a constant C which depends on $(\|b\|_{f_X}, \sigma^2, m_0)$ such that*

$$T_2 \leq \frac{C}{n}.$$

Indeed by the definition (2.8) of $\tilde{b}$;

$$\begin{aligned} T_2 &\leq 2\|b\|_{f_X}^2 P\left[A \cap \{\|\widehat{b}_{\widehat{m}}\|_{f_X}^2 > n^2 m_0\}\right] \\ &\leq 2\|b\|_{f_X}^2 P\left[\|\widehat{b}_{\widehat{m}}\|_n^2 > n^2 m_0/2\right] \\ &= 2\|b\|_{f_X}^2 P\left[\sum_{i=1}^n \widehat{b}_{\widehat{m}}^2(X_i) > n^3 m_0/2\right]. \end{aligned}$$

$\left(\widehat{b}_{\widehat{m}}(X_i)\right)_{i=1,\dots,n}$ is the projection of $(Y_i)_{i=1,\dots,n}$ on the linear subspace of $\mathbb{R}^n$ $\{(t(X_i))_{i=1,\dots,n}, t \in S_{\widehat{m}}\}$, hence $\sum_{i=1}^n \widehat{b}_{\widehat{m}}^2(X_i) \leq \sum_{i=1}^n Y_i^2$ and

$$T_2 \leq 4\|b\|_{f_X}^2 \frac{\mathbb{E}[Y_1^2]}{n^2 m_0} = 4\|b\|_{f_X}^2 \frac{\|b\|_{f_X}^2 + \sigma^2}{n^2 m_0} \leq \frac{C}{n}.$$

Claim 2.3 *There exists a constant C which depends on the parameters of the problem such that*

$$T_3 \leq \frac{C}{n}.$$

Indeed $T_3 \leq 2\left(\|b\|_{f_X}^2 + n\right) P[A^c]$ and

$$\begin{aligned} P[A^c] &\leq P\left[\sup_{t \in S_n, \|t\|_{f_X}^2 = 1} \left|\|t\|_n^2 - \|t\|_{f_X}^2\right| > \frac{1}{2}\right] \\ &= P\left[\sup_{\sum_{k \in I_n} a_k^2 = 1} \left|\sum_{k,l \in I_n} a_k a_l \left(\frac{1}{n}\sum_{i=1}^n \psi_k(X_i)\psi_l(X_i) - \mathbb{E}\left[\psi_k(X_i)\psi_l(X_i)\right]\right)\right| > \frac{1}{2}\right] \end{aligned}$$

where $(\psi_1, \dots, \psi_{N_n})$ is a $\|.\|_{f_X}$-orthonormal basis of the global space S_n. For every $(k,l) \in I_n^2$, let

$$\begin{cases} S_{k,l} = \frac{1}{n}\sum_{i=1}^n \psi_k(X_i)\psi_l(X_i) - \mathbb{E}\left[\psi_k(X_i)\psi_l(X_i)\right] \\ v_{k,l} = \mathbb{E}\left[\psi_k^2(X_i)\psi_l^2(X_i)\right] \\ c_{k,l} = \|\psi_k(X_i)\psi_l(X_i)\|_\infty, \end{cases}$$

and

$$\rho(V) = \sup_{\sum_{k\in I_n} a_k^2 = 1} \sum_{k,l\in I_n} |a_k||a_l|\sqrt{v_{k,l}}$$
$$\rho(C) = \sup_{\sum_{k\in I_n} a_k^2 = 1} \sum_{k,l\in I_n} |a_k||a_l| c_{k,l}.$$

Moreover, let

$$x = \frac{3-2\sqrt{2}}{2} \min\left(\frac{1}{\rho^2(V)}, \frac{1}{\rho(C)}\right), \tag{2.15}$$

then $\sqrt{2x}\rho(V) + x\rho(C) \le 1/2$ and

$$\begin{aligned} P[A^c] &\le P\left[\sup_{\sum_{k\in I_n} a_k^2 = 1} \sum_{k,l\in I_n} |a_k||a_l||S_{k,l}| > \sqrt{2x}\rho(V) + x\rho(C)\right] \\ &\le \sum_{k,l\in I_n} P[|S_{k,l}| > \sqrt{2xv_{k,l}} + xc_{k,l}]. \end{aligned}$$

With Berstein Inequality (Theorem 1.2.4), $P[A^c] \le 2\exp(-nx)$.

To bound x from below, we upper bound $\rho(V)$ and $\rho(C)$. Under Assumption $(\mathbf{H_1})$, there exists a $\|.\|_{f_X}$-orthonormal basis which satisfies $(\mathbf{H_{loc}})$ (see Section 1.2.5). As the basis $\{\psi_k\}$ can be chosen arbitrarly, we assume that it satisfies $(\mathbf{H_{loc}})$. Then we prove that

$$\rho^2(V) \le CN_n \quad \text{and} \quad \rho(C) \le CN_n$$

where C depends on K and m_0. Therefore,

$$P[A^c] \le 2\exp\left(-C'\frac{n}{N_n}\right) = 2\exp(-C'\log^2 n) = 2n^{-C'\log n}$$

and

$$T_3 \le C'' n^{1-C'\log n} \le \frac{C'''}{n}.$$

Under Assumption $(\mathbf{H_2})$, we prove that

$$\rho(V)^2 \leq CN_n^2 \quad \text{and} \quad \rho(C) \leq CN_n^2$$

and the conclusion is similar. Then Claims 2.1, 2.2 and 2.3 conclude the proof of Theorem 2.3.2. □

Proof of Lemma 2.6.1

By definition of $\widehat{\sigma}_n^2$, for every $t \in W_n$,

$$\begin{aligned}
\mathbb{E}[\widehat{\sigma}_n^2] &\leq \frac{2}{n}\mathbb{E}\left[\sum_{i=-n}^{-1}(Y_i - t(X_i))^2\right] \\
&= \frac{2}{n}\mathbb{E}\left[\sum_{i=-n}^{-1}((b-t)(X_i) + \epsilon_i)^2\right] \\
&= 2\left(\|b-t\|_{f_X}^2 + \sigma^2\right)
\end{aligned}$$

which proves (1) in Lemma 2.6.1.

Besides, let $\|.\|$ denote the canonical norm of $\mathbb{R}^n$, and $\Pi_{W_n(X)}$ the projection on

$$W_n(X) = \{(t(X_1), \ldots, t(X_n)), t \in W_n\},$$

then $\widehat{\sigma}_n^2 = (1/n)\|Y - \Pi_{W_n(X)}Y\|^2$. Let

$$w_n = \frac{b(X) - \Pi_{W_n(X)}b(X)}{\|b(X) - \Pi_{W_n(X)}b(X)\|} \quad \text{and} \quad \tilde{W}_n(X) = W_n(X) + Vect(w_n)$$

be a vector and a linear subset of $\mathbb{R}^n$, and $\Pi_{\tilde{W}_n(X)}$ and Π_{w_n} be the projection operators on $\tilde{W}_n(X)$ and $Vect(w_n)$. Then with Pythagoras formula

$$\begin{aligned}
\widehat{\sigma}_n^2 &= \frac{1}{n - \lfloor n/2 \rfloor}\|\epsilon - \Pi_{\tilde{W}_n(X)}\epsilon + \Pi_{w_n}\epsilon + b(X) - \Pi_{W_n(X)}b(X)\|^2 \\
&= \frac{1}{n - \lfloor n/2 \rfloor}\left(\|\epsilon - \Pi_{\tilde{W}_n(X)}\epsilon\|^2 + \|\Pi_{w_n}\epsilon + b(X) - \Pi_{W_n(X)}b(X)\|^2\right) \\
&\geq \frac{1}{n - \lfloor n/2 \rfloor}\|\epsilon - \Pi_{\tilde{W}_n(X)}\epsilon\|^2 \\
&= \frac{1}{n - \lfloor n/2 \rfloor}\left(\|\epsilon\|^2 - \|\Pi_{\tilde{W}_n(X)}\epsilon\|^2\right).
\end{aligned}$$

Hence

$$P\left[\widehat{\sigma}_n^2 \leq \frac{\sigma^2}{2}\right] \leq P\left[\frac{1}{n}\|\epsilon\|^2 \leq \frac{7}{8}\sigma^2\right] + P\left[\frac{1}{n}\|\Pi_{\tilde{W}_n(X)}\epsilon\|^2 \geq \frac{5}{8}\sigma^2\right].$$

On the one hand, by Markov's Inequality

$$P\left[\frac{1}{n}\|\epsilon\|^2 \leq \frac{7}{8}\sigma^2\right] \leq P\left[\left|\frac{1}{n}\sum_{i=1}^{n}(\epsilon_i^2 - \mathbb{E}[\epsilon_i^2])\right| \geq \frac{\sigma^2}{8}\right] \leq \frac{C}{n}.$$

On the other hand, the upper bound of $P\left[\frac{1}{n}\|\Pi_{\tilde{W}_n(X)}\epsilon\|^2 \geq \frac{5}{8}\sigma^2\right]$ relies on the following heuristic:

$$Dim(\tilde{W}_n(X)) \leq Dim(W_n(X)) + 1 = \lfloor \frac{n}{2} \rfloor \leq \frac{n}{2}$$

so

$$\mathbb{E}\left[\frac{1}{n}\|\Pi_{\tilde{W}_n(X)}\epsilon\|^2\right] = \sigma^2\frac{n}{2}.$$

Then by considering a basis of $\mathbb{R}^n$ adapted to the projection operator $\Pi_{\tilde{W}_n(X)}$, for $(X_1, \ldots, X_n)$ fixed,

$$\begin{aligned}
& P\left[\frac{1}{n}\|\Pi_{\tilde{W}_n(X)}\epsilon\|^2 \geq \frac{5}{8}\sigma^2 | (X_1, \ldots, X_n)\right] \\
= \; & P\left[\frac{1}{\lfloor n/2 \rfloor}\sum_{j=1}^{\lfloor n/2 \rfloor}((\epsilon'_j)^2 - \sigma^2) \geq \frac{1}{\lfloor n/2 \rfloor}\left(\frac{5}{8}n\sigma^2 - \lfloor \frac{n}{2} \rfloor \sigma^2\right) | (X_1, \ldots, X_n)\right] \\
\leq \; & P\left[\frac{1}{\lfloor n/2 \rfloor}\sum_{j=1}^{\lfloor n/2 \rfloor}((\epsilon'_j)^2 - \sigma^2) \geq C\sigma^2 | (X_1, \ldots, X_n)\right]
\end{aligned}$$

for some positive C. Then with technical tools similar to Rosenthal inequalities, Baraud (2000) proves that

$$P\left[\frac{1}{n}\|\Pi_{\tilde{W}_n(X)}\epsilon\|^2 \geq \frac{5}{8}\sigma^2\right]\frac{C}{n}.$$

2.6.3 Proof of Theorem 2.4.1

The proof of Theorem 2.4.1 relies on the following decomposition of the risk:

$$\mathbb{E}\left[\|\widehat{f}^-_{\widehat{m}} - f\|^2\right] \leq 2\mathbb{E}\left[\|\widehat{f}^-_{\widehat{m}} - f^-\|^2\right] + 2\mathbb{E}\left[\|f^- - f\|^2\right]. \tag{2.16}$$

$\star$ On the one hand, the estimator $\widehat{f}^-_{\widehat{m}}$ is built applying the procedure of Section 2.3.1 to the residuals $\{\widehat{\epsilon}_i\}$'s, which are i.i.d. given Z^- with common density f^-, thus the term

$\mathbb{E}\left[\|\widehat{f}^{-}_{\widehat{m}} - f^{-}\|^2 | Z^{-}\right]$ is upper bounded with Theorem 2.3.1.

$\star$ On the other hand, for every $i = 1, \ldots, n$,

$$\widehat{\epsilon}_i = Y_i - \tilde{b}(X_i) = \epsilon_i + (b - \tilde{b})(X_i)$$

thus the difference between the densities of $\widehat{\epsilon}_i$ and ϵ_i depends on $(b - \tilde{b})$.

More precisely, there exist absolute constants C_1 and C_2' such that almost surely

$$\begin{aligned}
&\mathbb{E}\left[\|\widehat{f}^{-}_{\widehat{m}} - f^{-}\|^2 | Z^{-}\right] \\
&\leq C_1 \inf_{m \in J_n} \left\{ \inf_{t \in S_m} \|f^{-} - t\|^2 + \frac{K^2 D_m}{n} \right\} + \frac{C_2'}{n} \left(1 + \|f^{-}\| \Gamma \sum_{D \in \mathbb{N}^*} D^{R+1/2} \exp\left(- \frac{K\sqrt{D}}{\|f^{-}\|} \right) \right) \\
&\leq C_1 \inf_{m \in J_n} \left\{ \inf_{t \in S_m} 2\|f - t\|^2 + \frac{K^2 D_m}{n} \right\} + \frac{C_2'}{n} \left(1 + \|f^{-}\| \Gamma \sum_{D \in \mathbb{N}^*} D^{R+1/2} \exp\left(- \frac{K\sqrt{D}}{\|f^{-}\|} \right) \right) + 2\|f^{-} - f\|^2.
\end{aligned}$$

Hence,

$$\mathbb{E}\left[\|\widehat{f}^{-}_{\widehat{m}} - f^{-}\|^2\right] \leq C_1 \inf_{m \in J_n} \left\{ \inf_{t \in S_m} 2\|f - t\|^2 + \frac{K^2 D_m}{n} \right\} + \mathbb{E}[\|f^{-}\|] \frac{C_2}{n} + 2\mathbb{E}\left[\|f^{-} - f\|^2\right]. \quad (2.17)$$

Besides, according to the expression of f^{-} in (2.9), almost surely,

$$\begin{aligned}
\|f^{-} - f\|^2 &= \int_{\mathbb{R}} \left(f^{-}(y) - f(y)\right)^2 dy \\
&= \int_{\mathbb{R}} \left(\int_0^1 \left[f(y - (b - \tilde{b})(x)) - f(y) \right] f_X(x) dx \right)^2 dy \\
&\leq \int_{\mathbb{R}} \int_0^1 \left[f(y - (b - \tilde{b})(x)) - f(y) \right]^2 f_X(x) dx dy \\
&= \int_0^1 \left(\int_{\mathbb{R}} \left[f(y - (b - \tilde{b})(x)) - f(y) \right]^2 dy \right) f_X(x) dx \qquad (2.18)
\end{aligned}$$

If $\mathbf{H_{err(1)}}$ holds, f is supported on $[-1, 1]$ so for every $x \in [-1, 1]$, the support of the application

$$y \to f(y - (b - \tilde{b})(x)) - f(y)$$

is included in $V(x) = [-1, 1] \cup [(b - \tilde{b})(x) - 1, (b - \tilde{b})(x) + 1]$. Thus

$$\begin{aligned}\|f^- - f\|^2 &\leq \int_0^1 \left(\mathrm{Lip}(f)^2(b-\tilde{b})^2(x)\int_{V(x)} dy\right) f_X(x)dx \\ &\leq 4\mathrm{Lip}(f)^2\|b-\tilde{b}\|_{f_X}^2.\end{aligned}$$

Therefore

$$\mathbb{E}\left[\|f^- - f\|^2\right] \leq 4\mathrm{Lip}(f)^2\mathbb{E}\left[\|b-\tilde{b}\|_{f_X}^2\right]. \tag{2.19}$$

Now, assume that $\mathbf{H_{err(2)}}$ holds. For every function h, let us denote by h^* the Fourier transform of h. For every $x \in [0,1]$,

$$\begin{aligned}\int_{\mathbb{R}} \left[f(y-(b-\tilde{b})(x)) - f(y)\right]^2 dy &= \frac{1}{2\pi}\int_{\mathbb{R}} \left|f^*(\lambda)e^{-i\lambda(b-\tilde{b})(x)} - f^*(\lambda)\right|^2 d\lambda \\ &= \frac{1}{2\pi}\int_{\mathbb{R}} |f^*(\lambda)|^2 \left|2\sin\left(\frac{\lambda(b-\tilde{b})(x)}{2}\right)\right|^2 d\lambda.\end{aligned}$$

For every $u \in \mathbb{R}$, $|\sin(u)| \leq |u|$, hence

$$\begin{aligned}\int_{\mathbb{R}} \left[f(y-(b-\tilde{b})(x)) - f(y)\right]^2 dy &\leq \frac{4}{2\pi}\int_{\mathbb{R}} |f^*(\lambda)|^2 \left|\frac{\lambda(b-\tilde{b})(x)}{2}\right|^2 d\lambda \\ &\leq \frac{1}{2\pi}(b-\tilde{b})^2(x)\int_{\mathbb{R}} |f^*(\lambda)\lambda|^2 d\lambda \\ &= \frac{1}{2\pi}(b-\tilde{b})^2(x)\int_{\mathbb{R}} |(f')^*(\lambda)|^2 d\lambda \\ &= (b-\tilde{b})^2(x)\|f'\|^2\end{aligned}$$

since $(f')^*(\lambda) = \lambda f^*(\lambda)$ according to Assumption $\mathbf{H_{err(2)}}$. Then, by (2.18),

$$\begin{aligned}\|f^- - f\|^2 &\leq \int_0^1 \left(\int_{\mathbb{R}} \left[f(y-(b-\tilde{b})(x)) - f(y)\right]^2 dy\right) f_X(x)dx \\ &\leq \|f'\|^2\int_0^1 (b-\tilde{b})^2(x) f_X(x)dx \\ &= \|f'\|^2\|b-\tilde{b}\|_{f_X}^2.\end{aligned}$$

Therefore

$$\mathbb{E}\left[\|f^- - f\|^2\right] \leq \|f'\|^2\mathbb{E}\left[\|b-\tilde{b}\|_{f_X}^2\right]. \tag{2.20}$$

Thus (2.19) and (2.20), together with Theorem 2.3.2, entail:

$$\mathbb{E}\left[\|f^- - f\|^2\right] \leq C_3' \inf_{m \in J_n'} \left\{ \|b - b_m\|_{f_X}^2 + \sigma^2 \frac{D_m}{n} \right\} + \frac{C_4}{n} \tag{2.21}$$

for some constants C_3' and C_4.

Moreover, almost surely

$$\begin{aligned} \|f^-\| &= \int_{\mathbb{R}} \left(\int_0^1 f(y - (b - \tilde{b})(x)) f_X(x) dx \right)^2 dy \\ &\leq \int_0^1 \left(\int_{\mathbb{R}} \left[f(y - (b - \tilde{b})(x)) \right]^2 dy \right) f_X(x) dx \\ &= \int_0^1 \left(\int_{\mathbb{R}} (f(z))^2 dz \right) f_X(x) dx \\ &= \|f\|^2. \end{aligned} \tag{2.22}$$

Hence,

$$\left(1 + \|f^-\| \Gamma \sum_{D \in \mathbb{N}^*} D^{R+1/2} \exp\left(-\frac{K\sqrt{D}}{\|f^-\|} \right) \right) \leq \left(1 + \|f\| \Gamma \sum_{D \in \mathbb{N}^*} D^{R+1/2} \exp\left(-\frac{K\sqrt{D}}{\|f\|} \right) \right).$$

Then (2.16), (2.17), (2.21) and (2.22) end the proof of Theorem 2.4.1. □

Chapitre 3

Estimation de densité par sélection de modèle ponctuelle - application à l'estimation ponctuelle de l'erreur de régression

Ce chapitre comprend deux résultats. Tout d'abord, nous proposons une méthode d'estimation de densité par sélection de modèle ponctuelle, qui produit un estimateur adapté au risque quadratique en un point x_0 fixé. A partir d'un échantillon i.i.d. de densité g à support dans $\mathbb{R}$, une collection d'estimateurs $\{\widehat{g}_m, m \in J_n\}$ est construite par minimisation d'un contraste de projection sur une collection de sous-espaces vectoriels de $L^2(\mathbb{R})$. La procédure de sélection de modèle s'appuie sur une heuristique semblable à celle de Birgé and Massart (1998), même si la mise en oeuvre est très différente : pour tout m, la somme biais-variance du risque quadratique de $\widehat{g}_m$ en x_0 est estimé par une fonction de m appelée critère, et le modèle sélectionné est celui qui minimise ce critère. Comme dans le cadre de la sélection de modèle classique, la variance est majorée par un terme déterministe mais le biais est moins aisé à estimer. Pour cette raison, le résultat obtenu est une inégalité "presque" oracle, qui assure néanmoins que l'estimateur atteint la vitesse de convergence minimax adaptative sur des espaces de Besov.

Cette méthode est ensuite utilisée pour estimer la densité de l'erreur de régression ϵ dans le modèle de régression homoscédastique

$$Y = b(X) + \epsilon.$$

La démarche est semblable à celle adoptée au Chapitre 2 : l'échantillon observé (X_i, Y_i) est scindé en deux échantillons indépendants Z^- et Z^+. A partir de l'échantillon Z^-, nous calculons un estimateur $\tilde{b}$ de b puis nous appliquons la méthode d'estimation de densité aux résidus $\widehat{\epsilon}_i = Y_i - \tilde{b}(X_i)$ de l'échantillon Z^+. Le risque ponctuel de l'estimateur de la densité de l'erreur se décompose donc en deux termes : un risque d'estimation de densité majoré grâce au résultat vérifié par notre estimateur de densité ponctuel, et un terme lié à l'estimation des (ϵ_i) par les

$(\widehat{\epsilon}_i)$, qui fait apparaître le risque intégré de l'estimateur $\tilde{b}$. Enfin, les méthodes d'estimation de densité et de densité de l'erreur sont illustrées sur des données simulées.

Ce chapitre est une version legèrement modifiée de l'article Plancade (2009), paru dans Mathematical Methods of Statistics.

3.1 Introduction

Consider a sample (X_i, Y_i) from the homoscedastic regression framework:

$$Y_i = b(X_i) + \epsilon_i \tag{3.1}$$

where the (ϵ_i) are unobserved independent identically distributed (i.i.d.) variables with common density f, with zero mean and independent of the (X_i). The main goal of this chapter is to propose an estimator for the density of ϵ_i, and to provide an upper bound for the quadratic risk of this estimator at a fixed point x_0.

The main issue in regression problems is to predict Y_i by measuring only X_i. The first step in such study is the estimation of the regression function $b(x) = \mathbb{E}[Y|X = x]$. This question has already been studied at length. The second step consists in studying the variations of Y_i around its conditional mean, which are characterized by the density of the errors (ϵ_i).

The knowledge of an estimator of the error density has many applications: for example, it enables model validation and, combined with an estimator of the regression function, it provides confidence intervals for future observations Y. The reader is refered to Efromovich (2005) for practical applications. Many papers are devoted to density estimation but the difficulty in our problem is to estimate the density of a sample (ϵ_i) which is not observed. The natural approach, already developed in Chapter 2, consists in computing proxies of the (ϵ_i), i.e. quantities based on the data which estimate the true (ϵ_i), and applying to them a density estimation procedure as if they were the true error sample. Observing that $\epsilon_i = Y_i - b(X_i)$, we naturally estimate the errors by the residuals $(\widehat{\epsilon}_i = Y_i - \widehat{b}(X_i))$, where $\widehat{b}$ is an estimator of the regression function. To the author's knowledge, no paper studies pointwise estimation of the error density by any method.

The estimators presented in this chapter are based on a pointwise model selection procedure. We will use here the estimator $\widehat{b}$ of b proposed in Baraud (2002), constructed by a model selection procedure based on least square estimators. Although the principle of pointwise model selection is the same, the techniques to carry it out are different. In particular, the key tool to prove the adaptivity of classical model selection estimators is Talagrand Inequality, whereas the adaptivity of pointwise model selection estimators comes out of a simpler Bernstein Inequality. The techniques developed in this chapter are based on Laurent et al. (2008), in which they develop these methods in a different framework.

This chapter presents two results. On the one hand, we build a density estimator which proves to be adaptive for the pointwise risk over some classical classes of regularity. Such estimators have been constructed using kernel methods in Butucea (2001), with the same adaptivity properties, along with minimax results over Sobolev classes. Nevertheless, our estimator is completely data driven, whereas the estimation procedure in Butucea (2001) brings into play upper bounds on unknown quantities. The second result proceeds from the application of the above density estimation procedure to residuals from the framework (3.1). We get an estimator of the error density, whose pointwise rate of convergence is the maximum of these two rates:

the pointwise minimax rate of estimation of f we would get if the errors (ϵ_i) where observed and the L^2-minimax rate of estimation of b.

The chapter is organized as follows. In Section 3.2, we introduce the definitions and notations, in particular we define spaces of regularity and collections of models. Section 3.3 presents the density estimator, and its convergence properties. This density estimation procedure is used in Section 3.4 to produce an estimator of the error density. Section 3.5 is devoted to numerical results. The proofs are gathered in Section 3.6, 3.7 and 3.8. Section 3.6 is devoted to the results about density estimator, Section 3.7 contains the proof of the error density estimation theorem, and proofs of minor results are gathered in Section 3.8.

3.2 Definitions and notations

3.2.1 Notations

Let t be a function defined on an interval I of $\mathbb{R}$ and μ be a density on I. We consider several norms of t:

$$\|t\|_\infty = \sup_{x\in I} |t(x)|, \quad \|t\| = \left(\int_I t^2(x)dx\right)^{1/2}, \quad \|t\|_\mu = \left(\int_I t^2(x)\mu(x)dx\right)^{1/2}.$$

Besides, we consider the following spaces of functions over I:

$$L^2(I) = \{t : I \to \mathbb{R}, \|t\| < +\infty\}, \quad L^\infty(I) = \{t : I \to \mathbb{R}, \|t\|_\infty < +\infty\}.$$

Moreover, we denote by $Supp(t)$ the closure of the set $\{x \in I, t(x) \neq 0\}$. If t is a function k times differentiable, we denote by $t^{(k)}$ its k-th derivative.
For every set S, we denote by $\mathbb{1}_S$ the indicator function of S, that is $\mathbb{1}_S(x) = 1$ if $x \in S$ and $\mathbb{1}_S(x) = 0$ otherwise.
For every function $t : \mathbb{R} \to \mathbb{R}$, we denote by t^* the Fourier transform of t:

$$t^*(u) = \int_{x\in\mathbb{R}} t(x)e^{-iux}dx, \quad \forall u \in \mathbb{R}.$$

For every linear space S_m we denote by t_m the L^2-orthogonal projection of t onto S_m.

Finally, for every $x \in \mathbb{R}$, we denote by $\lfloor x \rfloor$ its integer part, that is $\lfloor x \rfloor \in \mathbb{Z}$ and:

$$\lfloor x \rfloor \leq x < \lfloor x \rfloor + 1.$$

Let $I \subset J$ be two subsets of $\mathbb{R}$, we denote by $J \setminus I = \{x \in J, x \notin I\}$. Finally, we denote by $o(1)$ a quantity such that $\lim_{n\to+\infty} o(1) = 0$.

3.2.2 Spaces of functions

We consider the following Sobolev classes, for every $\alpha, L > 0$:

$$W(\alpha, L) = \{F \in L^2(\mathbb{R}), \frac{1}{2\pi} \int_{\mathbb{R}} |F^*(u)|^2 u^{2\alpha} du \leq L^2\}.$$

The Hölder classes are defined as follows. For every $\beta, L > 0$, and r the largest integer less than β, let:

$$\mathcal{H}(\beta, L) = \{F \in L^2(\mathbb{R}), |F^{(r)}(x) - F^{(r)}(y)| \leq L|x - y|^{\beta - r}, \forall x, y \in \mathbb{R}\}.$$

3.2.3 Collections of models

Sine-cardinal basis

Let ϕ be the function defined on $\mathbb{R}$ by

$$\phi(x) = \frac{\sin(\pi x)}{\pi x}, \quad \forall x \in \mathbb{R}^*$$

and $\phi(0) = 1$. For every $m > 0$, $k \in \mathbb{Z}$, set

$$\phi_{m,k}(x) = \sqrt{m}\phi(mx - k), \quad \forall x \in \mathbb{R},$$

and set

$$A_m = \text{Vect}\{\phi_{m,k}, k \in \mathbb{Z}\} \tag{3.2}$$

Let $\mathfrak{A}_n$ be the collection of models which incorporates the models (S_m) for m belonging to a grid of step $1/B$, B being a fixed positive integer:

$$\mathfrak{A}_n = \{A_m, m \in \frac{1}{B}\mathbb{N}, m \leq M_n\}$$

and $M_n \leq n$. The following results hold:

Proposition 3.2.1 *1. The family $\{\phi_{m,k}, k \in \mathbb{Z}\}$ is orthonormal.*

2. For every $m > 0$,

$$\|\sum_{k \in \mathbb{Z}} \phi_{m,k}^2\|_\infty \leq m.$$

3. For every $0 < m < m'$, $A_m \subset A_{m'}$.

This result is proved in Section 3.8.

Wavelet basis

We consider also a collection of functions on $[-1,1]$ constructed from the compact wavelet decomposition. Let ψ be a r times differentiable function, called mother wavelet, supported on a compact set $[-B,B]$ and which satisfies the following conditions:

1) $\psi, \ldots, \psi^{(r)}$ are bounded on $[-B,B]$.

2) For every $0 \leq k \leq r$, and $\ell \geq 1$ there exists a constant C_ℓ such that
$$|\psi^{(k)}(x)| \leq C_\ell (1+|x|)^{-\ell} \ , \ \forall x \in [-B,B].$$

3) $\int_{-B}^{B} x^k \psi(x) dx = 0 \ , \ \forall \ 0 \leq k \leq r.$

4) The set of functions $\{\psi_{j,k} : x \to 2^{j/2}\psi(2^{j/2}x - k), (j,k) \in \mathbb{Z}^2\}$ is an orthonormal basis of $L^2(\mathbb{R})$.

Consider a r times differentiable function φ called the father wavelet, supported on $[-B,B]$ and which satisfies conditions 1) and 2) above, as well as the following conditions:

3') $\int_{-B}^{B} \varphi(x) dx = 1.$

4') The set of functions $\{\varphi_k : x \to \varphi(x-k), k \in \mathbb{Z}\} \cup \{\psi_{j,k}, j \in \mathbb{N}, k \in \mathbb{Z}\}$ is an orthonormal basis of $L^2(\mathbb{R})$.

See Meyer (1990) for examples of such functions ψ and φ. The set $\{\psi_{j,k}, j \geq 0, k \in \mathbb{Z}\} \cup \{\varphi_k, k \in \mathbb{Z}\}$ is an orthonormal basis of $L^2[-1,1]$. As ψ is supported on $[-B,B]$, the restriction of $\psi_{j,k}$ to $[-1,1]$ is identically equal to zero for all $j \in \mathbb{N}$ and $k \notin [-2^j - B, 2^j + B]$. Let us denote $\Gamma(j) = \mathbb{Z} \cap [-2^j - B, 2^j + B]$. Similarly, φ_k is identically equal to zero for all $k \notin [-B-1, B+1] = \Gamma'(0)$. Set

$$B_m = \text{Vect}(\{\psi_{j,k}, j = 0, \ldots, m-1, k \in \Gamma(j)\} \cup \{\varphi_k, k \in \Gamma'(0)\}). \tag{3.3}$$

It is clear that for every positive integers $m' \geq m$, $B_m \subset B_{m'}$. Now, we define

$$\mathfrak{B}_n = \{B_m, m \in \mathbb{N}^*, 2^m \leq M_n\}$$

with $M_n \leq n$. The following result holds:

Proposition 3.2.2 *There exists a constant K which only depends on the father and mother wavelets ψ and φ, such that, for every $m \in \mathbb{N}^*$,*

$$\| \sum_{j=0}^{m-1} \sum_{k \in \Gamma(j)} \psi_{j,k}^2 + \sum_{k \in \Gamma'(0)} \varphi_k^2 \|_\infty \leq K^2 2^m. \tag{3.4}$$

This result is proved in Section 3.8.

3.3 Density estimation by pointwise model selection

In this section, we present a density estimator which is adaptive for the pointwise risk, over classical classes of regularity. In Section 3.4, this procedure will be applied to the pseudo observations $\widehat{\epsilon}_i$ of ϵ_i to get an estimator of the error density.

3.3.1 Framework and assumptions

Let

$$(V_1, \dots, V_{2n}) \tag{3.5}$$

be a i.i.d. sample drawn from a density g supported on $I \subset \mathbb{R}$, which satisfies:

$\mathbf{H_{dens}}$: $\sup_{x \in I} |g(x)| = \nu < +\infty$.

Let $\mathcal{M}_n = \{S_m, m = 1, \dots, N_n\}$ be a collection of subsets of $L^2(I)$, and $\{D_m, m = 1, \dots, N_n\}$ a collection of positive integers smaller than or equal to n, such that the following assumption holds.

$\mathbf{H_{mod}}$: The collection $\mathcal{M}_n$ is nested, that is:

$$S_1 \subset S_2 \subset \cdots \subset S_{N_n}. \tag{3.6}$$

Thus, there exists an L^2-orthonormal basis $\{\chi_\lambda, \lambda \in I_n\}$ of S_{N_n}, such that, for every model m, S_m is spanned by $\{\chi_\lambda, \lambda \in I_m\}$ where I_m is a subset of I_n. Besides, we assume that $D_m \leq D_{m'}$ for every $m \leq m'$.

Moreover, assume that for some positive constant K, the following condition holds:

$$\| \sum_{\lambda \in I_m} \chi_\lambda^2 \|_\infty \leq K^2 D_m, \quad \forall m \in \{1, \dots, N_n\} \tag{3.7}$$

Finally, we assume that there exists a constant $M \geq 1$ such that for every $n \in \mathbb{N}$ and every $\alpha \in]0, 1[$ with $n^\alpha M \leq D_{N_n}$, there exists a model m which satisfies

$$\left(\frac{n}{\log n}\right)^\alpha \leq D_m \leq M \left(\frac{n}{\log n}\right)^\alpha. \tag{3.8}$$

Let $\beta > 0$, we assume that the bias term satisfies the following assumption.

$\mathbf{H_{bias}}(\beta)$: Denoting by g_m the L^2-projection of g on S_m, we assume that for some positive constant C_0,

$$\|g - g_m\|_\infty \leq C_0 D_m^{-\beta}, \quad \forall m \in \{1, \dots, N_n\} \tag{3.9}$$

3.3.2 A preliminary risk bound for non adaptive estimators

We split the sample (3.5) into two independent sequences:

$$Z_0 = (V_i)_{i\in\{1,\dots,n\}}, \quad Z_1 = (V_i)_{i\in\{n+1\dots,2n\}}. \tag{3.10}$$

The sequence Z_0 is used to compute the collection $\{\widehat{g}_m, m = 1,\dots,N_n\}$ of non adaptive estimators, and the sequence Z_1 to estimate the parameter $\nu = \|g\|_\infty$ that appears in the penalty. Let x_0 be a fixed point in I. For every model $m \in \{1,\dots,N_n\}$, the projection estimator $\widehat{g}_m$ of g on S_m, computed from the sample Z_0 is defined by

$$\widehat{g}_m = \sum_{\lambda\in I_m} \left(\frac{1}{n}\sum_{i=1}^n \chi_\lambda(V_i)\right)\chi_\lambda. \tag{3.11}$$

Observing that $\mathbb{E}[\widehat{g}_m(x_0)] = g_m(x_0)$ for every model m, the squared risk of the estimator $\widehat{g}_m$ at the point x_0 can be written as:

$$\mathbb{E}[(\widehat{g}_m - g)^2(x_0)] = \mathbb{E}[(\widehat{g}_m - g_m)^2(x_0)] + (g_m - g)^2(x_0).$$

Moreover,

$$\mathbb{E}[(\widehat{g}_m - g_m)^2(x_0)] = Var\left(\sum_{\lambda\in I_m}\left(\frac{1}{n}\sum_{i=1}^n \chi_\lambda(V_i)\right)\chi_\lambda(x_0)\right) = Var\left(\frac{1}{n}\sum_{i=1}^n\left(\sum_{\lambda\in I_m}\chi_\lambda(V_i)\chi_\lambda(x_0)\right)\right).$$

The (V_i) are i.i.d, thus

$$\begin{aligned}
\mathbb{E}[(\widehat{g}_m - g_m)^2(x_0)] &= \frac{1}{n}Var\left(\sum_{\lambda\in I_m}\chi_\lambda(V_1)\chi_\lambda(x_0)\right) \le \frac{1}{n}\mathbb{E}\left[\left(\left(\sum_{\lambda\in I_m}\chi_\lambda(V_1)\chi_\lambda(x_0)\right)^2\right]\right. \\
&= \frac{1}{n}\int_{x\in I}\left(\sum_{\lambda\in I_m}\chi_\lambda(x)\chi_\lambda(x_0)\right)^2 g(x)dx \\
&\le \frac{\nu}{n}\int_{x\in I}\left(\sum_{\lambda\in I_m}\chi_\lambda(x)\chi_\lambda(x_0)\right)^2 dx.
\end{aligned}$$

By developing the square in the integral, we get

$$\mathbb{E}[(\widehat{g}_m - g_m)^2(x_0)] \le \frac{\nu}{n}\sum_{\lambda,\lambda'\in I_m}\left(\int_{x\in I}\chi_\lambda(x)\chi_{\lambda'}(x)dx\right)\chi_\lambda(x_0)\chi_{\lambda'}(x_0)$$

Besides, the family $\{\chi_\lambda, \lambda \in I_m\}$ is orthonormal which ensures that

$$\mathbb{E}[(\widehat{g}_m - g_m)^2(x_0)] \le \frac{\nu}{n}\sum_{\lambda\in I_m}\chi_\lambda^2(x_0)$$

and inequality (3.7) yields

$$\mathbb{E}[(\widehat{g}_m - g_m)^2(x_0)] \leq K^2 \nu \frac{D_m}{n}. \tag{3.12}$$

This bound is standard for a variance term. Finally, for every model $m \in \{1, \ldots, N_n\}$ we have the following non adaptive bound for $\widehat{g}_m$:

$$\mathbb{E}[(\widehat{g}_m - g)^2(x_0)] \leq (g - g_m)^2(x_0) + K^2 \nu \frac{D_m}{n}. \tag{3.13}$$

In Section 3.3.4, we will select a model by a penalised criterion which requires to estimate the variance term $K^2 \nu D_m / n$. Thus, we present an estimator $\widehat{\nu}_n$ of ν.

3.3.3 Estimation of ν

In this section, we propose an estimator $\widehat{\nu}_n$ of $\nu = \|g\|_\infty$ constructed from the sample Z_1. We consider a collection of models which satisfies the following properties.

> $\mathbf{H}_\nu(\beta)$: Let $\mathcal{M}'_n = \{S'_m, m = 1, \ldots, N'_n\}$ be a collection of models. We suppose that for every model m, $\{\xi_\lambda, \lambda \in I'_m\}$ is an L^2-orthonormal basis of S'_m and the (ξ_λ) are continuous on I. Moreover, let $g_m^{(1)} = \arg\min_{t \in S'_m} \|g - t\|^2$, we assume that
>
> $$\|g - g_m^{(1)}\|_\infty \leq C_0 {D'_m}^{-\beta}$$
>
> for some positive integers $(D'_m)_{m=1,\ldots,N'_n}$.

Let m_0 be a model such that $p_0 = D_{m_0}$ satisfies

$$\left(\frac{n}{\log n}\right)^\gamma \leq p_0 \leq M \left(\frac{n}{\log n}\right)^\gamma$$

for some $\gamma \in]0, 1/2[$, where M is defined in (3.8). We define

$$\widehat{g}_m^{(1)} = \sum_{\lambda \in I'_m} \left(\frac{1}{n} \sum_{i=n+1}^{2n} \xi_\lambda(V_i) \right) \xi_\lambda \quad \text{and} \quad \widehat{\nu}_n = \|\widehat{g}_{m_0}^{(1)}\|_\infty.$$

The following result holds.

Proposition 3.3.1 *Suppose that Assumptions* $\mathbf{H_{dens}}$ *and* $\mathbf{H}_\nu(\beta)$ *hold for some* $\beta > 0$. *Then for every* n *such that*

$$(\mathbf{A_1}) \quad C_0 p_0^{-\beta} < \frac{\nu}{6},$$

we have

$$P\left[\widehat{\nu}_n \leq \frac{1}{2}\nu\right] \leq 2 \exp\left(-\frac{n\nu}{84 K^2 p_0}\right). \tag{3.14}$$

If in addition

$$(\mathbf{A_2}) \quad \frac{p_0}{\sqrt{n}} \leq \frac{\nu}{12K^2},$$

then

$$P\left[\widehat{\nu}_n \geq 2\nu\right] \leq \exp\left(-\frac{n\nu}{456K^2p_0}\right). \tag{3.15}$$

This result is proved in Section 3.6.2.

Comment 1.

1) There exists an integer N which depends on (K, β, C_0) such that for every $n \geq N$, $(\mathbf{A_1})$ and $(\mathbf{A_2})$ hold.

2) The collections of model in which $\widehat{\nu}_n$ and $\widehat{g}_{\widehat{m}}$ are computed can be different.

3.3.4 Construction of the adaptive estimator

The model selection procedure developed by Birgé and Massart relies on the following idea: the "best" model among the collection $\mathcal{M}_n$ is the one which minimizes the bias-variance sum in the right hand side of (3.13), thus the natural idea consists in building an estimator of this sum and selecting the model $\widehat{m}$ which minimizes it.

On the one hand, the variance term $K^2\nu D_m/n$ is estimated by $K^2\widehat{\nu}_n D_m/n$.

On the other hand, the estimation of the bias term $(g - g_m)^2(x_0)$ is the main distinct point between pointwise and global model selection procedures. In classical L^2-model selection, the bias term $\|g - g_m\|^2$ is estimated, up to a quantity independent of m, by $-\|\widehat{g}_m\|^2$ (see Massart (2007)), but this procedure cannot be transposed for the pointwise bias.

We note that, as j tends to infinity, the model S_j grows, and g_j tends to g. Therefore, instead of estimating $(g - g_m)^2(x_0)$, we estimate the term $\sup_{j,m\leq j\leq N_n}(g_j - g_m)^2(x_0)$ which has same order. This heuristic is confirmed as follows. By (3.9) in Assumption $\mathbf{H_{bias}}(\beta)$,

$$\begin{aligned}\sup_{j,m\leq j\leq N_n}(g_j - g_m)^2(x_0) &\leq 2\sup_{j,m\leq j\leq N_n}[(g_j - g)^2(x_0) + (g_m - g)^2(x_0)] \\ &\leq 2\sup_{j,m\leq j\leq N_n}[C_0D_j^{-2\beta} + C_0D_m^{-2\beta}] \leq 4C_0D_m^{-2\beta}\end{aligned}$$

and $(g - g_m)^2(x_0)$ has order $D_m^{-2\beta}$ as well.

Now, let the best theoretical model m_{opt} be defined by

$$m_{opt} = \arg\min_{m\in\{1,\ldots,N_n\}}[\sup_{j,m<j\leq N_n}(g_j(x_0) - g_m(x_0))^2 + pen(m)] = \arg\min_{m\in\{1,\ldots,N_n\}}[Crit(m)]$$

where $pen(m) = AK^2 x_m \widehat{\nu}_n \frac{D_m}{n}$, A is a constant greater than or equal to 1 and

$$x_m = \frac{45}{2} \log(1 + D_m) \max \left\{ 1, \frac{9K^2}{\widehat{\nu}_n} \log(1 + D_m) \frac{D_m}{n} \right\}. \tag{3.16}$$

Remark about the numerical constant in x_m: If the constant 45/2 is replaced by any constant $B > 8$, Theorem 3.3.1 would still hold, but with different constants (θ_i) (see section 3.3.5). Moreover, the condition $B > 8$ appears in theoretical upper bounds but in numerical simulations (see Section 3.5), the value $B = 10$ seems to perform well. Nevertheless, the empirical calibration of this constant, as well as the constant in the penalty below, involves a lot of simulation experiments. This is a general problem in model selection and it is not specific to pointwise model selection.

In view to estimate $Crit(m)$, the natural idea would be to replace $(g_j - g_m)^2(x_0)$ by $(\widehat{g}_j - \widehat{g}_m)^2(x_0)$, but this proceeding is clearly biased. In fact,

$$\mathbb{E}[(\widehat{g}_m - \widehat{g}_j)^2(x_0)] = (g_j - g_m)^2(x_0) + \mathbb{E}[((\widehat{g}_j - \widehat{g}_m)(x_0) - (g_j - g_m)(x_0))^2].$$

The term $\mathbb{E}[((\widehat{g}_j - \widehat{g}_m)(x_0) - (g_j - g_m)(x_0))^2]$ is upper bounded similarly to the variance term in (3.12). More precisely,

$$\begin{aligned}
& \mathbb{E}[((\widehat{g}_j - \widehat{g}_m)(x_0) - (g_j - g_m)(x_0))^2] = Var\left((\widehat{g}_j - \widehat{g}_m)(x_0)\right) \\
= \;& Var\left(\sum_{\lambda \in I_j \setminus I_m} (\frac{1}{n} \sum_{i=1}^{n} \chi_\lambda(V_i)) \chi_\lambda(x_0) \right) = \frac{1}{n} Var\left(\sum_{\lambda \in I_j \setminus I_m} \chi_\lambda(V_1) \chi_\lambda(x_0) \right) \\
\leq \;& \frac{\nu}{n} \int_{x \in I} \left(\sum_{\lambda \in I_j \setminus I_m} \chi_\lambda(x) \chi_\lambda(x_0) \right)^2 dx = \frac{\nu}{n} \sum_{\lambda \in I_j \setminus I_m} \chi_\lambda^2(x_0) \leq \frac{\nu}{n} \sum_{\lambda \in I_j} \chi_\lambda^2(x_0) \leq \nu K^2 \frac{D_j}{n}.
\end{aligned}$$

Now, the theoretical criterion $Crit(m)$ is estimated by

$$\widehat{Crit}(m) = \sup_{j, m < j \leq N_n} \left[(\widehat{g}_j - \widehat{g}_m)^2(x_0) - K^2 \widehat{\nu}_n x_j \frac{D_j}{n} \right]_+ + pen(m) \tag{3.17}$$

and $\widehat{m} = \arg\min_{m \in \{1, \ldots, N_n\}} \widehat{Crit}(m)$.

Our estimator of g is $\widehat{g}_{\widehat{m}}$.

3.3.5 Results

We can prove the following result about the risk of $\widehat{g}_{\widehat{m}}$ at x_0.

Theorem 3.3.1 *Suppose that Assumptions* $\mathbf{H_{bias}}(\beta)$, $\mathbf{H}_\nu(\beta)$ *and* $\mathbf{H_{mod}}$ *hold for some* $\beta > 0$ *with the constraint* $M(n/\log n)^{1/(2\beta+1)} \leq N_n$. *Suppose that* $(\mathbf{A_1})$, $(\mathbf{A_2})$ *and the following condition hold:*

$$(\mathbf{A_3}): \quad 1+\frac{M}{n\log n}\leq n \quad \textit{and} \quad \left(\frac{n}{\log n}\right)^{2\beta/(2\beta+1)} \geq \frac{18MK^2}{\nu}. \tag{3.18}$$

Then,

$$\mathbb{E}[(\widehat{g}_{\widehat{m}}-g)^2(x_0)] \leq \theta_1 \left(\frac{n}{\log n}\right)^{-\frac{2\beta}{2\beta+1}} + \mathcal{R}_n$$

where

$$\mathcal{R}_n = \frac{\theta_2}{n} + (\nu + K^2 D_{N_n})^2 \exp\left(-\frac{n\nu}{84K^2p_0}\right) + \theta_3 p_0^2 \exp\left(-\frac{n\nu}{456K^2p_0}\right)$$

and

$$\theta_1 = \max\left\{15, 4\left(3+\frac{2}{45\log(1+D_1)}\right)\right\} \times (2C_0^2 + 45AK^2\nu) + 4C_0^2$$
$$\theta_2 = 40K^2(\nu+16K^2)\left(\sum_{m=1}^{N_n}(1+D_m)^{-(1+1/4)}\right)$$
$$\theta_3 = \frac{45A}{2}(M+1)K^4\nu \max\left\{15, 4\left(3+\frac{2}{45\log(1+D_1)}\right)\right\}$$

Comment 2. Clearly, $\mathcal{R}_n$ is negligible with respect to the rate $(n/\log n)^{-2\beta/(2\beta+1)}$.

Assumption $\mathbf{H_{bias}}(\beta)$ couples the collection of models and the fact that g belongs to a certain space of regularity (through the exponent β). The following Proposition gives examples for which this assumption is satisfied.

Proposition 3.3.2 *1. Let (β, L) be two positive numbers, let A_m be the linear subset of $L^2(\mathbb{R})$ defined in (3.2) and $h_m = \arg\min_{t\in A_m}\|h-t\|$, for every $h \in L^2(\mathbb{R})$. There exists a constant $K(\beta)$ such that*

$$\|h-h_m\|_\infty \leq K(\beta)Lm^{-\beta}, \quad \forall h \in W(\beta+1/2, L).$$

2. Let (β, L) be two positive numbers, and r be an integer greater than β, let B_m be the linear subset of $L^2([-1,1])$ defined in (3.3) and $h_m = \arg\min_{t\in B_m}\|h-t\|$ for every $h \in L^2([-1,1])$. There exists a constant $K'(\beta)$ such that

$$\|h-h_m\|_\infty \leq K'(\beta)L(2^m)^{-\beta}, \quad \forall h \in \mathcal{H}(\beta, L).$$

This Proposition is proved in Section 3.8. Moreover, by Propositions 3.2.1 and 3.2.2, the collections $\mathfrak{A}_n$ and $\mathfrak{B}_n$ satisfy Assumption $\mathbf{H_{mod}}$ for $M=2$.

Comment 3. It is well-known that the minimax rate of convergence for pointwise density estimation over $W(\beta+1/2,L)$ or $\mathcal{H}(\beta,L)$ is $n^{-2\beta/(2\beta+1)}$ (see e.g. Tsybakov (2004) for Hölder classes, and Butucea (2001) for Sobolev spaces). Our estimator reaches this rate up to a logarithmic factor. Nevertheless, Lepski (1991) defines the adaptive minimax rate, which is the best rate of convergence for adaptive estimators over a range of classes of regularity, and proves that the logarithmic loss is unavoidable in adaptive estimation, in several frameworks (see Introduction, Section 1.2.3). Following this line, Butucea (2001) proves that the adaptive minimax rate over the classes $\{W(\beta+1/2,L),\beta>0\}$ for pointwise density estimation is $(n/\log n)^{-2\beta/(2\beta+1)}$. Hence if we consider the collection of models $\mathfrak{A}_n$, $\widehat{g}_{\widehat{m}}$ is adaptive minimax over Sobolev classes. Similar results are proved over Hölder classes, for example in a white noise model (see Lepski and Spokoiny (1997)), so we expect that the adaptive minimax rate in pointwise density estimation has the same order. Then if we consider the collection $\mathfrak{B}_n$, our estimator should be adaptive minimax over Hölder classes.

3.3.6 Comparison with Lepski method

The reference method in pointwise estimation is the one originally presented by Lepski (1991) and developed in many other papers (see Introduction, Section 1.2.7). In particular it was adapted to density estimation by Butucea (2001). This procedure provides adaptive rates of convergence, and even exact adaptive results (see Butucea (2001)). This means that the estimator gets the adapative rate of convergence, and also the best asymptotic constant on given classes of functions. Lepski estimators have better asymptotic properties than the estimator presented in this chapter, but the theoretical results remain asymptotic whereas the results presented here are non asymptotic. One can object than the large constants which appear in the term $\mathcal{R}_n$ in Theorem 3.3.1 require large-size samples, but these constants are much larger than the effective ones, as proved by simulations.

In more recent works, Lepskii and Goldenshluger (2009) prove oracle inequalities in the gaussian white noise framework, but as far as the author knows, these results have not been developed in density estimation framework.

3.4 Error density estimation

3.4.1 Framework, outline and preliminary results

We consider a $3n$-sample

$$(X_i,Y_i)_{i\in\{-n,\dots,-1\}\cup\{1,\dots,2n\}} \tag{3.19}$$

from the regression framework (4.1), where the (X_i) are i.i.d with density f_X supported on $[0,1]$, the (ϵ_i) are i.i.d, independent of the (X_i) and $\mathbb{E}(\epsilon_1)=0$. This section presents a procedure of estimation of the density f of the (ϵ_i). Let us formulate the outline of this procedure, which decomposes in three steps.

Step 1: From the sequence

$$Z^- = (X_i, Y_i)_{i\in\{-n,\dots,-1\}}, \tag{3.20}$$

we compute an estimator $\widehat{b}$ of the regression function b.

In Section 3.4.4, we recall an example of adaptive estimation procedure of the regression function, but the result that we establish in Theorem 3.4.1 holds for any estimator $\widehat{b}$ of b computed from the sequence Z^-.

Step 2: We compute the residuals of the sequence $(X_i, Y_i)_{\{1,\dots,2n\}}$, namely

$$\widehat{\epsilon}_i = Y_i - \widehat{b}(X_i), \quad \forall i \in \{1, \dots, 2n\}$$

Noting that $\epsilon_i = Y_i - b(X_i)$, the $\widehat{\epsilon}_i$ are natural proxies for the unobserved (ϵ_i). Given Z^-, the $(\widehat{\epsilon}_i)$ are i.i.d.. Let us denote by f^- their common density, which only depends on the sequence $(X_i, Y_i)_{i\in\{-n,\dots,-1\}}$. For every integrable function $t : \mathbb{R} \to \mathbb{R}$

$$\begin{aligned}
\mathbb{E}[t(\widehat{\epsilon}_1)|Z^-] &= \mathbb{E}[t((b-\widehat{b})(X_1)+\epsilon_1)|Z^-] \\
&= \int_{x=0}^{1}\int_{y\in\mathbb{R}} t((b-\widehat{b})(x)+y) f_X(x) f(y) dy dx \\
&= \int_{x=0}^{1}\int_{z\in\mathbb{R}} t(z) f_X(x) f(z-(b-\widehat{b})(x)) dz dx \\
&= \int_{z\in\mathbb{R}} t(z) \left[\int_{x=0}^{1} f(z-(b-\widehat{b})(x)) f_X(x) dx\right] dz
\end{aligned}$$

Hence,

$$f^-(z) = \int_{x=0}^{1} f(z-(b-\widehat{b})(x)) f_X(x) dx, \quad \forall z \in \mathbb{R}. \tag{3.21}$$

Step 3: We apply the density estimation procedure described in Section 3.3 to the residuals $(\widehat{\epsilon}_i)$.

Thus, the risk of the estimator of f results from two consecutive approximations of different nature: the first one consists in replacing the true (ϵ_i) by the residuals, and the second one is a density estimation error. These two approximations appear in the following inequality:

$$\mathbb{E}[(\widehat{f}^-_{\widehat{m}} - f)^2(x_0)] \le 2\{\mathbb{E}[(\widehat{f}^-_{\widehat{m}} - f^-)^2(x_0)] + \mathbb{E}[(f^- - f)^2(x_0)]\}. \tag{3.22}$$

We assume that the error density f satisfies the following Assumption.

$\mathbf{H_{error}}$: f is Lipschitz with constant $Lip(f)$, that is

$$|f(x) - f(y)| \le Lip(f)|x - y|, \quad \forall x, y \in I.$$

Besides, $\sup_{x\in I} |f(x)| = \nu < +\infty$.

We consider a collection of models $\mathcal{M}_n$ which satisfies Assumption $\mathbf{H_{mod}}$, and such that one of these two alternative assumptions holds.

$\mathbf{H^{(1)}_{bias-error}}(\beta)$: $f \in \mathcal{H}(\beta, L)$ and there exists a constant $C_0(\beta, L)$ such that, for every model $S_m \in \mathcal{M}_n$,

$$\|h - h_m\|_\infty \leq C_0(\beta, L) D_m^{-\beta}, \quad \forall h \in \mathcal{H}(\beta, L). \tag{3.23}$$

$\mathbf{H^{(2)}_{bias-error}}(\beta)$: $f \in W(\beta + 1/2, L)$ and there exists a constant $C_0(\beta, L)$ such that, for every model $S_m \in \mathcal{M}_n$,

$$\|h - h_m\|_\infty \leq C_0(\beta, L) D_m^{-\beta}, \quad \forall h \in W(\beta + 1/2, L). \tag{3.24}$$

Remark 5 *According to Proposition* 3.3.2, (3.23) *is satisfied if* $f \in \mathcal{H}(\beta, L)$ *and the collection* $\mathcal{M}_n$ *that we consider is the wavelet collection* $\mathfrak{B}_n$, *and* (3.24) *is satisfied if* $f \in W(\beta + 1/2, L)$ *and* $\mathcal{M}_n$ *is the sine-cardinal collection* $\mathfrak{A}_n$.

Remark 6 *Assumptions* $\mathbf{H^{(1)}_{bias-error}}(\beta)$ *and* $\mathbf{H^{(2)}_{bias-error}}(\beta)$ *are less general than Assumption* $\mathbf{H_{bias}}$ *in the density estimation Theorem. In fact, in order to apply the result of Section* 3.3, *we need the density* f^- *of the residuals to satisfy the Assumptions of Theorem* 3.3.1, *which is guaranteed under* $\mathbf{H^{(1)}_{bias-error}}(\beta)$ *or* $\mathbf{H^{(2)}_{bias-error}}(\beta)$. *This fact comes out of the following proposition.*

Proposition 3.4.1 *1) For every* $x \in \mathbb{R}$, $|f^-(x)| \leq \nu$ *a.s.*

2) For every β, L *positive,* $f \in \mathcal{H}(\beta, L) \quad \Rightarrow \quad f^- \in \mathcal{H}(\beta, L)$ *a.s.*

3) For every β, L *positive,* $f \in W(\beta + 1/2, L) \quad \Rightarrow \quad f^- \in W(\beta + 1/2, L)$ *a.s.*

Proposition 3.4.1 is proved in Section 3.8.

We consider another collection $\mathcal{M}'_n = \{S'_m, m = 1, \ldots, N_n\}$ (which can be equal to or different from $\mathcal{M}_n$) and for every m, $S'_m = Vect\{\xi_\lambda, \lambda \in I'_m\}$ and the (ξ_λ) are continuous on I. For every $h \in L^2(I)$, let $h_m^{(1)} = \arg\min_{t \in S'_m} \|h - t\|^2$. We suppose that one of these two alternative assumptions holds.

$\mathbf{H^{(1)}_{\nu-error}}(\beta) : f \in \mathcal{H}(\beta, L)$ and for every model $S'_m \in \mathcal{M}_n$,

$$\|h - h_m^{(1)}\|_\infty \leq C_0(\beta, L) D_m^{-\beta}, \quad \forall h \in \mathcal{H}(\beta, L)$$

$\mathbf{H^{(2)}_{\nu-error}}(\beta) : f \in W(\beta + 1/2, L)$ and for every model $S'_m \in \mathcal{M}_n$,

$$\|h - h_m^{(1)}\|_\infty \leq C_0(\beta, L) D_m^{-\beta}, \quad \forall h \in W(\beta + 1/2, L)$$

3.4.2 Definition of the estimator

Let us consider the $3n$-sample (3.19). Let $\widehat{b}$ be any estimator of b constructed from the sequence Z^- defined in (3.20). State

$$\widehat{\epsilon}_i = Y_i - \widehat{b}(X_i), \quad \forall i = 1, \dots, 2n.$$

Let $\mathcal{M}_n = \{S_m, m = 1, \dots, N_n\}$ be a collection of subsets of $L^2(I)$, $\{D_m, m = 1, \dots, N_n\}$ a collection of positive integers, and β a positive number such that Assumptions $\mathbf{H_{mod}}$, and $\mathbf{H^{(1)}_{bias-error}}(\beta)$ or $\mathbf{H^{(2)}_{bias-error}}(\beta)$ hold.

For every model $S_m = Vect\{\chi_\lambda, \lambda \in I_m\}$, let

$$\widehat{f}^-_m = \sum_{\lambda \in I_m} \left(\frac{1}{n} \sum_{i=1}^{n} \chi_\lambda(\widehat{\epsilon}_i) \right) \chi_\lambda. \tag{3.25}$$

Let $\mathcal{M}'_n = \{S'_m, m = 1, \dots, N'_n\}$ be a collection of subsets of $L^2(I)$, $\{D'_m, m = 1, \dots, N_n\}$ a collection of positive integers, such that Assumption $\mathbf{H^{(1)}_{\nu-error}}(\beta)$ or $\mathbf{H^{(2)}_{\nu-error}}(\beta)$ holds. Let m_0 be in $\{1, \dots, N'_n\}$ such that $p_0 = D_{m_0}$ satisfies

$$\left(\frac{n}{\log n} \right)^\gamma < p_0 < M \left(\frac{n}{\log n} \right)^\gamma \tag{3.26}$$

for some $\gamma \in]0, 1/2[$ and

$$\widehat{\nu}^-_n = \|(\widehat{f}^-_{m_0})^{(1)}\|_\infty \quad \text{where} \quad (\widehat{f}^-_{m_0})^{(1)} = \sum_{\lambda \in I'_m} \left(\frac{1}{n} \sum_{i=n+1}^{2n} \xi_\lambda(\widehat{\epsilon}_i) \right) \xi_\lambda. \tag{3.27}$$

Finally, let

$$\widehat{m} = \arg \min_{m=1,\dots,N_n} \left\{ \left[\sup_{j, m \le j \le N_n} (\widehat{f}^-_j - \widehat{f}^-_m)^2(x_0) - K^2 x_j \widehat{\nu}^-_n \frac{D_j}{n} \right]_+ + pen^-(m) \right\}$$

where $pen^-(m) = AK^2 x^-_m \widehat{\nu}^-_n \frac{D_m}{n}$ with

$$x^-_m = \frac{45}{2} \log(1 + D_m) \max \left\{ 1, \frac{9K^2}{\widehat{\nu}^-_n} \log(1 + D_m) \frac{D_m}{n} \right\}.$$

3.4.3 Result

The estimator $\widehat{f}^-_{\widehat{m}}$ satisfies the following result.

Theorem 3.4.1 *Suppose that Assumptions* $\mathbf{H}^{(\mathbf{i})}_{\mathbf{bias-error}}(\beta)$ *and* $\mathbf{H}^{(\mathbf{i})}_{\nu\mathbf{-error}}(\beta)$ *hold, for* $i = 1$ *or* 2 *and for some* $\beta \geq \beta' > 3/4$ *where* β' *is known. Suppose that Assumption* $\mathbf{H_{mod}}$ *holds with*

$$\left(\frac{n}{\log n}\right)^{1/(2\beta'+1)} \leq D_{N_n} \leq M\left(\frac{n}{\log n}\right)^{1/(2\beta'+1)}.$$

In the definition of $\widehat{\nu}_n^-$ *(see* (3.26)*), consider* γ *which satisfies*

$$\gamma \in \left]\frac{1}{\beta'(2\beta'+1)}, \min\left\{\frac{1}{\beta'+1}, \frac{4\beta'+1}{3(2\beta'+1)}\right\}\right[.$$

Then, for every n *such that* $1 + Mn/\log n \leq n$*, we have*

$$\mathbb{E}[(\widehat{f}^-_{\widehat{m}} - f)^2(x_0)] \leq \theta'_1\left(\frac{n}{\log n}\right)^{-2\beta/(2\beta+1)} + \mathcal{C}_n\mathbb{E}[\|\widehat{b} - b\|^2_{f_X}] + \mathcal{R}_n \tag{3.28}$$

where

$$\theta'_1 = \left((2C_0^2 + 45A\nu K^2)\max\left(15, , 3 + \frac{2}{45\log(1+D_1)}\right) + C_0^2\right)(M+1),$$

$$\mathcal{C}_n = Lip(f)^2 + 2\theta'_4\log n\left(\left(\frac{n}{\log n}\right)^{2/(2\beta'+1)-2\beta'\gamma} + \left(\frac{n}{\log n}\right)^{(2-4\beta')/(2\beta'+1)}\right)$$

for some constant C*,*

$$\mathcal{R}_n = 2\left(\nu + K^2M\left(\frac{n}{\log n}\right)^{1/(2\beta'+1)}\right)^2\exp\left(-\frac{\sqrt{n}}{7}\right) + \frac{\theta_2}{n}.$$

and θ_2 *is defined in Theorem* 3.3.1*.*

Remark 7 *We have* $\mathcal{C}_n = Lip(f)^2 + o(1)$*, and* $\mathcal{R}_n \leq \kappa'_1/n$ *where* κ'_1 *depends on* (ν, M, K, β')*.*

Comment 4. By (3.28), the rate of convergence of our estimator is upper bounded by the maximum of the two following rates:

- the rate of convergence of the estimator $\widehat{b}$ of b.
- the minimax rate of estimation we would obtain for f if the (ϵ_i) were directly observed, that is $(n/\log n)^{-2\beta/(2\beta+1)}$.

According to Comment 3 in Section 3.3.5, the rate of convergence of $\widehat{f}^{-}_{\widehat{m}}$ is clearly lower bounded by $(n/\log n)^{-2\beta/(2\beta+1)}$. On the other hand, the term $\mathbb{E}[\|\widehat{b}-b\|^2_{f_X}]$ seems to be avoidable. In an integrated risk context, Efromovich (2005) proposes an error density estimator whose rate of convergence does not depend on the risk of $\widehat{b}$. Nevertheless, stronger conditions are required. In particular, the densities of X_i and ϵ_i are supposed to be two times differentiable and the errors (ϵ_i) are supposed to be symmetrical. The convergence results in Efromovich (2005) are based on properties of trigonometric basis and are not easily transposable in a pointwise context.

Besides, in numerical examples, our error density estimator performs nearly as well as the estimator we would obtain if the (ϵ_i) were observed (see Figure 3.3, Section 3.5).

3.4.4 An estimator of b

In this section, we briefly recall the definition of the estimator $\widehat{b}$ of b built in Chapter 2 from Baraud (2002). This is the estimator which is implemented in the simulations. Consider the following conditions.

> $\mathbf{H_b}$: The density f_X of X_1 is supported on a compact J, and is lower bounded by $m_0 > 0$ and upper bounded by $m_1 < +\infty$. Besides, $\mathbb{E}[\epsilon_1^4] < +\infty$.

Let us consider a collection of finite dimensional models Σ_n which satisfies the following assumption.

> $\mathbf{H_{mod-b}}$: Σ_n is included in a global model S_n with dimension smaller than $n^{1/2-d}$ for some $d > 0$. Furthermore, there exists some nonnegative constants Γ and R such that for every integer n,
>
> $$Card\left(\{m \in \Sigma_n : D_m = D\}\right) \leq \Gamma D^R$$
>
> for every $D \in \mathbb{N}^*$. Finally, there exists a constant K such that:
>
> $$\|t\|_\infty \leq K\sqrt{N_n}\|t\|, \ \forall t \in S_n.$$

For every model $m \in \Sigma_n$, let $\widehat{b}_m$ be the least square estimator of b:

$$\widehat{b}_m = \arg\min_{t \in S_m} \gamma_n(t)^- \quad \text{where} \quad \gamma_n(t)^- = \frac{1}{n}\sum_{i=-n}^{-1}(Y_i - t(X_i))^2,$$

and the selected model is $\widehat{m} = \arg\min_{m\in\Sigma_n}[\gamma_n^-(\widehat{b}_m) + A'\widehat{\sigma}_n^2 D_m/n]$ where $A' > 1$ and $\widehat{\sigma}_n^2$ is an estimator of the variance of ϵ_1: let $\mathcal{V}_n$ be a space of dimension $\lfloor n/2 \rfloor$ which includes the global model S_n, then:

$$\widehat{\sigma}_n^2 = \frac{1}{n - E\lfloor n/2 \rfloor} \inf_{t\in\mathcal{V}_n}(Y_i - t(X_i))^2$$

Let us define $\tilde{b} = \widehat{b}_{\widehat{m}}$ if $\|\widehat{b}_{\widehat{m}}\| \leq n$ and $\widehat{b} = 0$ otherwise.

Proposition 3.4.2 *Under Assumptions* $\mathbf{H_b}$ *and* $\mathbf{H_{mod-b}}$,

$$\mathbb{E}[\|\tilde{b} - b\|_{f_X}^2] \leq C \inf_{m \in \Sigma_n} \left\{ \|b - b_m\|^2 + \sigma^2 \frac{D_m}{n} \right\} + \frac{C'}{n}$$

for some constant C *depending on* A' *and* m_1, *and* C' *depending on* $(\sigma, \mathbb{E}[\epsilon_1^4], m_0, m_1)$.

Finally, classical results about approximation theory in Besov spaces lead to the following statement: if b belong to the Besov ball $\mathcal{B}_2^{\alpha,\infty}(L)$ with $\alpha > 1/2$, $\mathbb{E}[\|\widehat{b} - b\|_{f_X}^2] \leq Cn^{-2\alpha/(2\alpha+1)}$. This entails the following Corollary:

Corollary 3.4.1 *Suppose that the Assumptions of Theorem* 3.4.1 *hold, as well as* $\mathbf{H_b}$ *and* $\mathbf{H_{mod-b}}$. *Then, if* b *belongs to the Besov space* $\mathcal{B}_p^{\alpha,\infty}$ *for some* $p > 0$ *and* $\alpha \geq \beta > 1/2$,

$$\mathbb{E}[(\widehat{f}_{\widehat{m}}^{-} - f)^2(x_0)] \leq \theta \left(\frac{n}{\log n} \right)^{-\frac{2\beta}{2\beta+1}}$$

for some constant θ *independent of* n.

In other words, if b is smoother than f, the rate of convergence of $\widehat{f}_{\widehat{m}}^{-}$ is the optimal rate we would get if the (ϵ_i) were directly observed.

3.5 Simulations

3.5.1 Density estimation

This section illustrates the density estimation procedure presented in Section 3.3, with the sine-cardinal collection of models $\mathcal{A}_n$ described in (3.2). According to the remark following the definition (3.16) of x_m, the calibration of the constants in x_m and $pen(m)$ are determined from a lot of simulation experiments: we choose a constant $B = 10$ in x_m and $A = 2$ in $pen(m)$. Moreover, we consider $M_n = \sqrt{n}$. We draw 50 samples $(V_1, \ldots, V_n)$ of size $n = 200$, 500, 2000 of i.i.d. variables with gaussian distribution (denoted by $\mathcal{N}(0,1)$) and with Laplace density $g(x) = (1/2)\exp(-|x|)$ (denoted by $\mathcal{L}(1)$). Let J be the set of 150 regularly spaced points in $[-5, 5]$. For each sample and for every point $x \in J$ we compute an estimator $\widehat{g}_{\widehat{m}}(x)$ as follows, assuming that the maximum of the density ν is known.

- First we compute the projection density estimators $(\widehat{g}_m(x))$ defined in (3.11) for every $m \in (1/10)\mathbb{N}$, $m \leq M_n$ and every $x \in J$.
- Then for every $x \in J$, we select the best model as:

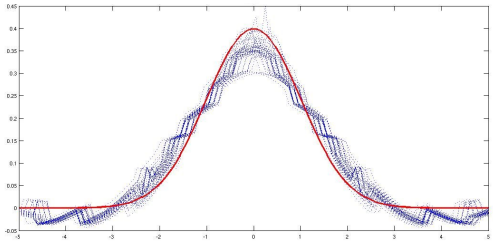
$$\widehat{m} = \arg\min \left\{ \sup_{m \leq j \leq N_n} [(\widehat{g}_j - \widehat{g}_m)^2(x) - \alpha\nu \log(1+j)\frac{j}{n}]_+ + \beta\nu \log(1+m)\frac{m}{n} \right\}$$

  with $\alpha = 5$ and $\beta = 10$.

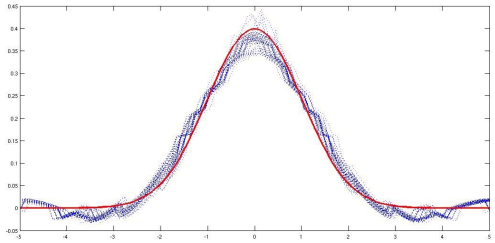
- We plot the set of points $\{(x, \widehat{g}_{\widehat{m}}(x)), x \in J\}$.

In Figure 3.1, each graph presents 50 estimated curves of $\widehat{g}_{\widehat{m}}$ for a given density g_i and a given n.

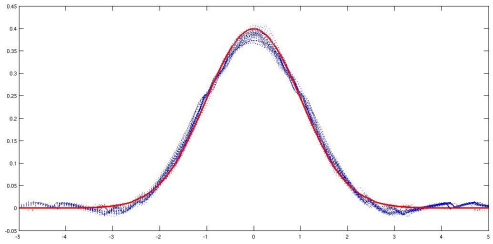
$\mathbf{V_i} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$



n=200

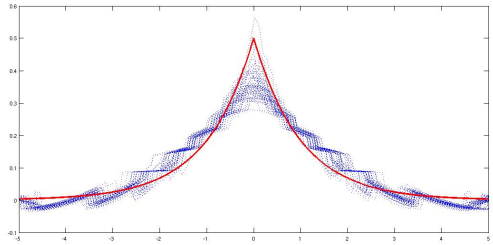


n=500

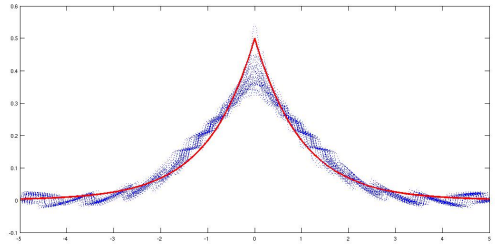


n=2000

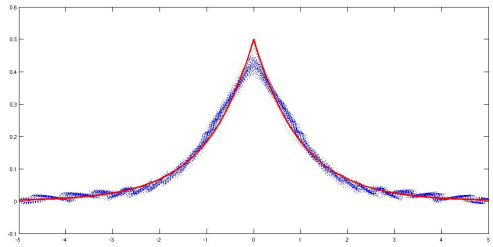
$\mathbf{V_i} \sim \mathcal{L}(\mathbf{1})$



n=200



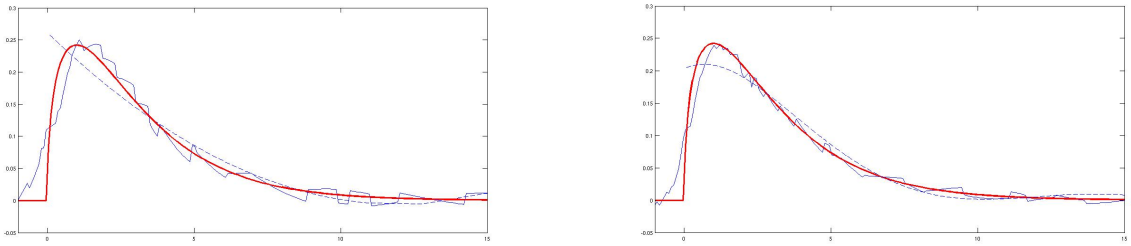
n=500



n=2000

Figure 3.1: Beam of 50 density estimators curves (blue dotted lines) built from i.i.d. samples of size n=200, 500 and 2000 of density $\mathcal{N}(0,1)$ and $\mathcal{L}(1)$ (red thick line), in sine-cardinal bases.

Figure 3.2 presents a comparison between our pointwise model selection estimator, and a global model selection estimator, computed following the procedure developed by Massart (2007), Section 7, for sample of size $n = 500$, 2000 with common density $\chi^2(3)$. The global model selection estimator (dotted blue line) is computed in a mixed piecewise polynomial and trigonometric polynomial basis using matlab programs available on Yves Rozenholc's web page (http://www.math-info.univ-paris5.fr/ rozen/), from a paper by Comte et al. (2008). The pointwise model selection estimator (solid blue line) is built following the procedure described above, on the set J of 150 regularly spaced points on $[-1, 15]$. We observe that the pointwise model selection estimator (solid blue line) fits the true density (red thick line) for a smaller sample size than the global model selection estimator.



n=500 n=2000

Figure 3.2: Pointwise model selection estimator (solid blue line) and global model selection estimator (dotted blue line) for a sample of size n=500, 2000 of density $\chi^2(3)$ (red thick line)

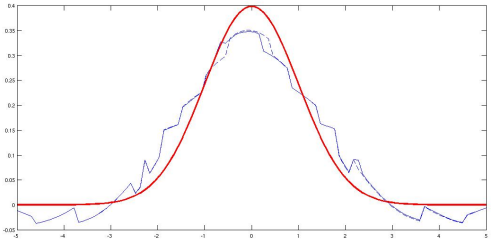
3.5.2 Error density estimation

This section proposes illustrations of the error density estimator described in Section 3.4, with the following procedure.

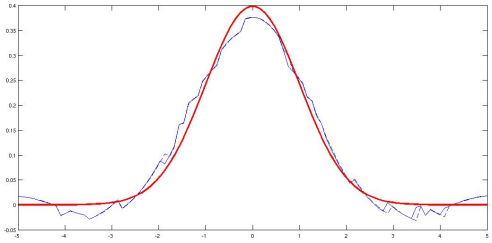
- We draw a sample $(X_1, \ldots, X_{2n})$ with common density f_X uniform on $[0, 1]$ and $\chi^2(3)$ and a sample $(\epsilon_1, \ldots, \epsilon_{2n})$ with common density f from a distribution $\mathcal{N}(0, 1)$ and $\mathcal{L}(1)$. We choose a regression function $b(x) = x^3 + 5x$ and $b(x) = \exp(-|x|)$ and compute the sample $(Y_1, \ldots, Y_{2n})$ where $Y_i = b(X_i) + \epsilon_i$.
- From the sample $\{(X_i, Y_i)\}_{i=1\ldots n}$, we compute an estimator $\widehat{b}$ of b following the procedure described in Section 3.4.4, using mixed piecewise polynomial and trigonometric polynomial basis.
- We compute the residuals from the second sample $(\widehat{\epsilon}_i)_{i=n+1,\ldots,2n}$, where $\widehat{\epsilon}_i = Y_i - \widehat{b}(X_i)$.
- Let J be a set of 150 regularly spaced points on $[-5, 5]$. We apply the density estimation procedure described in Section 3.5.1 to the residuals $(\widehat{\epsilon}_i)_{i=n+1,\ldots,2n}$.

Figure 3.3 presents the error density estimator (dotted blue line) and the theoretical estimator we obtain by applying the density estimation procedure of Section 3.5.1 directly to the sample $(\epsilon_i)_{i=n+1,\dots,2n}$. The thick line is the true density of ϵ_1.

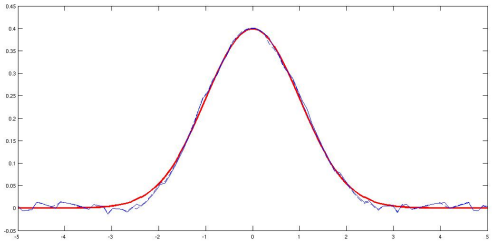
$\mathbf{X_i} \sim \mathcal{U}[\mathbf{0}, \mathbf{1}],\ \epsilon_\mathbf{i} \sim \mathcal{N}(\mathbf{0}, \mathbf{1}),\ \mathbf{b}(\mathbf{x}) = \mathbf{x^3} + \mathbf{5x}$



n=200

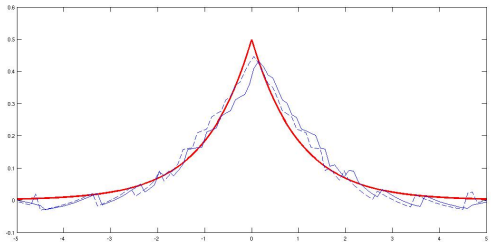


n=500

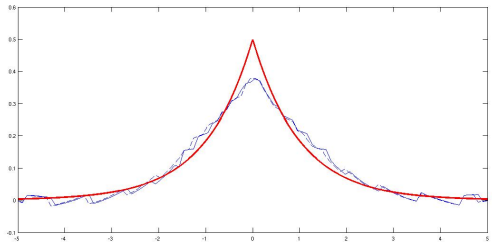


n=2000

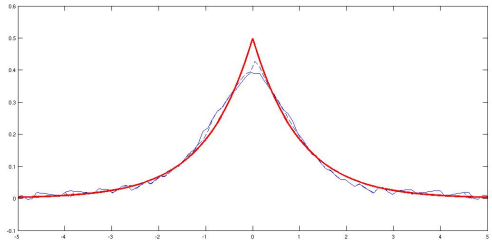
$\mathbf{X_i} \sim \chi^\mathbf{2}(\mathbf{3}),\ \epsilon_\mathbf{i} \sim \mathcal{L}(\mathbf{1}),\ \mathbf{b}(\mathbf{x}) = \exp(-|\mathbf{x}|)$



n=200



n=500



n=2000

Figure 3.3: Error density estimator (solid blue line), theoretical estimator we would get if the errors were observed (dotted blue line) and true density (thick red line).

We have also checked that the error density estimator hardly depends on the designs' distribution.

3.6 Proofs of Section 3.3

3.6.1 Proof of Theorem 3.3.1

The proof is divided in four claims. Let us denote by $\mathbb{E}_1[.] = \mathbb{E}[.|Z_1]$ the conditional expectation given Z_1, and $P_1[.] = P[.|Z_1]$ the probability given Z_1 (which is defined in (3.10)).

Claim 3.1 *Suppose that Assumptions* $\mathbf{H_{dens}}$ *and* $\mathbf{H_{mod}}$ *hold.*

$$\mathbb{E}_1[(\widehat{g}_{\widehat{m}}-g)^2(x_0)]\times \mathbb{1}_{\{\widehat{\nu}_n\geq \nu/2\}} \leq 2(g_{m_{opt}}-g)^2(x_0)+15\sup_{j,m_{opt}\leq j\leq N_n}(g_j-g_{m_{opt}})^2+(12+\frac{4}{x_{m_{opt}}})pen(m_{opt})+\frac{\theta_2}{n}.$$

This entails the following result.

Claim 3.2 *Under Assumptions* $\mathbf{H_{dens}}$, $\mathbf{H_{mod}}$, $\mathbf{H_{bias}}(\beta)$, *and* $(\mathbf{A_3})$

$$\mathbb{E}_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0)] \times \mathbb{1}_{\{\widehat{\nu}_n\geq \nu/2\}} \leq$$
$$\max(\kappa_2, \widehat{\nu}_n\kappa_3) \inf_{\{m=1,\ldots,N_n,(9K^2/\widehat{\nu}_n)\log(1+D_m)D_m/n\leq 1\}} \left[D_m^{-2\beta} + \log D_m \frac{D_m}{n}\right] + \frac{\theta_2}{n}$$

where

$$\kappa_1 = 4(3 + \frac{2}{45\log(1+D_1)}), \quad \kappa_2 = 2C_0^2[\max(15,\kappa_1)+2], \quad \kappa_3 = \max(15,\kappa_1)\frac{45AK^2}{2}.$$

We can deduce from Claim 3.2 the following inequality.

Claim 3.3 *Suppose that Assumptions* $\mathbf{H_{dens}}$, $\mathbf{H_{mod}}$, $\mathbf{H_{bias}}(\beta)$ *and* $(\mathbf{A_3})$ *hold. Moreover, assume that* $M(\log n/n)^{1/(2\beta+1)} \leq D_{N_n}$, *then*

$$\mathbb{E}_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0)] \times \mathbb{1}_{\{\widehat{\nu}_n\geq \nu/2\}} \leq \max(\kappa_2, \widehat{\nu}_n\kappa_3)(M+1)\left(\frac{n}{\log n}\right)^{\frac{-2\beta}{2\beta+1}} + \frac{\theta_2}{n}.$$

Besides, the following result holds.

Claim 3.4 *Under Assumption* $\mathbf{H_{dens}}$ *and* $\mathbf{H_{mod}}$, *for every model* $m \in \{1,\ldots,N_n\}$, *and every* $x \in I$,

$$|\widehat{g}_m(x)| \leq K^2 D_m \quad a.s.$$

The inequalities stated in Claims 3.3 and 3.4 enable us to prove Theorem 3.3.1. Indeed, on the one hand, by Claim 3.3,

$$\mathbb{E}[(\widehat{g}_{\widehat{m}} - g)^2(x_0)\mathbb{1}_{\{\widehat{\nu}_n \geq \nu/2\}}] \;\leq\; \mathbb{E}[\max(\kappa_2, \widehat{\nu}_n\kappa_3)](M+1)\left(\frac{n}{\log n}\right)^{-\frac{2\beta}{2\beta+1}} + \frac{\theta_2}{n}.$$

Moreover

$$\mathbb{E}[\max(\kappa_2, \widehat{\nu}_n\kappa_3)] \;\leq\; \mathbb{E}[\kappa_2 + \kappa_3\widehat{\nu}_n] \leq \kappa_2 + 2\kappa_3\nu + \kappa_3\mathbb{E}[\widehat{\nu}_n \mathbb{1}_{\{\widehat{\nu}_n \geq 2\nu\}}]$$

By Claim 3.4, $\widehat{\nu}_n \leq K^2 p_0$ almost surely, hence $\mathbb{E}[\max(\kappa_2, \widehat{\nu}_n\kappa_3)] \leq \kappa_2 + 2\nu\kappa_3 + \kappa_3 K^2 p_0 P[\widehat{\nu}_n \geq 2\nu]$, and under assumptions $(\mathbf{A_1})$ and $(\mathbf{A_2})$, inequality (3.15) of Proposition 3.3.1 yields

$$\mathbb{E}[\max(\kappa_2, \widehat{\nu}_n\kappa_3)] \leq \kappa_2 + 2\nu\kappa_3 + \kappa_3 K^2 p_0 \exp\left(-\frac{n\nu}{456K^2p_0}\right)$$

which induces that

$$\mathbb{E}[(\widehat{g}_{\widehat{m}} - g)^2(x_0)\mathbb{1}_{\{\widehat{\nu}_n \geq \nu/2\}}] \leq \left(\kappa_2 + 2\nu\kappa_3 + \kappa_3 K^2 p_0 \exp[-\frac{n\nu}{456K^2p_0}]\right)(M+1)\left(\frac{n}{\log n}\right)^{-\frac{2\beta}{2\beta+1}} + \frac{\theta_2}{n} \tag{3.29}$$

On the other hand, by Claim 3.4 and inequality (3.14) in Proposition 3.3.1,

$$\begin{aligned}\mathbb{E}[(\widehat{g}_{\widehat{m}} - g)^2(x_0)\mathbb{1}_{\{\widehat{\nu}_n < \nu/2\}}] &\leq (\nu + K^2 \max_{m=1,\ldots,N_n} D_m)^2 P\left[\widehat{\nu}_n < \frac{\nu}{2}\right] \\ &\leq (\nu + K^2 \max_{m=1,\ldots,N_n} D_m)^2 \exp\left(-\frac{n\nu}{84K^2p_0}\right) \qquad (3.30)\end{aligned}$$

and inequalities (3.29) and (3.30) provide the result of Theorem 3.3.1. □

Proof of Claim 3.1.

For every $j \in \{1, \ldots, N_n\}$, denote by

$$H(j) = K^2 x_j \widehat{\nu}_n \frac{D_j}{n}.$$

The proof of Claim 3.1 is based on the following steps: we exhibit a quantity $\mathcal{U}_{opt}$ such that

- $\mathbb{E}_1[\mathcal{U}_{opt}]$ has order $Crit(m_{opt})$.
- $\int_0^{+\infty} P_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0) - \mathcal{U}_{opt} \geq x]\mathbb{1}_{\{\widehat{\nu}_n \geq \nu/2\}} dx$ decreases to 0 with rate $1/n$.

Thus, inequality

$$\mathbb{E}_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0)]\mathbb{1}_{\{\widehat{\nu}_n \geq \nu/2\}} \leq \left(\mathbb{E}_1[((\widehat{g}_{\widehat{m}} - g)^2(x_0) - \mathcal{U}_{opt})_+] + \mathbb{E}_1[\mathcal{U}_{opt}]\right)\mathbb{1}_{\{\widehat{\nu}_n \geq /2\nu\}}$$
$$\leq \left(\int_0^{+\infty} P_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0) - \mathcal{U}_{opt} \geq x]dx + \mathbb{E}_1[\mathcal{U}_{opt}]\right)\mathbb{1}_{\{\widehat{\nu}_n \geq \nu/2\}} \tag{3.31}$$

yields the result of Claim 3.1. Let us consider a first result:

Lemma 3.6.1 *For every $\delta > 0$, $x > 0$ and for every model m*

$$P_1[\widehat{Crit}(m) \geq (1+\delta)Crit(m) + x] \leq 2\sum_{j=m}^{N_n} \exp\left(-C(x,j,\delta)\right) \qquad \textit{where}$$

$$C(x,j,\delta) = \min\left\{ \frac{1}{4\nu K^2(1+1/\delta)}\left(\frac{xn}{D_j} + K^2 x_j \widehat{\nu}_n\right), \frac{1}{4\sqrt{2(1+1/\delta)}K^2}\left(\frac{\sqrt{x}n}{D_j} + K\sqrt{x_j\widehat{\nu}_n \frac{n}{D_j}}\right)\right\}$$

Proof of Lemma 3.6.1

The empirical criterion $\widehat{Crit}(m)$ (defined in (3.17)) is built from $Crit(m)$ (defined in (3.16)) by replacing the unknown $(g_j - g_m)$ by its empirical counterpart $(\widehat{g}_j - \widehat{g}_m)$, so the deviation between $\widehat{Crit}(m)$ and $Crit(m)$ is upper bounded with Bernstein Inequality (see Introduction, Theorem 1.2.4). More precisely,

$$\begin{aligned} & P_1[\widehat{Crit}(m) \geq (1+\delta)Crit(m) + x] \\ = \; & P_1\left[\sup_{j,m \leq j \leq N_n} ((\widehat{g}_j - \widehat{g}_m)^2(x_0) - H(j))_+ \geq (1+\delta)\sup_{j,m \leq j \leq N_n} (g_j - g_m)^2(x_0) + x\right]. \end{aligned}$$

As $\sup_{j,m \leq j \leq n_N}(g_j - g_m)^2(x_0) + x$ is positive, we omit the positive part $(.)_+$.

$$\begin{aligned} & P_1[\widehat{Crit}(m) \geq (1+\delta)Crit(m) + x] \\ = \; & P_1\left[\sup_{j,m \leq j \leq N_n} ((\widehat{g}_j - \widehat{g}_m)^2(x_0) - H(j)) \geq (1+\delta)\sup_{j,m \leq j \leq N_n} (g_j - g_m)^2(x_0) + x\right] \\ \leq \; & \sum_{j=m}^{N_n} P_1[(\widehat{g}_j - \widehat{g}_m)^2(x_0) \geq (1+\delta)(g_j - g_m)^2(x_0) + x + H(j)] = \sum_{j=m}^{N_n} P_{j,m} \end{aligned}$$

and for every (j,m),

$$P_{j,m} = P_1\left[(\widehat{g}_j - \widehat{g}_m)^2(x_0) \geq (1+\delta)(g_j - g_m)^2(x_0) + (1+\frac{1}{\delta})\left(\sqrt{\frac{x + H(j)}{(1+1/\delta)}}\right)^2\right].$$

We recall that for every $x, y \in \mathbb{R}$, $(x+y)^2 \leq x^2(1+1/\delta) + y^2(1+\delta)$, thus

$$
\begin{aligned}
P_{j,m} &\leq P_1\left[(\widehat{g}_j - \widehat{g}_m)^2(x_0) \geq \left(|(g_j - g_m)(x_0)| + \sqrt{\frac{x+H(j)}{1+1/\delta}}\right)^2\right] \\
&= P_1\left[|(\widehat{g}_j - \widehat{g}_m)(x_0)| \geq |(g_j - g_m)(x_0)| + \sqrt{\frac{x+H(j)}{1+1/\delta}}\right] \\
&\leq P_1\left[|(\widehat{g}_j - \widehat{g}_m)(x_0) - (g_j - g_m)(x_0)| + |(g_j - g_m)(x_0)| \geq |(g_j - g_m)(x_0)| + \sqrt{\frac{x+H(j)}{1+1/\delta}}\right] \\
&= P_1\left[\left|\frac{1}{n}\sum_{i=1}^{n}(U_i - \mathbb{E}(U_i))\right| \geq \sqrt{\frac{x+H(j)}{1+1/\delta}}\right]. \qquad (3.32)
\end{aligned}
$$

where

$$U_i = \sum_{\lambda \in I_j} \chi_\lambda(V_i)\chi_\lambda(x_0) - \sum_{\lambda \in I_m} \chi_\lambda(V_i)\chi_\lambda(x_0) = \sum_{\lambda \in I_j \setminus I_m} \chi_\lambda(V_i)\chi_\lambda(x_0)$$

and $\mathbb{E}(U_i) = (g_j - g_m)(x_0)$. We have in view to upper bound the term (3.32) with Bernstein Inequality. Let us compute the terms v and c involved. Similarly to (3.12) we get:

$$\mathbb{E}_1(U_1^2) \leq \nu \sum_{\lambda \in I_j \setminus I_m} \chi_\lambda^2(x_0) \leq \nu \sum_{\lambda \in I_j} \chi_\lambda^2(x_0) \leq \nu K^2 D_j = v. \qquad (3.33)$$

Let ℓ be an integer greater than 2, then,

$$
\begin{aligned}
\mathbb{E}_1[(U_1)_+^\ell] &\leq \mathbb{E}_1[U_1^2] \times \|U_1\|_\infty^{\ell-2} \leq v \| \sum_{\lambda \in I_j \setminus I_m} \chi_\lambda(V_1)\chi_\lambda(x_0)\|_\infty^{\ell-2} \\
&\leq v\left[\| \sqrt{\sum_{\lambda \in I_j \setminus I_m} \chi_\lambda^2(V_1)}\|_\infty \sqrt{\sum_{\lambda \in I_j \setminus I_m} \chi_\lambda^2(x_0)}\right]^{\ell-2}
\end{aligned}
$$

and according to (3.7) in $\mathbf{H_{mod}}$, $\mathbb{E}_1[(U_1)_+^l] \leq v[K^2 D_j]^{l-2}$. So, we set

$$c = K^2 D_j. \qquad (3.34)$$

Finally, we denote by

$$\epsilon = \sqrt{\frac{x+H(j)}{1+1/\delta}} \geq \frac{1}{\sqrt{2(1+1/\delta)}}(\sqrt{x} + \sqrt{H(j)}). \qquad (3.35)$$

Then by Bernstein Inequality,

$$P_{j,m} \leq 2\exp\left(-\min(\frac{n\epsilon^2}{4v};\frac{n\epsilon}{4c})\right).$$

Moreover,

$$\frac{n\epsilon^2}{4v} = \frac{1}{4\nu K^2(1+1/\delta)}\left(\frac{xn}{D_j} + K^2 x_j \widehat{\nu}_n\right)$$
$$\frac{n\epsilon}{4c} \geq \frac{1}{4\sqrt{2(1+1/\delta)}K^2}\left(\frac{\sqrt{x}n}{D_j} + K\sqrt{x_j\widehat{\nu}_n\frac{n}{D_j}}\right).$$

This provides an upper bound of $P_{j,m}$ for every (j,m) which, inserted in inequality (3.32), ends the proof of Lemma 3.6.1. □

We derive from Lemma 3.6.1 the following result.

Lemma 3.6.2 *For every positive numbers δ and x, and every sequence Z_1,*

1) $P_1\left[\left\{(\widehat{g}_{\widehat{m}} - g)^2(x_0) \geq (1+\delta)\left(\sup_{j,m_{opt}\leq j\leq N_n}(g_j - g)^2(x_0) + Crit(m_{opt})\right) + 2x\right\} \cap \{\widehat{m} > m_{opt}\}\right]$
$\leq 4\sum_{m=1}^{N_n}\exp\left(-C(x,m,\delta)\right).$

2) $P_1[\{(\widehat{g}_{\widehat{m}} - g)^2(x_0) \geq 2(\widehat{g}_{m_{opt}} - g)^2(x_0) + 2H(m_{opt}) + 2(1+\delta)Crit(m_{opt}) + 2x\}$
$\cap\{\widehat{m} \leq m_{opt}\}] \leq 2\sum_{j=m_{opt}}^{N_n}\exp(-C(x,j,\delta)).$

Proof of Lemma 3.6.2

- Let us prove inequality 1).

$P_1[\{(\widehat{g}_{\widehat{m}} - g)^2(x_0) \geq (1+\delta)(\sup_{j,m_{opt}\leq j\leq N_n}(g_j - g)^2(x_0) + Crit(m_{opt})) + x\} \cap \{\widehat{m} > m_{opt}\}]$
$\leq P_1[\{(\widehat{g}_{\widehat{m}} - g)^2(x_0) \geq (1+\delta)\sup_{j,m_{opt}\leq j\leq N_n}(g_j - g)^2(x_0) + \widehat{Crit}(\widehat{m}) + x\} \cap \{\widehat{m} > m_{opt}\}]$
$+P_1[\widehat{Crit}(\widehat{m}) \geq (1+\delta)Crit(m_{opt}) + x]$ (3.36)

By definition of $\widehat{m}$, $\widehat{Crit}(\widehat{m}) = \inf_{m=1,\ldots N_n}\widehat{Crit}(m) \leq \widehat{Crit}(m_{opt})$. Hence, by Lemma 3.6.1,

$P_1[\widehat{Crit}(\widehat{m}) \geq (1+\delta)Crit(m_{opt}) + x] \leq P[\widehat{Crit}(m_{opt}) \geq (1+\delta)Crit(m_{opt}) + x]$
$\leq 2\sum_{j=m_{opt}}^{N_n}\exp\left(-C(x,j,\delta)\right) \leq 2\sum_{m=1}^{N_n}\exp\left(-C(x,m,\delta)\right).$ (3.37)

Besides it is clear that for every model m, $Crit(m) \geq pen(m)$, and if $\widehat{m} > m_{opt}$, $\sup_{j,m_{opt}\leq j\leq N_n}(g_j - g)^2(x_0) \geq (g_{\widehat{m}} - g)^2(x_0)$. So,

$$\begin{aligned}
P_1[\{(\widehat{g}_{\widehat{m}} - g)^2(x_0) \geq (1+\delta) \sup_{j,m_{opt}\leq j\leq N_n} (g_j - g)^2(x_0) + \widehat{Crit}(\widehat{m}) + x\} \cap \{\widehat{m} > m_{opt}\}] \\
\leq P_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0) \geq (1+\delta)(g_{\widehat{m}} - g)^2(x_0) + pen(\widehat{m}) + x] \\
\leq \sum_{m=1}^{N_n} P_1[(\widehat{g}_m - g)^2(x_0) \geq (1+\delta)(g_m - g)^2(x_0) + pen(m) + x] \qquad (3.38) \\
= \sum_{m=1}^{N_n} P_m. \qquad (3.39)
\end{aligned}$$

For every $m \in \{1, \ldots, N_n\}$, we have almost surely

$$(\widehat{g}_m - g)^2(x_0) \leq (1+\delta)(g - g_m)^2(x_0) + \left(1 + \frac{1}{\delta}\right)(\widehat{g}_m - g_m)^2(x_0)$$

and $pen(m) = AH(m) \geq H(m)$, so

$$\begin{aligned}
P_m &\leq P_1\left[(1+\frac{1}{\delta})(\widehat{g}_m - g_m)^2(x_0) \geq pen(m) + x\right] \\
&\leq P_1\left[|(\widehat{g}_m - g_m)^2(x_0)| \geq \sqrt{\frac{H(m)+x}{1+1/\delta}}\right] \\
&= P_1\left[\left|\frac{1}{n}\sum_{i=1}^{n}(U_i - \mathbb{E}(U_i))\right| \geq \sqrt{\frac{H(m)+x}{1+1/\delta}}\right]
\end{aligned}$$

where $U_i = \sum_{\lambda\in I_m} \chi_\lambda(V_i)\chi_\lambda(x_0)$. Similarly to the proof of Lemma 3.6.1, we apply Bernstein Inequality (Theorem 1.2.4) with the parameters defined in (3.33), (3.34) and (3.35), and obtain

$$P_m \leq 2\exp(-C(x, m, \delta)). \qquad (3.40)$$

Combining inequalities (3.36), (3.37), (3.38) and (3.40), the result of Lemma 3.6.2, 1) follows.

• Let us prove now inequality 2) in Lemma 3.6.2. First of all, by definition of $\widehat{m}$, $\widehat{Crit}(\widehat{m}) \leq \widehat{Crit}(m_{opt})$, so

$$\begin{aligned}
&P_1[\widehat{Crit}(m_{opt}) \geq (1+\delta)Crit(m_{opt}) + x] \geq P_1[\widehat{Crit}(\widehat{m}) \geq (1+\delta)Crit(m_{opt}) + x] \\
&\geq P_1[\sup_{j,\widehat{m}\leq j\leq N_n}[(\widehat{g}_j - \widehat{g}_{\widehat{m}})^2(x_0) - H(j)] + pen(\widehat{m}) \geq (1+\delta)Crit(m_{opt}) + x] \\
&\geq P_1[\{(\widehat{g}_{m_{opt}} - \widehat{g}_{\widehat{m}})^2(x_0) - H(m_{opt})] + pen(\widehat{m}) \geq (1+\delta)Crit(m_{opt}) + x\} \cap \{\widehat{m} \leq m_{opt}\}] \qquad (3.41)
\end{aligned}$$

Besides, $(\widehat{g}_{\widehat{m}} - g)^2(x_0) \leq 2(\widehat{g}_{m_{opt}} - \widehat{g}_{\widehat{m}})^2(x_0) + 2(\widehat{g}_{m_{opt}} - g)^2(x_0)$, therefore

$$(\widehat{g}_{m_{opt}} - \widehat{g}_{\widehat{m}})^2(x_0) \geq \frac{1}{2}(\widehat{g}_{\widehat{m}} - g)^2(x_0) - (\widehat{g}_{m_{opt}} - g)^2(x_0).$$

So we derive from (3.41) that

$$\begin{aligned} P_1[\widehat{Crit}(m_{opt}) \geq (1+\delta)Crit(m_{opt}) + x] &\geq P_1[\{\frac{1}{2}(\widehat{g}_{\widehat{m}} - g)^2(x_0) - (\widehat{g}_{m_{opt}} - g)^2(x_0) \\ &\geq (1+\delta)Crit(m_{opt}) + H(m_{opt}) - pen(\widehat{m}) + x\} \cap \{\widehat{m} \leq m_{opt}\}]. \end{aligned}$$

As $pen(\widehat{m})$ is positive, we get

$$\begin{aligned} &P_1[\widehat{Crit}(m_{opt}) \geq (1+\delta)Crit(m_{opt}) + x] \geq \\ &P_1[\{(\widehat{g}_{\widehat{m}} - g)^2(x_0) \geq 2(\widehat{g}_{m_{opt}} - g)^2(x_0) + 2H(m_{opt}) + 2(1+\delta)Crit(m_{opt}) + 2x\} \cap \{\widehat{m} \leq m_{opt}\}]. \end{aligned}$$

By Lemma 3.6.1, inequality 2) in Lemma 3.6.2 follows. □

Let us prove Claim 3.1. Consider

$$\mathcal{U}_{opt} = 2(\widehat{g}_{m_{opt}} - g)^2(x_0) + 2(1+\delta)Crit(m_{opt}) + 2H(m_{opt}) + (1+\delta) \sup_{j, m_{opt} \leq j \leq N_n} (g_j - g)^2(x_0).$$

Then, by inequalities 1) and 2) in Lemma 3.6.2, we get

$$P_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0) \geq \mathcal{U}_{opt} + x] \leq 4 \sum_{m=1}^{N_n} \exp[-C(x, m, \delta)].$$

Take $\delta = 4$, then

$$\begin{aligned} \mathbb{E}_1[((\widehat{g}_{\widehat{m}} - g)^2(x_0) - \mathcal{U}_{opt})_+] &\leq \int_0^{+\infty} P_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0) \geq \mathcal{U}_{opt} + x]dx \\ &= 2\int_0^{+\infty} P_1\left[(\widehat{g}_{\widehat{m}} - g)^2(x_0) \geq \mathcal{U}_{opt} + 2y\right] dy \qquad (3.42) \\ &\leq 8\int_0^{+\infty} \left(\sum_{m=1}^{N_n} \exp[-C(y, m, 4)]\right) dy \qquad (3.43) \end{aligned}$$

We recall that, for every positive constant C'

$$\int_0^{+\infty} \exp(-C'y)dy = \frac{1}{C'}, \quad \int_0^{+\infty} \exp(-C'\sqrt{y})dy = \frac{2}{C'^2}.$$

Therefore, according to the expression of $C(y,m,\delta)$ defined in Lemma 3.6.1,

$$\int_0^{+\infty} \sum_{m=1}^{N_n} \exp\left(-C(y,m,4)\right) dx$$
$$\leq \sum_{m=1}^{N_n} \left[5\nu K^2 \frac{D_m}{n} \exp\left(-\frac{x_m \widehat{\nu}_n}{5\nu}\right) + 80K^4 \frac{D_m^2}{n^2} \exp\left(-\frac{1}{2\sqrt{10}K} \sqrt{x_m \widehat{\nu}_n \frac{n}{D_m}}\right)\right].$$

Besides, assuming that $\widehat{\nu}_n \geq \nu/2$,

$$\begin{aligned} & x_m \geq \tfrac{45}{2} \log(1+D_m) \\ \Rightarrow \quad & x_m \geq \tfrac{45}{4} \tfrac{\nu}{\widehat{\nu}_n} \log(1+D_m) \\ \Leftrightarrow \quad & \exp\left(-\tfrac{x_m \widehat{\nu}_n}{5\nu}\right) \leq (1+D_m)^{-(2+1/4)} \\ \Leftrightarrow \quad & D_m \exp\left(-\tfrac{x_m \widehat{\nu}_n}{5\nu}\right) \leq (1+D_m)^{-(1+1/4)} \end{aligned} \tag{3.44}$$

and similarly,

$$\begin{aligned} & x_m \geq \tfrac{45}{2} \times \tfrac{9K^2}{\widehat{\nu}_n} \log^2(1+D_m) \tfrac{D_m}{n} \\ \Leftrightarrow \quad & \exp\left(-\tfrac{1}{2\sqrt{10}K} \sqrt{x_m \widehat{\nu}_n \tfrac{n}{D_m}}\right) \leq (1+D_m)^{-(2+1/4)} \\ \Leftrightarrow \quad & D_m \exp\left(-\tfrac{1}{2\sqrt{10}K} \sqrt{x_m \widehat{\nu}_n \tfrac{n}{D_m}}\right) \leq (1+D_m)^{-(1+1/4)}. \end{aligned} \tag{3.45}$$

Hence

$$\int_0^{+\infty} \sum_{m=1}^{N_n} \exp\left(-C(x,m,4)\right) dx \leq 5K^2(\nu + 16K^2) \left(\sum_{m=1}^{N_n} (1+D_m)^{1+1/4} \right) \frac{1}{n}.$$

Plugging these upper bounds in inequality (3.42) yields

$$\mathbb{E}_1[((\widehat{g}_{\widehat{m}} - g)^2(x_0) - \mathcal{U}_{opt})_+] \leq 40K^2(\nu + 16K^2) \left(\sum_{m=1}^{N_n} (1+D_m)^{1+1/4} \right) \frac{1}{n} = \frac{\theta_2}{n}. \tag{3.46}$$

It remains to upper bound $\mathbb{E}_1[\mathcal{U}_{opt}]$. As $\delta = 4$,

$$\begin{aligned} \mathbb{E}_1[\mathcal{U}_{opt}] &= 2\mathbb{E}[(\widehat{g}_{m_{opt}} - g)^2(x_0)] + 2\widehat{\nu}_n K^2 x_{m_{opt}} \frac{D_{m_{opt}}}{n} + 5 \sup_{j, m_{opt} \leq j \leq N_n} (g_j - g_{m_{opt}})^2(x_0) \\ &\quad + 10 Crit(m_{opt}) \\ &\leq 2[(g_{m_{opt}} - g)^2(x_0) + \nu K^2 \frac{D_{m_{opt}}}{n}] + 2\widehat{\nu}_n K^2 x_{m_{opt}} \frac{D_{m_{opt}}}{n} \\ &\quad + 5 \sup_{j, m_{opt} \leq j \leq N_n} (g_j - g_{m_{opt}})^2(x_0) + 10 Crit(m_{opt}). \end{aligned}$$

Thus on the set $\{\widehat{\nu}_n \geq \nu/2\}$, we have

$$\begin{aligned}\mathbb{E}_1[\mathcal{U}_{opt}]\mathbb{1}_{\{\widehat{\nu}_n \geq \nu/2\}} &\leq 2(g_{m_{opt}} - g)^2(x_0) + 4\widehat{\nu}_n K^2 \frac{D_{m_{opt}}}{n} + (2 + 10A)\widehat{\nu}_n K^2 x_{m_{opt}} \frac{D_{m_{opt}}}{n} \\ &\quad + 15 \sup_{j, m_{opt} \leq j \leq N_n} (g_j - g_{m_{opt}})^2 \\ &\leq 2(g_{m_{opt}} - g)^2(x_0) + 15 \sup_{j, m_{opt} \leq j \leq N_n} (g_j - g_{m_{opt}})^2 + (12 + \frac{4}{x_{m_{opt}}}) pen(m_{opt}).\end{aligned} \tag{3.47}$$

Putting together inequalities (3.31), (3.46) and (3.47), we get

$$\begin{aligned}&\mathbb{E}_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0)] \times \mathbb{1}_{\{\widehat{\nu}_n \geq \nu/2\}} \leq 2(g_{m_{opt}} - g)^2(x_0) + 15 \sup_{j, m_{opt} \leq j \leq N_n} (g_j - g_{m_{opt}})^2 \\ &+ \left(12 + \frac{4}{x_{m_{opt}}}\right) pen(m_{opt}) + \frac{\theta_2}{n}\end{aligned}$$

which ends the proof of Claim 3.1. □

Proof of Claim 3.2

First of all, note that

$$4\left(3 + \frac{1}{x_{m_{opt}}}\right) \leq 4\left(3 + \frac{2}{45\log(1 + D_{m_{opt}})}\right) \leq 4\left(3 + \frac{2}{45\log(1 + D_1)}\right) = \kappa_1.$$

Assume that $\widehat{\nu}_n \geq \nu/2$, and denote

$$\begin{cases} F(m) = D_m^{-2\beta} + \log(1 + D_m)\frac{D_m}{n} \\ m_1 = \arg\min\{F(m), m = 1, \ldots, N_n, (9K^2/\widehat{\nu}_n)\log(1 + D_m)D_m/n \leq 1\} \end{cases}$$

Thus, $x_{m_1} = (45/2)\log(1 + D_{m_1})$. We consider two situations: the case where $m_{opt} \geq m_1$, and the case where $m_{opt} < m_1$.

- If $m_{opt} \geq m_1$, by $\mathbf{H_{bias}}(\beta)$,

$$(g_{m_{opt}} - g)^2(x_0) \leq C_0^2 D_{m_{opt}}^{-2\beta} \leq C_0^2 D_{m_1}^{-2\beta}.$$

Besides, it is obvious that $11 \leq \kappa_1$. Thus by Claim 3.1, we get

$$\mathbb{E}_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0)]\mathbb{1}_{\{\widehat{\nu}_n \geq \nu/2\}} \leq 2C_0^2 D_{m_1}^{-2\beta} + \kappa_1 Crit(m_{opt}) + \frac{\theta_2}{n}.$$

As $m_{opt} = \arg\min_{m=1,\dots,N_n} Crit(m)$, $Crit(m_{opt}) \leq Crit(m_1)$. Then,

$$\begin{aligned}
\mathbb{E}_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0)]\mathbb{1}_{\{\widehat{\nu}_n \geq \nu/2\}} &\leq 2C_0^2 D_{m_1}^{-2\beta} + \kappa_1 Crit(m_1) + \frac{\theta_2}{n} \\
&\leq 2C_0^2(1+\kappa_1)D_{m_1}^{-2\beta} + \kappa_1 \frac{45AK^2}{2}\widehat{\nu}_n \log(1+D_{m_1})\frac{D_{m_1}}{n} + \frac{\theta_2}{n} \\
&\leq \max\{2C_0^2(1+\kappa_1), \kappa_1 \frac{45AK^2}{2}\widehat{\nu}_n\} F(m_1) + \frac{\theta_2}{n}. \qquad (3.48)
\end{aligned}$$

- If $m_{opt} < m_1$,

$$\begin{aligned}
(g_{m_{opt}} - g)^2(x_0) &\leq 2(g_{m_{opt}} - g_{m_1})^2(x_0) + 2(g_{m_1} - g)^2(x_0) \\
&\leq 2 \sup_{j, m_{opt} \leq j \leq N_n} (g_j - g_{m_{opt}})^2(x_0) + 2C_0^2 D_{m_1}^{-2\beta}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathbb{E}_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0)]\mathbb{1}_{\{\widehat{\nu}_n \geq \nu/2\}} \\
\leq\ & 15 \sup_{j, m_{opt} \leq j \leq N_n} (g_j - g_{m_{opt}})^2(x_0) + \kappa_1 pen(m_{opt}) + 4C_0^2 D_{m_1}^{-2\beta} + \frac{\theta_2}{n} \\
\leq\ & \max(15, \kappa_1) Crit(m_{opt}) + 4C_0^2 D_{m_1}^{-2\beta} + \frac{\theta_2}{n} \\
\leq\ & \max(15, \kappa_1) Crit(m_1) + 4C_0^2 D_{m_1}^{-2\beta} + \frac{\theta_2}{n} \\
\leq\ & \max(15, \kappa_1)[2C_0^2 D_{m_1}^{-2\beta} + pen(m_1)] + 4C_0^2 D_{m_1}^{-2\beta} + \frac{\theta_2}{n} \\
\leq\ & \max\left\{2C_0^2\left(\max(15,\kappa_1)+2\right), \max(15,\kappa_1)\frac{45AK^2}{2}\widehat{\nu}_n\right\} F(m_1) + \frac{\theta_2}{n}. \qquad (3.49)
\end{aligned}$$

Moreover, it is clear that

$$\begin{aligned}
2C_0^2(1+\kappa_1) &\leq 2C_0^2\left(\max(15,\kappa_1)+2\right) \\
\kappa_1 \frac{45AK^2}{2} &\leq \max(15,\kappa_1)\frac{45AK^2}{2}.
\end{aligned}$$

Therefore, inequalities (3.48) and (3.49) yield the proof of Claim 3.2. □

Proof of Claim 3.3

Let m_2 be a model such that

$$\left(\frac{n}{\log n}\right)^{\frac{1}{2\beta+1}} \leq D_{m_2} \leq M\left(\frac{n}{\log n}\right)^{\frac{1}{2\beta+1}}. \qquad (3.50)$$

On the set $\{\widehat{\nu}_n \geq \nu/2\}$, by Assumption $(\mathbf{A_3})$,

$$\frac{9K^2}{\widehat{\nu}_n}\log(1+D_{m_2})\frac{D_{m_2}}{n} \leq \frac{18K^2}{\nu}\left(\frac{n}{\log n}\right)^{-2\beta/(2\beta+1)} \leq 1.$$

By definition of m_1,

$$F(m_1) \leq F(m_2) \leq M\frac{\log n}{n}\left(\frac{n}{\log n}\right)^{\frac{1}{2\beta+1}} + \left(\frac{n}{\log n}\right)^{-\frac{2\beta}{2\beta+1}} \leq (M+1)\left(\frac{n}{\log n}\right)^{-\frac{2\beta}{2\beta+1}}.$$

Thus we derive from Claim 3.2 that

$$\mathbb{E}_1[(\widehat{g}_{\widehat{m}}-g)^2(x_0)]\mathbb{1}_{\{\widehat{\nu}_n\geq\nu/2\}} \leq \max(\kappa_2,\widehat{\nu}_n\kappa_3)(M+1)\left(\frac{n}{\log n}\right)^{\frac{-2\beta}{2\beta+1}} + \frac{\theta_2}{n}. \quad \square$$

Proof of Claim 3.4

For every model m, $(\widehat{g}_m - g)^2(x_0) \leq (|\widehat{g}_m(x_0)| + \nu)^2$ almost surely. Besides,

$$\begin{aligned}(\widehat{g}_m)^2(x_0) &= \sum_{\lambda\in I_m}\left(\frac{1}{n}\sum_{i=1}^n \chi_\lambda(V_i))\chi_\lambda(x_0)\right)^2 \leq \frac{1}{n}\sum_{i=1}^n\left(\sum_{\lambda\in I_m}\chi_\lambda(V_i)\chi_\lambda(x_0)\right)^2 \leq \left\|\sum_{\lambda\in I_m}\chi_\lambda^2\right\|_\infty^2 \\ &\leq K^4 D_m^2 \end{aligned} \tag{3.51}$$

which provides the result of Claim 3.4. $\square$

3.6.2 Proof of Proposition 3.3.1

Let us prove inequality (3.14). Let $x_1 \in I$ be such that $g(x_1) \geq 5\nu/6$, then by definition of $\widehat{\nu}_n$,

$$\begin{aligned}P\left[\widehat{\nu}_n \leq \frac{\nu}{2}\right] &\leq P\left[\widehat{g}^{(1)}_{m_0}(x_1) \leq \frac{\nu}{2}\right] = P\left[(\widehat{g}^{(1)}_{m_0} - g_{m_0})(x_1) \leq \frac{5\nu}{6} - g_{m_0}(x_1) - \frac{\nu}{3}\right] \\ &\leq P\left[(\widehat{g}^{(1)}_{m_0} - g_{m_0})(x_1) \leq (g-g_{m_0})(x_1) - \frac{\nu}{3}\right].\end{aligned}$$

By Assumption $\mathbf{H_{bias}}(\beta)$,

$$P\left[\widehat{\nu}_n \leq \frac{\nu}{2}\right] \leq P\left[(\widehat{g}^{(1)}_{m_0} - g_{m_0})(x_1) \leq C_0 p_0^{-\beta} - \frac{\nu}{3}\right]$$

and by condition ($\mathbf{A_1}$),

$$\begin{aligned} P\left[\widehat{\nu}_n \leq \frac{\nu}{2}\right] &\leq P\left[(\widehat{g}_{m_0}^{(1)} - g_{m_0})(x_1) \leq -\frac{\nu}{6}\right] \leq P\left[|(\widehat{g}_{m_0}^{(1)} - g_{m_0})(x_1)| \geq \frac{\nu}{6}\right] \\ &= P\left[\left|\frac{1}{n}\sum_{i=n+1}^{2n} U_i - \mathbb{E}(U_i)\right| \geq \frac{\nu}{6}\right]. \end{aligned} \tag{3.52}$$

Now, apply Bernstein Inequality (Theorem 1.2.4) with the following parameters v and c.

$$\begin{aligned} \mathbb{E}[U_1^2] &= \mathbb{E}\left[\left(\sum_{\lambda\in I_{m_0}} \xi_\lambda(V_1)\xi_\lambda(x_1)\right)^2\right] = \int_I \left(\sum_{\lambda\in I_{m_0}} \xi_\lambda(x)\xi_\lambda(x_1)\right)^2 g(x)dx \\ &\leq \nu \sum_{\lambda,\lambda'\in I_{m_0}} \left(\int_I (\xi_\lambda(x)\xi_{\lambda'}(x)dx\right)\xi_\lambda(x_1)\xi_{\lambda'}(x_1)) = \nu \sum_{\lambda\in I_{m_0}} \xi_\lambda^2(x_1) \end{aligned}$$

since the family $\{\xi_\lambda\}$ is orthonormal. Finally, Assumption (3.7) in $\mathbf{H_{mod}}$ yields

$$\mathbb{E}[U_1^2] \leq \nu K^2 p_0 = v.$$

Let l be an integer greater than 2,

$$\begin{aligned} \mathbb{E}[(X_1)_+^l] &\leq \mathbb{E}[U_1^2] \times \|U_1\|_\infty^{l-2} \leq v \left\|\sum_{\lambda\in I_{m_0}} \xi_\lambda(V_1)\xi_\lambda(x_0)\right\|_\infty^{l-2} \\ &\leq v\left(\sqrt{\left\|\sum_{\lambda\in I_{m_0}} \xi_\lambda^2(V_1)\right\|_\infty}\sqrt{\sum_{\lambda\in I_{p_0}} \xi_\lambda^2(x_0)}\right)^{l-2} \leq v(K^2p_0)^{l-2} \end{aligned}$$

Hence we set $c = K^2p_0$. By Bernstein Inequality (Theorem 1.2.4, we derive from inequality (3.52) that

$$P\left[\widehat{\nu}_n \leq \frac{\nu}{2}\right] \leq 2\exp\left(-\frac{n\nu}{84K^2p_0}\right)$$

which is the result we wanted to prove.

Let us prove inequality (3.15). Let $\widehat{x}_1 \in I$ be such that $\widehat{g}_{m_0}(\widehat{x}_1) \geq 5\widehat{\nu}_n/6$. Similarly to (3.52), under condition ($\mathbf{A_1}$),

$$P\left[\nu \leq \frac{\widehat{\nu}_n}{2}\right] \leq P\left[|(\widehat{g}_{m_0}^{(1)} - g_{m_0})(\widehat{x}_1)| \geq \frac{\nu}{6}\right].$$

Moreover,

$$\begin{aligned} P\left[\nu \leq \frac{\widehat{\nu}_n}{2}\right] &\leq P\left[\sup_{x\in I} |(\widehat{g}^{(1)}_{m_0} - g_{m_0})(x)| \geq \frac{\nu}{6}\right] \\ &= P\left[\sup_{x\in I} \frac{1}{n}\sum_{i=n+1}^{2n}\left(\sum_{\lambda\in I_{m_0}} (\xi_\lambda(V_i)\xi_\lambda(x) - \mathbb{E}[\xi_\lambda(V_i)\xi_\lambda(x)])\right) \geq \frac{\nu}{6}\right] \\ &= P\left[\sup_{x\in I} \frac{1}{n}\sum_{i=n+1}^{2n} \varphi_x(V_i) \geq \frac{\nu}{6}\right] \end{aligned}$$

We have in view to apply Talagrand Inequality (see Introduction, Theorem 1.2.2), but the set of functions

$$\mathcal{F} = \left\{ \varphi_x : u \to \sum_{\lambda\in I_{m_0}} \xi_\lambda(x)\xi_\lambda(u) - \mathbb{E}[\xi_\lambda(x)\xi_\lambda(V_1)], x \in I \right\}$$

is not countable. Nevertheless, the (ξ_λ) are continuous, thus for every u the application $x \to \varphi_x(u)$ is continuous. Hence, since the set $\mathbb{Q}\cap I$ is dense in I, we have

$$Z = \sup_{x\in I} \frac{1}{n}\sum_{i=n+1}^{2n} \varphi_x(V_i) = \sup_{x\in I\cap\mathbb{Q}} \frac{1}{n}\sum_{i=n+1}^{2n} \varphi_x(V_i),$$

so

$$P\left[\nu \leq \frac{\widehat{\nu}_n}{2}\right] \leq P\left[\sup_{x\in I\cap\mathbb{Q}} \frac{1}{n}\sum_{i=n+1}^{2n} \varphi_x(V_i) \geq \frac{\nu}{6}\right].$$

and $\mathbb{Q}\cap I$ is countable. Let $x \in I$, by Cauchy Schwartz Inequality,

$$\begin{aligned} \left(\frac{1}{n}\sum_{i=n+1}^{2n} \varphi_x(V_i)\right)^2 &\leq \frac{1}{n}\sum_{i=n+1}^{2n} (\varphi_x(V_i))^2 = \frac{1}{n}\sum_{i=n+1}^{2n}\left(\sum_{\lambda\in I_{m_0}} (\xi_\lambda(V_i) - \mathbb{E}[\xi_\lambda(V_i)])\xi_\lambda(x)\right)^2 \\ &\leq \frac{1}{n}\sum_{i=n+1}^{2n}\left(\sum_{\lambda\in I_{m_0}} \xi^2_\lambda(x)\right)\left(\sum_{\lambda\in I_{m_0}} (\xi_\lambda(V_i) - \mathbb{E}[\xi_\lambda(V_i)])^2\right). \end{aligned}$$

Then, by Assumption $\mathbf{H_{mod}}$,

$$\left(\frac{1}{n}\sum_{i=n+1}^{2n} \varphi_x(V_i)\right) \leq K^2 p_0 \frac{1}{n}\sum_{i=n+1}^{2n}\left(\sum_{\lambda\in I_{m_0}} (\xi_\lambda(V_i) - \mathbb{E}[\xi_\lambda(V_i)])^2\right).$$

Hence,

$$\begin{aligned}
\left(\mathbb{E}\left[\left|\sup_{x\in I\cap\mathbb{Q}}\frac{1}{n}\sum_{i=n+1}^{2n}\varphi_x(V_i)\right|\right]\right)^2 &\leq \mathbb{E}\left[\left(\sup_{x\in I\cap\mathbb{Q}}\frac{1}{n}\sum_{i=n+1}^{2n}\varphi_x(V_i)\right)^2\right]\\
&\leq K^2p_0\sum_{\lambda\in I_{m_0}}\mathbb{E}\left[\frac{1}{n}\sum_{i=n+1}^{2n}(\xi_\lambda(V_1)-\mathbb{E}[\xi_\lambda(V_1)])^2\right]\\
&= \frac{K^2p_0}{n}\mathbb{E}\left[\sum_{\lambda\in I_{m_0}}(\xi_\lambda(V_1)-\mathbb{E}[\xi_\lambda(V_1)])^2\right]\\
&= \frac{K^2p_0}{n}\sum_{\lambda\in I_{m_0}}Var(\xi_\lambda(V_1))\\
&\leq \frac{K^2p_0}{n}\mathbb{E}\left[\sum_{\lambda\in I_{m_0}}\xi_\lambda^2(V_1))\right]\leq\frac{K^4p_0^2}{n}.
\end{aligned}$$

Thus,

$$\mathbb{E}\left[\left|\sup_{x\in I\cap\mathbb{Q}}\frac{1}{n}\sum_{i=n+1}^{2n}\varphi_x(V_i)\right|\right]\leq\frac{K^2p_0}{\sqrt{n}}=\mathbb{H}.$$

Let us compute the terms v and c involved in Talagrand Inequality. For every $x\in I$,

$$Var\left(\sum_{\lambda\in I_{m_0}}\xi_\lambda(V_1)\xi_\lambda(x)\right)\leq\mathbb{E}\left[\left(\sum_{\lambda\in I_{m_0}}\xi_\lambda(V_1)\xi_\lambda(x)\right)^2\right]=\int_I\left(\sum_{\lambda\in I_{m_0}}\xi_\lambda(u)\xi_\lambda(x)\right)^2g(u)du$$
$$\leq\nu\int_I\left(\sum_{\lambda\in I_{m_0}}\xi_\lambda(u)\xi_\lambda(x)\right)^2du=\nu\sum_{\lambda,\lambda'\in I_{m_0}}\left(\int_I\xi_\lambda(u)\xi_{\lambda'}(u)du\right)\xi_\lambda(x)\xi_{\lambda'}(x).$$

The family $\{\xi_\lambda,\lambda\in I_{m_0}\}$ is orthonormal, so

$$Var\left(\sum_{\lambda\in I_{m_0}}\xi_\lambda(V_1)\xi_\lambda(x)\right)\leq\nu\sum_{\lambda\in I_{m_0}}\xi_\lambda^2(x)\leq\nu K^2p_0=v.$$

Besides,

$$\left\|\sum_{\lambda\in I_{m_0}}\xi_\lambda(x)\xi_\lambda\right\|_\infty\leq\sqrt{\sum_{\lambda\in I_{m_0}}\xi_\lambda^2(x)}\times\left\|\sqrt{\sum_{\lambda\in I_{m_0}}\xi_\lambda^2}\right\|_\infty\leq K^2p_0=b.$$

Moreover, Assumption ($\mathbf{A_2}$) yields

$$P\left[\nu \leq \frac{\widehat{\nu}_n}{2}\right] \leq P\left[Z \geq \frac{\nu}{6}\right] = P\left[Z \geq \mathbb{H} + \left(\frac{\nu}{6} - \frac{K^2 p_0}{\sqrt{n}}\right)\right] \leq P\left[Z \geq \mathbb{H} + \frac{\nu}{12}\right].$$

Finally, Talagrand Inequality (Theorem 1.2.2) provides the following upper bound:

$$P\left[Z \geq \mathbb{H} + \frac{1}{12}\nu\right] \leq \exp\left(-\frac{n(\nu/12)^2}{2(\nu K^2 p_0 + 4(K^2 p_0)^2/\sqrt{n} + 3K^2 p_0(\nu/12))}\right).$$

Applying once again Assumption ($\mathbf{A_2}$), we get

$$\begin{aligned}
&\exp\left(-\frac{n(\nu/12)^2}{2(\nu K^2 p_0 + 4(K^2 p_0)^2/\sqrt{n} + 3K^2 p_0(\nu/12))}\right) \\
&\leq \exp\left(-\frac{n(\nu/12)^2}{2(\nu K^2 p_0 + 4K^2 p_0(\nu/12) + 3K^2 p_0(\nu/12))}\right) = \exp\left(-\frac{n\nu}{456K^2 p_0}\right) \quad \square
\end{aligned}$$

3.7 Proof of Theorem 3.4.1

The proof is based on the decomposition (3.22).

3.7.1 Upper bound of $\mathbb{E}[(f - f^-)^2(x_0)]$

The following Proposition holds.

Proposition 3.7.1 *Suppose that f is Lipschitz, then*

$$\mathbb{E}[(f^- - f)^2(x_0)] \leq Lip(f)^2 \mathbb{E}[\|b - \widehat{b}\|_{f_X}^2]. \tag{3.53}$$

Indeed, for every Z^-

$$\begin{aligned}
(f - f^-)^2(x_0) &= \Big(\int_0^1 [f(x_0) - f(x_0 - (b - \widehat{b})(x))] f_X(x)dx\Big)^2 \\
&\leq \int_0^1 [f(x_0) - f(x_0 - (b - \widehat{b})(x))]^2 f_X(x)dx \\
&\leq Lip(f) \int_0^1 [(b - \widehat{b})(x)]^2 f_X(x)dx \\
&= Lip(f)\|b - \widehat{b}\|_{f_X}^2
\end{aligned}$$

and by considering the expectation of the above inequality, we get the result of Proposition 3.7.1. $\square$

3.7.2 Upper bound of $\mathbb{E}[(\widehat{f}^-_{\widehat{m}} - f^-)^2(x_0)]$

Now, the term $\mathbb{E}[(\widehat{f}^-_{\widehat{m}} - f^-)^2(x_0)]$ in (3.22) is upper bounded with the results of Section 3.3. By Proposition 3.4.1, under the assumptions of Theorem 3.4.1, for every fixed sequence Z^-, f^- satisfies the assumptions of Theorem 3.3.1. Indeed, let Z^- be fixed, and suppose that Assumption $\mathbf{H}^{(\mathbf{1})}_{\mathbf{bias-error}}(\beta)$ holds, then $f \in \mathcal{H}(\beta, L)$, and by Proposition 3.4.1, $f^- \in \mathcal{H}(\beta, L)$. Besides, for every $t \in \mathcal{H}(\beta, L)$, $\|t - t_m\|_\infty \leq L D_m^{-\beta}$ thus $\|(f^-)_m - f^-\|_\infty \leq L D_m^{-\beta}$ and f^- satisfies Assumption $\mathbf{H}_{\mathbf{bias}}(\beta)$. The same argument holds with Assumption $\mathbf{H}^{(\mathbf{2})}_{\mathbf{bias-error}}(\beta)$. Similarly, if f satisfies Assumption $\mathbf{H}^{(\mathbf{1})}_{\nu\mathbf{-error}}(\beta)$ or $\mathbf{H}^{(\mathbf{2})}_{\nu\mathbf{-error}}(\beta)$, then f^- satisfies $\mathbf{H}_\nu$. Thus the following result holds.

Proposition 3.7.2 *Suppose that Assumption $\mathbf{H}^{(\mathbf{1})}_{\mathbf{bias-error}}(\beta)$ or $\mathbf{H}^{(\mathbf{2})}_{\mathbf{bias-error}}(\beta)$ holds for some $\beta \geq \beta' > 3/4$. In the definition (3.27) of $\widehat{\nu}^-_n$, consider an integer p_0 such that*

$$\left(\frac{n}{\log n}\right)^\gamma \leq p_0 \leq M \left(\frac{n}{\log n}\right)^\gamma$$

for some

$$\gamma \in \left]\frac{1}{\beta'(2\beta'+1)}, \min\left\{\frac{1}{\beta'+1}, \frac{4\beta'+1}{3(2\beta'+1)}\right\}\right[. \tag{3.54}$$

Then

$$\mathbb{E}[(\widehat{f}^-_{\widehat{m}} - f^-)^2(x_0)] \leq \theta'_1 \left(\frac{n}{\log n}\right)^{-2\beta/(2\beta+1)} + \mathcal{C}'_n \mathbb{E}[\|\widehat{b} - b\|^2_{f_X}] + \mathcal{R}_n$$

with

$$\theta'_1 = (\kappa_2 + 2\nu\kappa_3)(M+1)$$

$$\mathcal{C}'_n = 2\log n \left[\left(\nu\left(\frac{n}{\log n}\right)^{-\frac{1}{2\beta'+1}} + K^2 M\right)^2 \left(36 C_0^2 \left(\frac{n}{\log n}\right)^{\frac{2}{2\beta'+1}-2\beta'\gamma} + (18MK^2)^2 \left(\frac{n}{\log n}\right)^{\frac{2-4\beta'}{2\beta'+1}}\right)\right.$$
$$\left. + (12K^3)^2 \left(\frac{n}{\log n}\right)^{3\gamma - \frac{4\beta'+1}{2\beta'+1}}\right]$$

$$\mathcal{R}_n = 2\left[\nu + K^2 M \left(\frac{n}{\log n}\right)^{1/(2\beta'+1)}\right]^2 \exp\left(-\frac{C_0}{14K^2} n^{1-\gamma(1+\beta')}\right)$$
$$+ 2\kappa_3 K^2 (M+1) p_0 \left(\frac{n}{\log n}\right)^{-\frac{2\beta}{2\beta+1}} \exp\left(-\sqrt{n}38\right) + \frac{\theta_2}{n}.$$

Moreover, $\lim_{n\to+\infty} \mathcal{C}'_n = 0$ and $\mathcal{R}_n \leq \kappa'_1/n$ for some constant κ'_1 which depends on (M, K, β', ν).

Let us define the following sets, which depend on the sequence Z^-.

$$A_1^- = \left\{ C_0 p_0^{-\beta} \leq \frac{\nu^-}{6} \right\}, \quad A_2^- = \left\{ 12K^2 \frac{p_0}{\sqrt{n}} \leq \nu^- \right\}, \quad A_3^- = \left\{ \frac{18MK^2}{\nu^-} \leq \left(\frac{n}{\log n} \right)^{\frac{2\beta}{2\beta+1}} \right\}.$$

where $\nu^- = \|f^-\|_\infty$. The proof of Proposition 3.7.2 comes out of the following decomposition:

$$\begin{aligned}
&\mathbb{E}[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0)] = \mathbb{E}\left[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0) \mathbb{1}\!\mathrm{I}_{A_1^- \cap A_3^-}\right] + \mathbb{E}\left[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0) \mathbb{1}\!\mathrm{I}_{(A_1^-)^c \cup (A_3^-)^c}\right] \\
&\leq \mathbb{E}\left[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0) \mathbb{1}\!\mathrm{I}_{\{\widehat{\nu}_n^- \geq \nu^-/2\} \cap A_3^-}\right] + \mathbb{E}\left[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0) \mathbb{1}\!\mathrm{I}_{\{\widehat{\nu}_n^- < \nu^-/2\} \cap A_1^-}\right] \\
&\quad + \mathbb{E}\left[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0) \mathbb{1}\!\mathrm{I}_{\ (A_1^-)^c \cup (A_3^-)^c}\right] \qquad (3.55)
\end{aligned}$$

Then, these Claims provide an upper bound for each term in the right side of (3.55). There exists an integer n_0 which depends on (σ^2, β) such that for every $n \geq n_0$, the following results hold:

Claim 3.5

$$\begin{aligned}
&\mathbb{E}\left[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0) \mathbb{1}\!\mathrm{I}_{(A_1^-)^c \cup (A_3^-)^c}\right] \leq \\
&2\log n \left(\nu + K^2 M \left(\frac{n}{\log n} \right)^{\frac{1}{2\beta'+1}} \right)^2 \times \left[\left(\frac{6C_0}{p_0^\beta} \right)^2 + (18MK^2)^2 \left(\frac{n}{\log n} \right)^{-\frac{4\beta}{2\beta+1}} \right] \mathbb{E}[\|\widehat{b} - b\|_{f_X}^2].
\end{aligned}$$

Claim 3.6

$$\mathbb{E}\left[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0) \mathbb{1}\!\mathrm{I}_{\{\widehat{\nu}_n^- < \nu^-/2\} \cap A_1^-}\right] \leq 2 \left(\nu + MK^2 \left(\frac{n}{\log n} \right)^{\frac{1}{2\beta'+1}} \right)^2 \exp\left(-\frac{C_0}{14K^2} \frac{n}{p_0^{1+\beta}} \right).$$

Claim 3.7

$$\begin{aligned}
&\mathbb{E}\left[(\widehat{f}_{\widehat{m}}^- - f^-)^2(x_0) \mathbb{1}\!\mathrm{I}_{\{\widehat{\nu}_n^- \geq \nu^-/2\} \cap A_3^-}\right] \\
&\leq \left\{ \kappa_2 + \kappa_3 \left(2\nu + 2K^2 p_0 \exp\left(-\frac{\sqrt{n}}{38} \right) + 2(12K^3)^2 \log n \frac{p_0^3}{n} \mathbb{E}[\|\widehat{b} - b\|_{f_X}^2] \right) \right\} \\
&\times (M+1) \left(\frac{n}{\log n} \right)^{-\frac{2\beta}{2\beta+1}} + \frac{\theta_2}{n}.
\end{aligned}$$

These Claims provide the proof of Proposition 3.7.2. Indeed,

$$\mathbb{E}[(\widehat{f}^-_{\widehat{m}} - f^-)^2(x_0)] \leq (\kappa_2 + 2\nu\kappa_3)(M+1)\left(\frac{n}{\log n}\right)^{-\frac{2\beta}{2\beta+1}}$$

$$+2\log n\left[\left(\nu + K^2 M\left(\frac{n}{\log n}\right)^{\frac{1}{2\beta'+1}}\right)^2\left(\frac{36C_0^2}{p_0^{2\beta}} + (18MK^2)^2\left(\frac{n}{\log n}\right)^{-\frac{4\beta}{2\beta+1}}\right)\right.$$

$$\left.+\kappa_3(12K^3)^2\frac{p_0^3}{n}(M+1)\left(\frac{n}{\log n}\right)^{-\frac{2\beta}{2\beta+1}}\right]\mathbb{E}[\|\widehat{b}-b\|^2_{f_X}] + 2\left(\nu + K^2M\left(\frac{n}{\log n}\right)^{\frac{1}{2\beta'+1}}\right)^2\exp\left(-\frac{\sqrt{n}}{7}\right)$$

$$+2\kappa_3K^2(M+1)p_0\left(\frac{n}{\log n}\right)^{-\frac{2\beta}{2\beta+1}}\exp\left(-\sqrt{n}38\right) + \frac{\theta_2}{n}.$$

By the conditions $\beta \geq \beta'$ and $(n/\log)^\gamma \leq p_0 \leq M(n/\log n)^\gamma$, we have

$$\left(\nu + K^2M\left(\frac{n}{\log n}\right)^{\frac{1}{2\beta'+1}}\right)^2\left(\frac{36C_0^2}{p_0^{2\beta}} + (18MK^2)^2\left(\frac{n}{\log n}\right)^{-\frac{4\beta}{2\beta+1}}\right) + (12K^3)^2\frac{p_0^3}{n}(M+1)\left(\frac{n}{\log n}\right)^{-\frac{2\beta}{2\beta+1}}$$

$$\leq \left(\nu\left(\frac{n}{\log n}\right)^{-\frac{1}{2\beta'+1}} + K^2M\right)^2\left(36C_0^2\left(\frac{n}{\log n}\right)^{\frac{2}{2\beta'+1}-2\beta'\gamma} + (18MK^2)^2\left(\frac{n}{\log n}\right)^{\frac{2-4\beta'}{2\beta'+1}}\right)$$

$$+(12K^3)^2\left(\frac{n}{\log n}\right)^{3\gamma-\frac{4\beta'+1}{2\beta'+1}} = \mathcal{C}'_n.$$

According to Assumption (3.54), $2/(2\beta'+1) - 2\beta'\gamma < 0$ and $3\gamma - (4\beta'+1)/(2\beta'+1) < 0$, so $\lim_{n\to+\infty}\mathcal{C}'_n = 0$. Hence

$$\exp\left(-\frac{C_0}{14K^2}\frac{n}{p_0^{1+\beta'}}\right) = \exp\left(-\frac{C_0}{14K^2}n^{1-\gamma(1+\beta')}\right)$$

and $1-\gamma(1+\beta') > 0$ which entails that $\mathcal{R}_n \leq \kappa'_1$ for some constant κ'_1.

Let us prove these Claims. First of all, the probabilities $P[(A_1^-)^c]$, $P[(A_2^-)^c]$ and $P[(A_3^-)^c]$ are upper bounded via the following Lemma.

Lemma 3.7.1 *Let us consider a sequence (α_n) of positive number such that $\alpha_n = o(1/\sqrt{\log n})$. Then for every $n \in \mathbb{N}$ such that*

$$(C) \quad \frac{2}{\sqrt{\log n}} + \sigma^2\alpha_n^2\log n \leq \frac{1}{2}$$

where $\sigma^2 = \mathbb{E}[\epsilon_1^2]$,

$$P[\nu^- \leq \alpha_n] \leq 2\log n\alpha_n^2\mathbb{E}[\|\widehat{b}-b\|^2_{f_X}].$$

Hence there exists an integer n_0 which depends on $(\sigma^2, \beta, C_0, K)$ such that, for every $n \geq n_0$,

$$P[(A_1^-)^c] = P[\nu^- < 6C_0 p_0^{-\beta}] \leq 2\log n \left(\frac{6C_0}{p_0^\beta}\right)^2 \mathbb{E}[\|\widehat{b} - b\|_{f_X}^2] \tag{3.56}$$

$$P[(A_2^-)^c] = P\left[\nu^- \leq 12K^2 \frac{p_0}{\sqrt{n}}\right] \leq 2\log n (12K^2)^2 \frac{p_0^2}{n} \mathbb{E}[\|\widehat{b} - b\|_{f_X}^2] \tag{3.57}$$

and

$$\begin{aligned} P[(A_3^-)^c] &= P\left[\nu^- < 18MK^2 \left(\frac{n}{\log n}\right)^{-\frac{2\beta}{2\beta+1}}\right] \\ &\leq 2\log n (18MK^2)^2 \left(\frac{n}{\log n}\right)^{-\frac{4\beta}{2\beta+1}} \mathbb{E}[\|\widehat{b} - b\|_{f_X}^2]. \end{aligned} \tag{3.58}$$

Proof of Lemma 3.7.1

Given Z^-, ϵ_1 and $(b - \widehat{b})(X_1)$ are independent so

$$\mathbb{E}[\widehat{\epsilon}_1^2 | Z^-] = \mathbb{E}[\epsilon_1^2 | Z^-] + \mathbb{E}[(b - \widehat{b})^2(X_1) | Z^-] + 2\mathbb{E}[\epsilon_1 (b - \widehat{b})(X_1) | Z^-].$$

Moreover, $\mathbb{E}[\epsilon_1 | Z^-] = 0$, thus

$$\mathbb{E}[\widehat{\epsilon}_1^2 | Z^-] = \sigma^2 + \|b - \widehat{b}\|_{f_X}^2.$$

Then for every $A_n > 0$,

$$\int_{|y| > A_n} f^-(y)dy \leq \frac{1}{A_n^2} \int_{|y| > A_n} y^2 f^-(y)dy \leq \frac{1}{A_n^2}(\sigma^2 + \|b - \widehat{b}\|_{f_X}^2)$$

which entails

$$\int_{|y| \leq A_n} f^-(y)dy \geq 1 - \frac{\sigma^2 + \|b - \widehat{b}\|_{f_X}^2}{A_n^2}.$$

On the other hand, $\int_{|y| \leq A_n} f^-(y)dy \leq 2\nu^- A_n$, by definition of ν^-. Hence,

$$\nu^- \geq \frac{1}{2A_n}\left(1 - \frac{\sigma^2 + \|b - \widehat{b}\|_{f_X}^2}{A_n^2}\right)$$

for every $A_n > 0$. Thus,

$$\begin{aligned} P[\nu^- \leq \alpha_n] &\leq P\left[1 - \frac{\sigma^2 + \|b - \widehat{b}\|_{f_X}^2}{A_n^2} \leq 2A_n\alpha_n\right] \\ &= P\left[1 - (2A_n\alpha_n + \frac{\sigma^2}{A_n^2}) \leq \frac{\|b - \widehat{b}\|_{f_X}^2}{A_n^2}\right]. \end{aligned}$$

Let us consider $A_n = 1/(\alpha_n\sqrt{\log n})$, then condition (C) gives:

$$\begin{aligned} P[\nu^- \leq \alpha_n] &\leq P\left[1 - (\frac{2}{\sqrt{\log n}} + \sigma^2\alpha_n^2 \log n) \leq \|b - \widehat{b}\|_{f_X}^2 \log n\alpha_n^2\right] \\ &\leq P\left[\frac{1}{2} \leq \|b - \widehat{b}\|_{f_X}^2 \log n\alpha_n^2\right] \leq 2\log n\alpha_n^2 \mathbb{E}[\|b - \widehat{b}\|_{f_X}^2]. \quad \square \end{aligned}$$

Proof of Claim 3.5

According to Claim 3.4 in Section 3.6,

$$(\widehat{f}^-_{\widehat{m}} - f^-)^2(x_0) \leq (\nu^- + K^2 \max_{m=1,\ldots,N_n} D_m)^2 \quad \text{a.s.}$$

and $\nu^- \leq \nu$. Besides, by assumption, $\max_{m=1,\ldots,N_n} D_m \leq M(n/\log n)^{1/2\beta'+1}$. Hence

$$\mathbb{E}\left[(\widehat{f}^-_{\widehat{m}} - f^-)^2(x_0)\mathbb{1}_{(A_1^-)^c \cup (A_3^-)^c}\right] \leq \left(\nu + K^2 M \left(\frac{n}{\log n}\right)^{\frac{1}{2\beta'+1}}\right)^2 (P[(A_1^-)^c] + P[(A_3^-)^c])$$

and inequalities (3.56) and (3.58) end the proof of Claim 3.5. $\square$

Proof of Claim 3.6

For every Z^-,

$$\begin{aligned} \mathbb{E}\left[(\widehat{f}^-_{\widehat{m}} - f^-)^2(x_0)\mathbb{1}_{\{\widehat{\nu}_n^- < \nu^-/2\}}|Z^-\right]\mathbb{1}_{A_1^-} &\leq \left(\nu^- + K^2 \max_{m=1,\ldots,N_n} D_m\right)^2 P\left[\widehat{\nu}_n^- < \frac{\nu^-}{2}|Z^-\right]\mathbb{1}_{A_1^-} \\ &\leq 2\left(\nu + MK^2\left(\frac{n}{\log n}\right)^{\frac{1}{2\beta'+1}}\right)^2 \exp\left(-\frac{n\nu^-}{84K^2p_0}\right)\mathbb{1}_{A_1^-} \\ &\leq 2\left(\nu + MK^2\left(\frac{n}{\log n}\right)^{\frac{1}{2\beta'+1}}\right)^2 \exp\left(-\frac{C_0}{14K^2}\frac{n}{p_0^{1+\beta}}\right) \end{aligned}$$

since $\nu^- \geq 6C_0 p_0^{-\beta}$ on A_1^-. $\square$

Proof of Claim 3.7

According to Claim 3.3 in Section 3.6,

$$\begin{aligned}
&\mathbb{E}\left[(\widehat{f}^-_{\widehat{m}} - f^-)^2(x_0)\mathbb{1}_{\{\widehat{\nu}^-_n \geq \nu^-/2\}}|Z^-\right]\mathbb{1}_{A_3^-} \\
&\leq \mathbb{E}[\max(\kappa_2, \widehat{\nu}^-_n\kappa_3)|Z^-]\mathbb{1}_{A_3^-}(M+1)\left(\frac{n}{\log n}\right)^{-\frac{2\beta}{2\beta+1}} + \frac{\theta_2}{n} \\
&\leq (\kappa_2 + \kappa_3\mathbb{E}[\widehat{\nu}^-_n|Z^-])(M+1)\left(\frac{n}{\log n}\right)^{-\frac{2\beta}{2\beta+1}} + \frac{\theta_2}{n}
\end{aligned}$$

which entails that

$$\begin{aligned}
&\mathbb{E}\left[(\widehat{f}^-_{\widehat{m}} - f^-)^2(x_0)\mathbb{1}_{\{\widehat{\nu}^-_n \geq \nu^-/2\}\cap A_3^-}\right] \\
&\leq (\kappa_2 + \kappa_3\mathbb{E}[\widehat{\nu}^-_n])(M+1)\left(\frac{n}{\log n}\right)^{-\frac{2\beta}{2\beta+1}} + \frac{\theta_2}{n}. \qquad (3.59)
\end{aligned}$$

Besides,

$$\mathbb{E}[\widehat{\nu}^-_n|Z^-] \;\leq\; \mathbb{E}\left[\widehat{\nu}^-_n\mathbb{1}_{\widehat{\nu}^-_n \leq 2\nu^-}|Z^-\right] + \mathbb{E}\left[\widehat{\nu}^-_n\mathbb{1}_{\{\widehat{\nu}^-_n > 2\nu^-\}}|Z^-\right]\mathbb{1}_{A_2^-} + \mathbb{E}[\widehat{\nu}^-_n|Z^-]\mathbb{1}_{(A_2^-)^c}.$$

According to inequality (3.51), $\widehat{\nu}^-_n = \|\widehat{g}^{(1)}_{m_0}\|_\infty \leq K^2p_0$, thus

$$\mathbb{E}[\widehat{\nu}^-_n|Z^-] \;\leq\; 2\nu^- + K^2p_0\exp\left(-\frac{n\nu^-}{456K^2p_0}\right)\mathbb{1}_{(A_2^-)^c} + K^2p_0\mathbb{1}_{(A_2^-)^c}.$$

On A_2^-, $\exp(-n\nu^-/456K^2p_0) \leq \exp(-\sqrt{n}/38)$ a.s., so

$$\mathbb{E}[\widehat{\nu}^-_n] \leq 2\nu + K^2p_0\left(\exp\left(-\frac{\sqrt{n}}{38}\right) + P[(A_2^-)^c]\right)$$

and with (3.57),

$$\mathbb{E}[\widehat{\nu}^-_n] \leq 2\nu + K^2p_0\exp\left(-\frac{\sqrt{n}}{38}\right) + 2\log n(12K^3)^2\frac{p_0^3}{n}\mathbb{E}[\|\widehat{b} - b\|^2_{f_X}]. \qquad (3.60)$$

Then inequalities (3.59) and (3.60) provide the proof of Claim 3.7. □

3.8 Additionnal Proofs

3.8.1 Proof of Proposition 3.4.1

1) Let $x \in \mathbb{R}$,

$$|f^-(x)| \leq \int_0^1 |f(x-(b-\widehat{b})(y))| f_X(y)dy \leq \nu \int_0^1 f_X(y)dy = \nu \quad a.s.$$

2) Suppose that $f \in \mathcal{H}(\beta, L)$ and $\beta = r + \alpha$ with $\alpha \in]0,1]$. We have,

$$\begin{aligned}(f^-)^{(r)}(x) &= \frac{\partial^r}{\partial x^r}\left(\int_0^1 f(x-(b-\widehat{b})(y))f_X(y)dy\right) \\ &= \int_0^1 \frac{\partial^r}{\partial x^r}(f(x-(b-\widehat{b})(y)))f_X(y)dy \\ &= \int_0^1 f^{(r)}(x-(b-\widehat{b})(y))f_X(y)dy.\end{aligned}$$

Hence, for every $x, x' \in \mathbb{R}$,

$$\begin{aligned}|(f^-)^{(r)}(x) - (f^-)^{(r)}(x')| &\leq \int_0^1 |f^{(r)}(x-(b-\widehat{b})(y)) - f^{(r)}(x'-(b-\widehat{b})(y))| f_X(y)dy \\ &\leq \int_0^1 L|x-x'|^\alpha f_X(y)dy = L|x-y|^\alpha\end{aligned}$$

which proves that $f^- \in \mathcal{H}(\beta, L)$.

3) First of all, for every $u \in \mathbb{R}$, the Fourier transform of f^- is

$$(f^-)^*(u) = \int_{x\in\mathbb{R}} f^-(x)e^{-iux}dx = \int_{x\in\mathbb{R}}\int_{y=0}^1 f(x-(b-\widehat{b})(y))f_X(y)e^{-iux}dxdy.$$

Set $z = x - (b-\widehat{b})(y)$,

$$(f^-)^*(u) = \int_{y=0}^1 \int_{z\in\mathbb{R}} f(z)e^{-iuz}e^{-iu(b-\widehat{b})(y)}dz f_X(y)dy = f^*(u)\int_{y=0}^1 e^{-iu(b-\widehat{b})(y)}f_X(y)dy.$$

Hence,

$$|(f^-)^*(u)| \leq |f^*(u)| \int_{y=0}^1 |e^{-iu(b-\widehat{b})(y)}| f_X(y)dy = |f^*(u)|.$$

Then, if $f \in W(\beta, L)$,

$$\frac{1}{2\pi}\int_{u\in\mathbb{R}} |(f^-)^*(u)|^2 u^{2\beta+1}du \leq \frac{1}{2\pi}\int_{u\in\mathbb{R}} |f^*(u)|^2 u^{2\beta+1}du \leq L^2$$

so $f^- \in W(\beta, L)$. □

3.8.2 Proof of Proposition 3.2.1 and (1) in Proposition 3.3.2

Let us prove Proposition 3.2.1.

1) A simple calculus proves that the Fourier transform of $\mathbb{1}_{[-\pi,\pi]}$ is $2\pi\phi$, then for every $u \in \mathbb{R}$,

$$\phi^*(u) = \int_{\mathbb{R}} \phi(y) e^{-iuy} dy = \frac{1}{2\pi} \int_{\mathbb{R}} \mathbb{1}^*_{[-\pi,\pi]}(y) e^{iuy} dy = \frac{1}{2\pi} \mathbb{1}_{[-\pi,\pi]}(-u) = \frac{1}{2\pi} \mathbb{1}_{[-\pi,\pi]}(u).$$

Hence, for every (m,k),

$$\phi^*_{m,k}(u) = (1/\sqrt{m}) e^{-iku/m} \mathbb{1}_{[-\pi m,\pi m]}(u). \tag{3.61}$$

Then, let $m > 0$ and $k, l \in \mathbb{Z}$, according to the Parseval formula, we have

$$\langle \phi_{m,k}, \phi_{m,l} \rangle = \frac{1}{2\pi} \langle \phi^*_{m,k}, \phi^*_{m,l} \rangle = \frac{1}{2\pi m} \int_{-\pi m}^{\pi m} e^{-i(k-l)u/m} du = \mathbb{1}_{\{k=l\}}.$$

2) First of all, we recall that for every subset S_m of $L^2(I)$, the two following properties are equivalent (see Introduction, Section 1.2.5).

$$\Big(\|t\|_\infty \leq K\sqrt{D_m}\|t\|, \quad \forall t \in S_m \Big) \quad \Leftrightarrow \quad \| \sum_{k\in\mathbb{Z}} \phi^2_{m,k} \|_\infty \leq K^2 D_m. \tag{3.62}$$

So let $t \in S_m$ and $x \in \mathbb{R}$, we prove that $|t(x)| \leq \sqrt{m}\|t\|$. As $Supp(t^*) \subset [-\pi m, \pi m]$, with Parseval Equality

$$(t(x))^2 = \left(\frac{1}{2\pi} \int_{-\pi m}^{\pi m} t^*(u) e^{ixu} du \right)^2 \leq \frac{1}{(2\pi)^2} \left(\int_{-\pi m}^{\pi m} |t^*(u)|^2 du) \times 2\pi m \right) = m\|t\|^2$$

which proves (2) in Proposition 3.2.1.

3) By (3.61), it is obvious that $S_m \subset \{t \in L^2(\mathbb{R}), Supp(t^*) \subset [-\pi m, \pi m]\}$. Conversely, let $t \in L^2(\mathbb{R})$ be such that $Supp(t^*) \subset [-\pi m, \pi m]$, then t^* decomposes in Fourier series as

$$t^*(u) = \left(\sum_{k\in\mathbb{Z}} a_k e^{iku\pi/m} \right) \mathbb{1}_{[-\pi m,\pi m]} \in Vect\{\phi^*_{m,k}, k \in \mathbb{Z}\}$$

for some numbers $(a_k)_{k\in\mathbb{Z}}$, thus $t \in S_m$. Hence $S_m = \{t \in L^2(\mathbb{R}), Supp(t^*) \subset [-\pi m, \pi m]\}$. Then $S_m \subset S_{m'}$ for every $m \leq m'$.

Let us prove (1) in Proposition 3.3.2. For every $h \in L^2(\mathbb{R})$,

$$h_m = \arg\min_{t\in A_m} \|h-t\|^2 = \arg\min_{Supp(t^*)\subset[-\pi m,\pi m]} \frac{1}{2\pi}\|h^* - t^*\| = \frac{1}{2\pi}(h^* \mathbb{1}_{[-\pi m,\pi m]})^*.$$

Suppose that $h \in W(\beta+1/2, L)$, let $x \in \mathbb{R}$,

$$\begin{aligned}(h-h_m)^2(x) &= \left(\frac{1}{2\pi}\int_{\mathbb{R}}(h^*-h_m^*)(u)e^{iux}du\right)^2 = \left(\frac{1}{2\pi}\int_{|u|>\pi m} h^*(u)e^{iux}du\right)^2 \\ &\leq \frac{1}{(2\pi)^2}\int_{|u|>\pi m}|h^*(u)|^2|u|^{2\beta+1}du \times \int_{|u|>\pi m}\frac{1}{|u|^{2\beta+1}}du \\ &\leq \frac{L^2}{2\beta\pi^{2\beta+1}} \times m^{-2\beta}. \quad \square\end{aligned}$$

3.8.3 Proof of Proposition 3.2.2 and (2) in Proposition 3.3.2

Let us prove Proposition 3.2.2. For every $j \in \mathbb{N}$, $x \in \mathbb{R}$,

$$Card\left(\{k \in \Gamma(j) : \psi_{j,k}(x) \neq 0\}\right) \leq Card\left(\{k \in \mathbb{Z}, -B \leq 2^j x - k \leq B\}\right) \leq 2B+1$$

Thus, for every $m \in \mathbb{N}^*$, $t \in B_m$ and $x \in [-1,1]$, we have,

$$\begin{aligned}(t(x))^2 &= \left(\sum_{k\in\Gamma'(0)}\langle\varphi_k,t\rangle\varphi_k(x) + \sum_{j=0}^{m-1}\sum_{k\in\Gamma(j)}\langle\psi_{j,k},t\rangle\psi_{j,k}(x)\right)^2 \\ &\leq \left(\sum_{k\in\Gamma'(0)}\langle\varphi_k,t\rangle^2 + \sum_{j=0}^{m-1}\sum_{k\in\Gamma(j)}\langle\psi_{j,k},t\rangle^2\right) \times \left(\sum_{k\in\Gamma'(0)}\varphi_k^2(x) + \sum_{j=0}^{m-1}\sum_{k\in\Gamma(j)}\psi_{j,k}^2(x)\right) \\ &\leq \|t\|^2 \times (2B+1)\left(\|\varphi\|_\infty^2 + \sum_{j=0}^{m-1}2^j\|\psi\|_\infty^2\right) \\ &\leq K^2\|t\|^2 2^m\end{aligned}$$

where K depends only on the structure of the mother and father wavelets. According to (3.62), this proves the result of Proposition 3.2.2.

• Assertion (2) in Proposition 3.3.2 comes from Meyer (1990) (Section 9, chapter 2, Proposition 4). The result stated by Meyer is more general (for Besov spaces and a L^q-norm), and we only recall it in the form we require. Let $h \in L^2(\mathbb{R})$, then

$$h(x) = \sum_{k\in\Gamma'(0)}\langle h,\varphi_k\rangle\varphi_k + \sum_{j\geq 0}\sum_{k\in\Gamma(j)}\langle h,\psi_{j,k}\rangle\psi_{j,k}$$

and

$$h_m(x) = \sum_{k\in\Gamma'(0)}\langle h,\varphi_k\rangle\varphi_k + \sum_{j=0}^{m-1}\sum_{k\in\Gamma(j)}\langle h,\psi_{j,k}\rangle\psi_{j,k}.$$

On the one hand, if $h \in \mathcal{H}(\beta, L) = \mathcal{B}_{\infty}^{\beta,\infty}(L)$

$$\sup_{j\geq 0} 2^{j\beta} \left\| \sum_{k\in\Gamma(j)} \langle h, \psi_{j,k}\rangle \psi_{j,k} \right\|_{\infty} = |||h||| < +\infty.$$

Moreover, there exists a constant which only depends of ψ and φ such that $|||h||| \leq CL$ for every $h \in \mathcal{H}(\beta, L)$. Thus, for every $m \geq 1$,

$$\begin{aligned} \|h - h_m\|_{\infty} &= \| \sum_{j\geq m} \sum_{k\in\Gamma(j)} \langle \psi_{j,k}, h\rangle \psi_{j,k} \|_{\infty} \\ &\leq \sum_{j\geq m} C|||h|||2^{-j\beta} \\ &\leq CL \frac{2^{-m\beta}}{1-2^{-\beta}} \\ &= \frac{K'(\beta)L}{2^{m\beta}}. \quad \square \end{aligned}$$

Deuxième partie

Estimation à partir de données censurées

Chapitre 4

Estimation du risque instantané à partir de données censurées à droite

Ce chapitre est consacré à l'estimation du taux de risque instantané $h(x) = f_Y(x)/\overline{F}_Y(x)$ d'une variable positive Y, où f_Y et $\overline{F}_Y$ désignent la densité et la fonction de survie de Y, en présence de censure à droite : on suppose qu'il existe une variable C indépendante de Y telle que l'observation se limite au couple

$$\left(\min(Y, C), \mathbb{1}_{\{Y \leq C\}}\right).$$

Nous proposons dans ce chapitre une méthode d'estimation originale basée sur la minimisation d'un contraste de type régression. Une procédure de sélection de modèle produit ensuite un estimateur adaptatif. Enfin, nous illustrons les performances de l'estimateur de sélection de modèle sur des simulations et comparons les valeurs du risque avec celles fournies par d'autres estimateurs adaptatifs présents dans la littérature

Ce chapitre est une version legèrement modifiée de l'article Plancade (to appear), à paraître dans Metrika.

4.1 Introduction

In medical follow-up and other subjects, the observation of a variable of interest, for example the lifetime of an individual, can be right censored. This means that we only observe the minimum of the lifetime and a variable called censoring time (for example the time when a patient leaves the medical program), which is supposed independent of the lifetime. We also observe if this minimum corresponds to the variable of interest or to the censoring time. More precisely, we consider a sample $(Y_i)_{i=1,\dots,n}$ of non-negative variables, and a sample $(C_i)_{i=1,\dots,n}$ of non-negative censoring times. Then we observe a sample $(T_i, \delta_i)_{i=1,\dots n}$ with

$$T_i = \min(Y_i, C_i), \quad \delta_i = \mathbb{1}_{\{Y_i \leq C_i\}}. \tag{4.1}$$

A function of interest in such a study is the hazard rate function of Y which represents the risk of death at a time x knowing that the patient is alive until x. If we denote by $f_Y(x)$ and $\overline{F}_Y(x) = P[Y_1 \geq x]$ the density and the survival function of Y, the hazard rate function is

$$h(x) = \frac{f_Y(x)}{\overline{F}_Y(x)}. \tag{4.2}$$

A lot of papers are devoted to hazard rate estimation. In particular, it forms part of the most general study of counting processes (see Andersen et al. (1993)). Two general methods can be drawn in the non parametric context that we only consider.

The first one consists in estimating h by a ratio of two estimators. The most obvious is $\widehat{f_Y}/\widehat{F}_Y$ where $\widehat{f_Y}$ and $\widehat{F}_Y$ are estimators of f_Y and $\overline{F}_Y$. In general, $\overline{F}_Y$ is replaced by the well known Kaplan Meier estimator of $\overline{F}_Y$ (Kaplan and Meier (1958)). Another decomposition of h is

$$h = \frac{f_Y \overline{F}_C}{\overline{F}_T}. \tag{4.3}$$

Indeed, $\overline{F}_T(x) = P[\{Y \geq x\} \cap \{C \geq x\}] = \overline{F}_Y(x)\overline{F}_C(x)$. The function $\psi(x) = f_Y(x)\overline{F}_C(x)$, called the subdensity of Y, corresponds heuristically to the "density" of the observed variables Y_i, in the sense that for every function $t : \mathbb{R}^+ \to \mathbb{R}$ such that $t(0) = 0$,

$$\mathbb{E}[t(\delta_i T_i)] = \mathbb{E}[\delta_i t(Y_i)] = \int t(x)\psi(x)dx.$$

As the (δ_i, Y_i) are directly measured, ψ is easier to estimate than f_Y. Similarly, $\overline{F}_T$ is easier to estimate than $\overline{F}_Y$. Indeed it can simply be replaced by the empirical survival function of the observed (T_i). Patil (1993) proposes a kernel estimator of ψ with a bandwidth selection and gets an estimator of h via (4.3). Antoniadis et al. (1999) use a wavelet decomposition but their estimator is not really adaptive as the optimal resolution of the wavelets depends on the regularity of f_Y. Comte and Brunel (2005) build a projection estimator of ψ by model selection in more general bases, and obtain an adaptive estimator.

Other estimators of h are based on the cumulative hazard $H(x) = -\log(\overline{F}_Y(x))$. One of the most frequently used estimator of H is the Nelson-Aalen estimator (Nelson (1972)). Obviously, we have

$$h(x) = H'(x). \tag{4.4}$$

Thus Yandell (1983) and Tanner and Wong (1983) build an estimator of h by differentiating the Nelson-Aalen estimator of H with a delta-sequence method, and Muller and Wang (1994) introduce a variable bandwidth. Nielsen (2003) compares the numerical results from several variable bandwidth kernel estimators, and one of them is developed in Bagkavos and Patil (2009). Brunel and Comte (2008) propose a projection type estimator based on a approximation of cumulative hazard function. The method is very different from the one presented here, but leads also to an adaptive estimation procedure.

Let us mention also the estimator of Reynaud-Bouret (2006) built by model selection in a set of random models, which is adaptive on Hölder spaces with regularity smaller than 1.

The present chapter describes a regression type strategy, in a different spirit from other procedures. It leads to an adaptive estimator for the integrated squared risk on a set $[0, a]$ such that $P(T \geq a)$ is positive. The proofs are self contained (apart from Talagrand Inequality), and the key point is that the reference norm for the risk is chosen to be well suited to the problem.

The plan of the chapter is the following. Section 4.2 presents the framework, and the main assumptions. The estimation procedure is described in Section 4.3, as well as the main result. But the estimator built in Section 4.3 brings into play unknown quantities, which are estimated in Section 4.4. The performance of these estimators on simulated data are presented in Section 4.5. The proofs are gathered in Sections 4.6 and 4.7. Section 4.8 presents a technical algebra lemma.

4.2 Presentation of the framework, assumptions and notations

4.2.1 Framework

We consider a sample $(Y_1, \ldots, Y_n)$ of i.i.d. non negative random variables with common survival function $\overline{F}_Y(x) = P[Y_1 \geq x]$ and density f_Y, and a sample $(C_1, \ldots, C_n)$ of i.i.d. non negative random variables with common survival function $\overline{F}_C$, independent of the (Y_i)'s. The variables of interest are the (Y_i)'s, but we only observe the sample $((T_1, \delta_1), \ldots, (T_n, \delta_n))$ defined in (4.1). The aim of this chapter is to build an estimator of the hazard rate of Y_1 given by (4.2), on a compact interval A on which $\overline{F}_T = \overline{F}_C \overline{F}_Y$ is lower bounded by a positive number, which is a classical assumption in such studies. Theoretically, A is a known compact interval independent of the data, even if practically it is chosen by looking at the data. More precisely, we consider the following assumption.

$\mathbf{A_{frame}}$: $\overline{F}_T$ is lower bounded by $\overline{F}_0 > 0$ on $A = [0, a]$ for some positive number a, and h is upper bounded on A by $\|h\|_{\infty,A} = \sup_{x\in A} h(x) < \infty$.

4.2.2 Notations

We define the following scalar products and norms on $L^2(A)$. For every $s, t \in L^2(A)$,

$$\langle s,t\rangle = \int_A s(x)t(x)dx, \quad \|t\|^2 = \int_A t^2(x)dx,$$
$$\langle s,t\rangle_{\overline{F}_T} = \int_A s(x)t(x)\overline{F}_T(x)dx, \quad \|t\|^2_{\overline{F}_T} = \int_A t^2(x)\overline{F}_T(x)dx,$$
$$\langle s,t\rangle_n = \frac{1}{n}\sum_{i=1}^n \int_A s(x)t(x)\mathbb{1}_{\{T_i\geq x\}}dx, \quad \|t\|^2_n = \frac{1}{n}\sum_{i=1}^n \int_A t^2(x)\mathbb{1}_{\{T_i\geq x\}}dx.$$

Let M be a matrix, we denote by M^t the transpose of M. If M is a square matrix, let $Sp(M)$ be the set of the eigenvalues of M.

Let β and L be positive numbers, and r the greatest integer smaller than β, we define the Hölder space $\mathcal{H}(\beta, L)$ on A,

$$\mathcal{H}(\beta, L) = \{f : A \to \mathbb{R}, |f^{(r)}(x) - f^{(r)}(y)| \leq L|x-y|^{\beta-r}, \forall x, y \in A\}.$$

For every $x \in \mathbb{R}$, we denote by $E(x)$ the greatest integer smaller than or equal to x. For every subset S of $\mathbb{R}$ we denote by S^c the complementary of S and by $\mathbb{1}_S$ the function which is equal to 1 on S and to 0 on S^c.

All throughout the chapter, C_i denotes a universal numerical constant, and C, C' denote constants which depend on the given parameters of the problem and may change from one line to another.

4.2.3 Collections of models

We consider a collection $\mathcal{M}_n = \{S_m, m \in J_n\}$ of finite dimensional linear subsets of $L^2(A)$ with dimension $D_m = dim(S_m) \leq n/\log^2 n$. We suppose that $\mathcal{M}_n$ satisfies either Assumption $\mathbf{A^{(1)}_{mod}}$ or $\mathbf{A^{(2)}_{mod}}$.

$\mathbf{A^{(1)}_{mod}}$: We suppose that $J_n = \{1, \dots, N_n\}$ and

$$S_1 \subset S_2 \subset \cdots \subset S_{N_n}.$$

Besides, there exists a constant K such that for every model S_m, and for every $(\phi^m_1, \dots, \phi^m_{D_m})$ orthonormal basis of S_m for the $L^2(A)$-norm,

$$\sup_{x\in A}\left|\sum_{k=1}^{D_m}(\phi^m_k(x))^2\right| \leq K^2 D_m. \tag{4.5}$$

Moreover, for every positive constant c, there exists a constant $C > 0$ such that, for every $n \in \mathbb{N}^*$,

$$\sum_{m \in J_n} \exp(-c\sqrt{D_m}) \leq C. \tag{4.6}$$

$\mathbf{A}^{(\mathbf{2})}_{\mathbf{mod}}$: Condition (4.6) is satisfied and there exists a linear subset S_n of $L^2(A)$ with dimension $N_n \leq n/\log^2 n$ such that for every $m \in J_n$ $S_m \subset S_n$, and the global space S_n satisfies (4.5).

Remark 8 *1. Assumption $\mathbf{A}^{(\mathbf{1})}_{\mathbf{mod}}$ is clearly stronger than Assumption $\mathbf{A}^{(\mathbf{2})}_{\mathbf{mod}}$. Thus $\mathbf{A}^{(\mathbf{2})}_{\mathbf{mod}}$ allows more irregular collections of models. For example, take $A = [0,1]$, let I_n be the regular partition of $[0,1]$ of step $1/N_n$, and S_n the set of histograms on $[0,1]$ which are constant on I_n. Under Assumption $\mathbf{A}^{(\mathbf{2})}_{\mathbf{mod}}$, the collection $\mathcal{M}_n$ can include any set of histograms S_m based on a partition I_m of $[0,1]$ composed of union of intervals from I_n, whereas Assumption $\mathbf{A}^{(\mathbf{1})}_{\mathbf{mod}}$ only allows diadic regular set of histograms, namely*

$$S_m = Vect\left\{ \mathbb{1}_{I_{[\frac{j-1}{D_m}\frac{j}{D_m})}}, j = 1, \dots D_m \right\}, \text{ where } D_m = 2^m, \quad m \in \mathbb{N}, \quad 2^m \leq N_n.$$

2. Condition (4.6) is equivalent to

$$\sum_{D=1}^{N_n} \exp(-c\sqrt{D})\, Card\left(\{S_m \in \mathcal{M}_n, D_m = D\}\right) \leq C,$$

which restricts the number of models S_m in $\mathcal{M}_n$ of dimension $D_m = D$ for every $D \leq N_n$.

3. Under Assumption $\mathbf{A}^{(\mathbf{1})}_{\mathbf{mod}}$, (4.6) is clearly satisfied since $Card\left(\{S_m \in \mathcal{M}_n, D_m = D\}\right) = 1$ for every $D \leq N_n$.

4.3 Theoretical estimators

The estimators built in this section bring into play unknown quantities, which are replaced by estimators in Section 4.4. In Section 4.3.1, we present a non adaptive procedure which provides an estimator $\widehat{h}_m$ of h on each model S_m. The model selection procedure is described in Section 4.3.2 , with two different penalties corresponding to the two Assumptions $\mathbf{A}^{(\mathbf{1})}_{\mathbf{mod}}$ and $\mathbf{A}^{(\mathbf{2})}_{\mathbf{mod}}$. Section 4.3.3 presents the main result.

4.3.1 Minimum contrast estimators

We note that

$$h\mathbb{1}_A = \arg\min_{t\in L^2(A)} \|t-h\|^2_{\overline{F}_T} = \arg\min_{t\in L^2(A)} \|t\|^2_{\overline{F}_T} - 2\langle t,h\rangle_{\overline{F}_T}.$$

Now we build an estimator of $\|t\|^2_{\overline{F}_T} - 2\langle t,h\rangle_{\overline{F}_T}$. For every $t \in L^2(A)$, $\mathbb{E}[\|t\|^2_n] = \|t\|^2_{\overline{F}_T}$. In addition, for every $i = 1,\dots,n$,

$$\begin{aligned}
\mathbb{E}[\delta_i t(T_i)] &= \mathbb{E}[\mathbb{E}[\delta_i t(T_i)|Y_i]] = \mathbb{E}[t(Y_i)\mathbb{E}[\mathbb{1}_{Y_i\le C_i}|Y_i]] \\
&= \mathbb{E}[t(Y_i)\overline{F}_C(Y_i)] = \int_A t(x)\overline{F}_C(x)f_Y(x)dx \\
&= \int_A t(x)\overline{F}_C(x)\overline{F}_Y(x)h(x)dx = \langle t,h\rangle_{\overline{F}_T}. \qquad (4.7)
\end{aligned}$$

Then, we set

$$\gamma_n(t) = \|t\|^2_n - \frac{2}{n}\sum_{i=1}^n \delta_i t(T_i)$$

for every $t \in L^2(A)$, and $\mathbb{E}[\gamma_n(t)] = \|t\|^2_{\overline{F}_T} - 2\langle t,h\rangle_{\overline{F}_T}$.

For every model S_m, let $\tilde{h}_m = \arg\min_{t\in S_m}\gamma_n(t)$. Let $(\phi^m_1,\dots,\phi^m_{D_m})$ be an $L^2(A)$-orthonormal basis of S_m, then $\tilde{h}_m = \sum_{k=1}^{D_m}\tilde{a}^m_k\phi^m_k$ where the $\{\tilde{a}^m_k\}$'s satisfy

$$\frac{\partial\gamma_n(\sum_{k=1}^{D_m}a^m_k\phi^m_k)}{\partial a^m_k} = 0, \quad \forall k = 1,\dots D_m \quad \Leftrightarrow \quad \widehat{G}_m\tilde{A}_m = \widehat{V}_m$$

with $\tilde{A}_m = (\tilde{a}^m_1,\dots,\tilde{a}^m_{D_m})^t$ and

$$\widehat{G}_m = (\langle\phi^m_k,\phi^m_{k'}\rangle_n)_{k,k'=1,\dots,D_m}, \quad \widehat{V}_m = \left(\frac{1}{n}\sum_{i=1}^n \delta_i\phi^m_k(T_i)\right)_{k=1,\dots,D_m}. \qquad (4.8)$$

We note that $\tilde{h}_m$ is uniquely defined iff $\widehat{G}_m$ is invertible. Thus we construct a set of high probability on which the spectrum of $\widehat{G}_m$ is lower bounded. First of all, by Lemma 4.8.1,

$$\min(Sp(\widehat{G}_m)) = \min_{\{U\in\mathbb{R}^{D_m},\ U^tU=1\}} U^t\widehat{G}_mU = \min_{\{t=\sum_{k=1}^{D_m}u_k\phi^m_k,\ \|t\|=1\}}\|t\|^2_n.$$

Besides, $\|.\|_n$ is the empirical norm associated with $\|.\|_{\overline{F}_T}$, therefore the set

$$\Delta_1 = \left\{\left|\frac{\|t\|^2_n}{\|t\|^2_{\overline{F}_T}} - 1\right| < \frac{1}{4}, \forall t\in S_n\right\} \qquad (4.9)$$

has a probability close to 1, which is proved in Proposition 4.6.3. On Δ_1, $\min(Sp(\widehat{G}_m)) = \min_{\{t\in S_m, \|t\|=1\}}(3/4)\|t\|^2_{\overline{F}_T} \geq (3/4)\overline{F}_0$ for every $m \in J_n$, so the set

$$\Delta_2^{th} = \left\{ \min(Sp(\widehat{G}_m)) \geq \frac{3}{4}\overline{F}_0,\ \forall m \in J_n \right\} \tag{4.10}$$

contains Δ_1. Thus $P[\Delta_2^{th}]$ is close to 1, and $\widehat{G}_m$ is invertible on Δ_2^{th}. Finally, we define the estimator $\widehat{h}_m = \sum_{k=1}^{D_m} \widehat{a}_k^m \phi_k^m$ with

$$\widehat{A}_m = (\widehat{a}_1^m, \ldots, \widehat{a}_{D_m}^m)^t = \begin{cases} \widehat{G}_m^{-1}\widehat{V}_m & \text{on } \Delta_2^{th} \\ 0 & \text{otherwise .} \end{cases}$$

4.3.2 Adaptive estimators

The non adaptive estimation procedure described in Section 4.3.1 provides a collection of estimators $\{\widehat{h}_m, m \in \mathcal{M}_n\}$, among which one is automatically selected by a penalised model selection procedure. By Pythagoras Theorem, for every model S_m the risk of the estimator $\widehat{h}_m$ splits in two terms,

$$\mathbb{E}\left[\|\widehat{h}_m - h\|^2_{\overline{F}_T}\right] = \|h - h_m\|^2_{\overline{F}_T} + \mathbb{E}\left[\|\widehat{h}_m - h_m\|^2_{\overline{F}_T}\right]$$

where h_m is the $\|.\|_{\overline{F}_T}$-projection of h on S_m. The bias term $\|h - h_m\|^2_{\overline{F}_T}$ decreases when the model S_m grows, whereas the term $\mathbb{E}[\|\widehat{h}_m - h_m\|^2_{\overline{F}_T}]$ has the order D_m/n of a variance-type term, and increases with D_m. (Nevertheless, in our case, it is not exactly a variance term, for $\mathbb{E}[\widehat{h}_m(x)] \neq h_m(x)$.) Thus, the best model would be the one which generates the smallest risk, i.e. the one which realises the better trade-off between bias and variance.

We construct a data driven quantity which has the same order as the bias-variance sum (up to a constant independent of m) and select the model which minimises this quantity. On the one hand,

$$\|h - h_m\|^2_{\overline{F}_T} = \|h_m\|^2_{\overline{F}_T} - 2\langle h, h_m\rangle_{\overline{F}_T} + \|h\|^2_{\overline{F}_T}.$$

The term $\|h_m\|^2_{\overline{F}_T} - 2\langle h, h_m\rangle_{\overline{F}_T}$ is estimated by $\gamma_n(\widehat{h}_m)$ (see (4.7)) and the term $\|h\|^2_{\overline{F}_T}$ is independent of m. On the other hand, the variance term $\mathbb{E}[\|\widehat{h}_m - h_m\|^2_{\overline{F}_T}]$ is upper bounded by a deterministic term with order D_m/n, called the penalty. We do not explicitly prove this result here but a more general one (see Theorem 4.3.1 and Comment 1 hereafter).

We consider two penalties with order D_m/n, but with different constants.

$$pen_1^{th}(m) = \frac{BK^2}{\overline{F}_0}\frac{D_m}{n}, \quad pen_2^{th}(m) = B\|h\|_{\infty,A}\frac{D_m}{n} \tag{4.11}$$

with $B > 3$, and select the model

$$\widehat{m}_j = \arg\min_{m \in J_n} \gamma_n(\widehat{h}_m) + pen_j^{th}(m) \tag{4.12}$$

for $j = 1$ or 2. We get two almost data-driven estimators of h: $\widehat{h}_{\widehat{m}_1}$ and $\widehat{h}_{\widehat{m}_2}$. Each penalty corresponds to a set of assumptions. Penalty pen_2^{th} corresponds to Assumption $\mathbf{A}_{\mathbf{mod}}^{(\mathbf{2})}$ so it works under both $\mathbf{A}_{\mathbf{mod}}^{(\mathbf{1})}$ and $\mathbf{A}_{\mathbf{mod}}^{(\mathbf{2})}$ (see Remark 8). Penalty pen_2^{th} only works under Assumption $\mathbf{A}_{\mathbf{mod}}^{(\mathbf{1})}$, but is more computing-saving since $\overline{F}_0$ is estimated anyway to compute the non adaptive estimators (see Δ_2^{th} in (4.10)).

Remark 9 *Actually, any constant $B > 1$ could be allowed in the above penalties provided slight changes in the definition of Δ_2^{th}, but we fix $B > 3$ for simplicity's sake. This point is discussed more precisely in Section* 4.6.5. *Nevertheless, as B tends to 1, the constants C and C' involved in Theorem* 4.3.1 *tend to infinity.*

4.3.3 Result

The following Theorem states the adaptivity of $\widehat{h}_{\widehat{m}_j}$.

Theorem 4.3.1 *Let $j = 1$ or 2, and $\widehat{m}_j$ defined by* (4.12)*, for some constant $B > 3$. Then under $\mathbf{A}_{\mathbf{mod}}^{(\mathbf{j})}$ and $\mathbf{A}_{\mathbf{frame}}$,*

$$\mathbb{E}\left[\|\widehat{h}_{\widehat{m}_j} - h\|_{\overline{F}_T}^2\right] \leq C \inf_{m \in J_n} \left\{ \inf_{t \in S_m} \|t - h\|_{\overline{F}_T}^2 + pen_j^{th}(m) \right\} + \frac{C'}{n} \tag{4.13}$$

where C is a numerical constant and C' depends on $(K, \overline{F}_0, \|h\|_\infty)$.

Comments

1. We do not study explicitly the risk of $\widehat{h}_m$ for one model S_m but a particular case of (4.13) when $\mathcal{M}_n$ is restricted to $\{S_m\}$ provides the following inequality.

$$\mathbb{E}\left[\|\widehat{h}_m - h\|_{\overline{F}_T}^2\right] \leq C'' \left\{ \inf_{t \in S_m} \|t - h\|_{\overline{F}_T}^2 + pen_j^{th}(m) \right\}$$

   for $j = 1$ or 2.

2. Huber and MacGibbon (2004) prove that the minimax rate of convergence on the Hölder space $\mathcal{H}(\beta, L)$, for $\beta > 0$ and $L > 0$ is the classical rate $n^{-2\beta/(2\beta+1)}$. Besides, suppose that $h \in \mathcal{H}(\beta, L)$,

$$\inf_{t \in S_m} \|t - h\|_{\overline{F}_T} \leq \inf_{t \in S_m} \|t - h\| \leq C(L, \beta) D_m^{-\beta}.$$

Thus for a model of dimension $D_{m^*} = E(n^{1/(2\beta+1)})$,

$$\mathbb{E}\left[\|\widehat{h}_{m^*} - h\|^2_{\overline{F}_T}\right] \leq Cn^{-2\beta/(2\beta+1)}.$$

So, for this choice D_{m^*}, $\widehat{h}_{m^*}$ is optimal in the minimax sense on the space $\mathcal{H}(\beta, L)$. Thus, the collection $\mathcal{M}_n$ contains an estimator with optimal rate, but the choice of $D_{m^*} = n^{1/(2\beta+1)}$ is not accessible as β is unknown.

3. The model selection procedure enables us to choose automatically such a model, without estimating β. More precisely, Inequality (4.13) (called an oracle inequality) shows that the risk bound of $\widehat{h}_{\widehat{m}_j}$ has same order as the risk of the best estimator among the collection $\{\widehat{h}_m, m \in \mathcal{M}_n\}$. In particular, $\widehat{h}_{\widehat{m}_j}$ reaches the minimax rate of convergence $n^{2\beta/(2\beta+1)}$ over all Hölder classes $\mathcal{H}(\beta, L)$ for $\beta > 0$, $L > 0$.

4.4 Data-driven estimators

The estimators presented in this section are similar to the ones of Section 4.3, but the unknown quantities $\overline{F}_0$ and $\|h\|_{\infty,A}$ are replaced by estimators.

4.4.1 Estimator of $\overline{F}_0$

$\overline{F}_0$ is the lower bound of $\overline{F}_T$ on $A = [0, a]$, so $\overline{F}_0 = \overline{F}_T(a)$. Thus a natural estimator of $\overline{F}_0$ would be the value of the empirical survival function in a. In order to force the estimator of $\overline{F}_0$ to be lower bounded, we define:

$$\widehat{F}_0 = \max(\alpha_n, \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{T_i \geq a\}}), \qquad \text{where } \alpha_n = 1/\sqrt{n}.$$

Let

$$\Delta_2 = \left\{\min(Sp(\widehat{G}_m)) \geq \frac{3}{5}\widehat{F}_0\right\} \quad \text{and} \quad \Delta_3 = \left\{\frac{3}{4}\overline{F}_0 \leq \widehat{F}_0 \leq \frac{5}{4}\overline{F}_0\right\}.$$

The following result holds.

Proposition 4.4.1

$$\Delta_1 \cap \Delta_3 \quad \subset \quad \Delta_2^{th} \cap \Delta_3 \quad \subset \quad \Delta_2 \cap \Delta_3$$

Proof of Proposition 4.4.1

According to Section 4.3.1,

$$\Delta_1 \subset \Delta_2^{th} \quad \Rightarrow \quad \Delta_1 \cap \Delta_3 \subset \Delta_2^{th} \cap \Delta_3.$$

Moreover, on the set $\Delta_2^{th} \cap \Delta_3$, according to Lemma 4.8.1,

$$\min(Sp(\widehat{G}_m)) = \inf_{\{t \in S_m,\ t \neq 0\}} \frac{\|t\|_n^2}{\|t\|^2} \geq \frac{3}{4}\overline{F}_0 \geq \frac{3}{4} \times \frac{4}{5}\widehat{F}_0 = \frac{3}{5}\widehat{F}_0$$

hence $\Delta_2^{th} \cap \Delta_3 \subset \Delta_2 \cap \Delta_3$. □

4.4.2 Estimator of $\|h\|_{\infty,A}$

Let $\nu = \|h\|_{\infty,A}$. Let $D = \lfloor n^\gamma \rfloor$ be a middle-sized model with $0 < \gamma < 1$, and $S_D = Vect(\varphi_1^D, \ldots, \varphi_D^D)$ be the set of piecewise constant functions on $[0, a]$ with $\varphi_j^D = \sqrt{D/a}\mathbb{1}_{[a(j-1)/D, aj/D)}$. Let $\widehat{h}_D = \arg\min_{t \in S_D} \gamma_n(t)$. As the basis functions (φ_j^D) have disjoint supports, the matrix $\widehat{G}_D$ of the scalar product $\langle ., . \rangle_n$ in the basis $(\varphi_1^D, \ldots, \varphi_D^D)$ is diagonal, with diagonal coefficients $\left(\|\varphi_j^D\|_n^2\right)_{j=1,\ldots,D}$. On Δ_2,

$$\|\varphi_j^D\|_n^2 \geq \frac{3}{5}\widehat{F}_0 \geq \frac{3}{5}\alpha_n > 0$$

thus $\widehat{G}_D$ is invertible. Let $\widehat{h}_D = \sum_{j=1,\ldots,D} \widehat{a}_j^D \varphi_j^D$ where

$$\widehat{a}_j^D = \begin{cases} \frac{(1/n)\sum_{i=1}^n \delta_i \varphi_j^D(T_i)}{\|\varphi_j^D\|_n^2} & \text{on } \Delta_2 \\ \\ 0 \quad \text{otherwise.} \end{cases}$$

Let $\widehat{\nu}_n = \|\widehat{h}\|_\infty$, then

$$\widehat{\nu}_n = \sqrt{\frac{D}{a}} \max_{j=1\ldots,D} \widehat{a}_j.$$

Besides, let h_D be the $\|.\|_{\overline{F}_T}$-projection of h on S_D,

$$h_D = \sum_{j=1,\ldots,D} a_j^D \varphi_j^D \quad \text{where} \quad a_j^D = \frac{\int_A \varphi_j^D(x) h(x) \overline{F}_T(x) dx}{\|\varphi_j^D\|_{\overline{F}_T(x)}^2}. \tag{4.14}$$

Finally, we define the following set whose probability is close to 1 (see Proposition 4.7.3),

$$\Delta_4 = \left\{ \frac{3}{4}\nu \leq \widehat{\nu}_n \leq \frac{5}{4}\nu \right\}.$$

4.4.3 Data-driven estimator

Let S_m be a model of the collection $\mathcal{M}_n$. We follow a procedure similar to the one described in Section 4.3.1, but now the set Δ_2^{th} is replaced by Δ_2. Let $\widehat{h}_m = \sum_{k=1}^{D_m} \widehat{a}_k^m \phi_k$ where $\widehat{A}_m =$

$(\widehat{a}_1^m, \ldots, \widehat{a}_{D_m}^m)^t = \widehat{G}_m^{-1}\widehat{V}_m$ on Δ_2, and 0 otherwise, and $\widehat{G}_m$ and $\widehat{V}_m$ are defined in (4.8). Moreover, let

$$pen_1(m) = \frac{BK^2}{\widehat{F}_0}\frac{D_m}{n}, \quad pen_2(m) = B\widehat{\nu}_n\frac{D_m}{n}$$

with $B > 15/4$. Finally we consider the estimators $\widehat{h}_{\widehat{m}_1}$ and $\widehat{h}_{\widehat{m}_2}$ where

$$\widehat{m}_j = \arg\min_{m\in J_n} \gamma_n(\widehat{h}_m) + pen_j(m) \tag{4.15}$$

for $j = 1$ or 2.

4.4.4 Results

Now our estimators are completely data-driven when B is chosen, and we can generalize Theorem 4.3.1 as follows.

Theorem 4.4.1 *Let $j = 1$ or 2. Assume that $\mathbf{A}_{\mathbf{mod}}^{(\mathbf{j})}$ and $\mathbf{A}_{\mathbf{frame}}$ hold, as well as the following condition:*

$$\|h - h_D\|_\infty \leq \frac{\nu}{8} \tag{4.16}$$

where h_D is defined by (4.14). Let $\widehat{m}_j$ be defined by (4.15), then

$$\mathbb{E}\left[\|\widehat{h}_{\widehat{m}_j} - h\|^2_{\overline{F}_T}\right] \leq C \inf_{m\in J_n}[\inf_{t\in S_m} \|t - h\|^2_{\overline{F}_T} + pen_j^{th}(m)] + \frac{C'}{n}$$

where C is an absolute constant and C' depends on $(K, \overline{F}_0, \|h\|_\infty, a)$.

Remark 10 *1. If h is in the Hölder space $\mathcal{H}(\beta, L)$ for some $\beta \in]0,1[$, $L > 0$, then (4.16) is satisfied for n large enough. In fact, let $y \in A$:*

$$\begin{aligned}
|h(y) - h_D(y)| &= \left| h(y) - \frac{(D/a)\int_{a(j-1)/D}^{aj/D} h(x)\overline{F}_T(x)dx}{(D/a)\int_{a(j-1)/D}^{aj/D} \overline{F}_T(x)dx} \right| \\
&= \frac{\left|(D/a)\int_{a(j-1)/D}^{aj/D} h(y)\overline{F}_T(x)dx - (D/a)\int_{a(j-1)/D}^{aj/D} h(x)\overline{F}_T(x)dx\right|}{(D/a)\int_{a(j-1)/D}^{aj/D} \overline{F}_T(x)dx} \\
&\leq \frac{\int_{a(j-1)/D}^{aj/D} |h(y) - h(x)|\overline{F}_T(x)dx}{\int_{a(j-1)/D}^{aj/D} \overline{F}_T(x)dx} \leq \frac{L}{D^\beta}.
\end{aligned}$$

Comments of Section 3.3.3 hold, thus the adaptive estimators are minimax over Hölder spaces.

2. As notified in Remark 9, B could be choosen as any numerical constant, provided that it is greater than 1.

4.5 Numerical examples

In this section, we present the performances of the model selection estimator on simulated data. For the sake of simplicity, we suppose that the parameters $\overline{F}_0$ and $\|h\|_{\infty,A}$ are known, which corresponds to the estimators described in Section 4.2. Besides, preliminary numerical studies prove that replacing these parameters by estimators only slightly affects the result. Indeed, this change is equivalent to a small change of the constant involved in the penalty. Moreover, simulations prove that the method is quite robust with respect to this constant. We consider two collections of models of functions supported on $[0, a]$ for some $a > 0$.

1. For $m \in \mathbb{N}^*$, let

$$S_m^{1,a} = vect\left(\{\phi_{1,m}^{(1)} : x \to 1\} \cup \{\phi_{k,m}^{(1)} : x \to \sqrt{2/a}\cos(\pi(k-1)x/a), k = 2, \ldots, m\}\right).$$

   This basis is not exactly the classical trigonometric basis: only cosine elements are used, and instead of $\cos(2\pi kx/a)$ and $\sin(2\pi kx/a)$ we consider $\cos(\pi kx/a)$. This model is based on the following observation. Let t be a function defined on $[0, a]$, t can be extended as an even function t^* on $[-a, a]$ by setting $t^*(-x) = t(x)$. Thus by classical Fourier analysis, t^* can be expanded in the basis $\{\cos(\pi kx/a)\}$ since the sine Fourier coefficients of t^* vanishes, and so does t. (See Efromovich (2007), Section 3, Remark 1 for more details.)

2. For $m \in \mathbb{N}^*$, let

$$S_m^{2,a} = vect(\{\phi_{k,m}^{(2)} : x \to \sqrt{m/a}\mathbb{1}_{[a(k-1)/m, ak/m)}(x), k = 1, \ldots, m\})$$

   be the set of histograms of step a/m.

Each simulation run is constructed as follows.

• Sequences $(Y_1, \ldots, Y_n)$ and $(C_1, \ldots, C_n)$ are simulated following one of these distributions.

(i) Y_i has a bimodal density defined by $f_Y = 0.8u + 0.2v$ where u is the density of $\exp(W/2)$ with $W \sim \mathcal{N}(0, 1)$ and v is the density of a gaussian distribution with mean 2 and variance 0.17. The (C_i)'s are generated from an exponential distribution with mean 2.5.

(ii) Y_i is generated from a gamma distribution with shape parameter 5 and scale 1. The (C_i)'s are generated from an exponential distribution with mean 6.

The mean of the (C_i)'s distribution is numerically choosen such that $\mathbb{E}[(1/n)(\sum_{i=1}^n \mathbb{1}_{\{Y_i \geq C_i\}})] = 0.4$.

• We compute the sample $(T_i, \delta_i)_{i=1,\ldots,n}$ where $T_i = \min(Y_i, C_i)$ and $\delta_i = \mathbb{1}_{\{Y_i \leq C_i\}}$.

• We consider either the collection $\{S_m^{1,a}, m = 1, \ldots, \sqrt{n}\}$ or $\{S_m^{2,a}, m = 1, \ldots, \sqrt{n}\}$ supported on $[0, a]$ with: (i) a=3 (ii) a=10. In numerical computing we can restrict ourselves to models m smaller than $\sqrt{n}$ without altering the result, since in practice the selected model is much smaller than $\sqrt{n}$. For every $m \in \{1, \ldots, \sqrt{n}\}$, we compute the matrix

$$\widehat{G}_m = \left(\frac{1}{n} \sum_{i=1}^{n} \int_0^a \phi_{k,m}^{(j)}(x) \phi_{k',m}^{(j)}(x) \mathbb{1}_{\{x \leq T_i\}} dx \right)_{k,k'=1,\ldots,m},$$

the column vector

$$\widehat{V}_m = \left[\frac{1}{n} \sum_{i=1}^{n} \delta_i \phi_{k,m}^{(j)}(T_i) \right]_{k=1,\ldots,m},$$

and the coefficient vector

$$\widehat{A}_m = \begin{cases} \widehat{G}_m^{-1} \widehat{V}_m & \text{if } \max(Sp(\widehat{G}_m)) \geq \overline{F}_0/4 \\ \\ 0 & \text{otherwise} \end{cases}$$

where $\overline{F}_0 = \overline{F}_T(a)$ and $\overline{F}_T$ is the survival function of T.

• We select the model $\widehat{m} \in \{1, \ldots \sqrt{n}\}$ which minimizes

$$\widehat{A}_m^t \widehat{G}_m \widehat{A}_m - 2\widehat{A}_m^t \widehat{V}_m + \frac{3}{2} \|\widehat{h}\|_{\infty,A} \frac{m}{n} = -\widehat{A}_m^t \widehat{V}_m + \frac{3}{2} \|\widehat{h}\|_{\infty,A} \frac{m}{n}$$

where $\|h\|_{\infty,A}$ is the maximum of h on $A = [0, a]$. The constant $3/2$ in the penalty was chosen from simulations over a wide variety of distributions, only a few of them are presented here.

• Let I be the set of 100a+1 equispaced points in [0,a]. For every $x \in I$, we compute

$$\widehat{h}_{\widehat{m}}(x) = \sum_{k=0}^{\widehat{m}-1} \widehat{a}_k^{\widehat{m}} \phi_{k,\widehat{m}}^{(j)}(x)$$

where $\widehat{A}_m = (\widehat{a}_0^m, \ldots, \widehat{a}_{m-1}^m)^t$. We plot the set of points $\{(x, \widehat{h}_{\widehat{m}}(x)), x \in I\}$.

4.5.1 Bimodal distribution

Consider the model (i). Figure 4.1 illustrates the performance of the model selection estimator in trigonometric basis for sample sizes $n = 200$, 500 and 1000, and Figure 4.2 presents a beam of 20 estimators for $n = 500$. We notice that the estimation is bad at the end of the interval, which is classically observed in hazard rate estimation. Besides, this behaviour is consistent with the theoretical aspect. Indeed $\overline{F}_T$ is decreasing, so the $\|.\|_{\overline{F}_T}$-risk puts more weight on the

beginning of the interval and the bad estimation at the end of the interval has less influence on the risk $\|\widehat{h}_{\widehat{m}} - h\|_{\overline{F}_T}$. Thus, similarly to Comte and Brunel (2005) and Antoniadis et al. (1999), we also present in Figure 4.2 the beam of curves of 20 estimators restricted to the interval $[0, 2]$. We note that estimation is much better than on the full interval $[0, 3]$.

4.5.2 Gamma distribution

Consider the model (ii). Figure 4.3 illustrates the performance of the model selection estimator in histogram basis for sample sizes $n = 200$, 500 and 1000. Similarly to the bimodal model, we present in Figure 4.4 a beam of 20 estimators for $n = 1000$ on $[0, 10]$ and the same beam restricted to the interval $[0, 6]$.

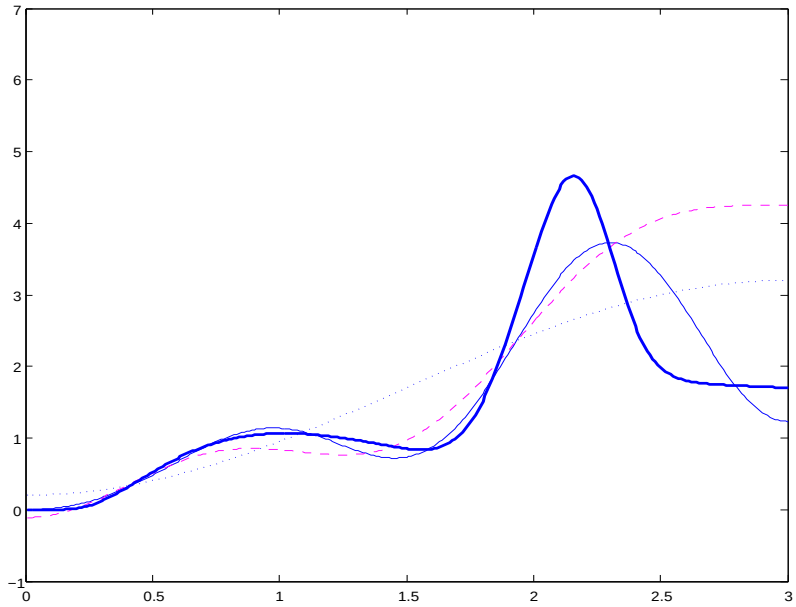


Figure 4.1: Model selection estimator in trigonometric basis for $n = 200$ (small dotted line), $n = 500$ (large dotted line) and $n = 1000$ (solid line), and true h (thick line) from the bimodal distribution.

4.5.3 Numerical results

Examples (i) and (ii) have been studied in Antoniadis et al. (1999), Reynaud-Bouret (2006) and Comte and Brunel (2005). Antoniadis et al. (1999) use wavelet methods, Reynaud-Bouret (2006) builds an histogram estimator, and Comte and Brunel (2005) also use model selection

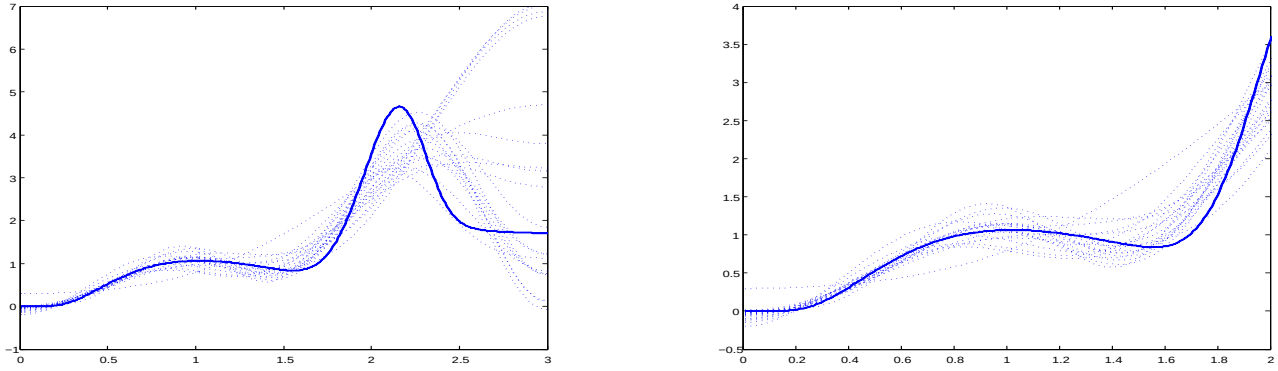


Figure 4.2: Beam of 20 estimators in trigonometric basis for $n = 500$ (small dotted line) from the bimodal distribution and true h (thick line).

but with a different contrast. The authors study the squared integrated risk of the estimator $\widehat{h}$: they draw a large number L of replications $\{(T_i^l, \delta_i^l)_{i=1,\dots,n}, l = 1, \dots, L\}$, and compute an estimator $\widehat{h}^l$ for each replication. Then the MSE is the quantity

$$\frac{1}{L}\sum_{l=1}^{L}\left[\frac{1}{J}\sum_{j=1}^{J}(\widehat{h}^l(t_j) - h(t_j))^2\right] \tag{4.17}$$

where the (t_j)'s are regularly spaced points in the interval $[0, \max(T_i)]$. It turns out to be difficult to compare their estimators with our since we do not estimate h on the same interval. Nevertheless, they also compute an error MSE2 similarly to MSE but on a restricted interval $[0, b]$ with $b = 6$ in the gamma case and $b = 2$ in the bimodal case. More precisely, MSE2 is equal to (4.17) where the (t_k)'s are regularly spaced points in the interval $[0, b]$. In our examples, we take 100a +1 equispaced points in [0,a] and $L = 500$ replications.

The performances of the three above-mentionned estimators for the MSE2 are gathered in Table 4.1, whereas Table 4.2 shows the performance of our estimator in bimodal and gamma case. We notice that our estimator provides slightly better results for the MSE2 in the gamma model, and slightly less good results in the bimodal model.

We also compute the mean and the empirical variance of the selected model for $L = 500$

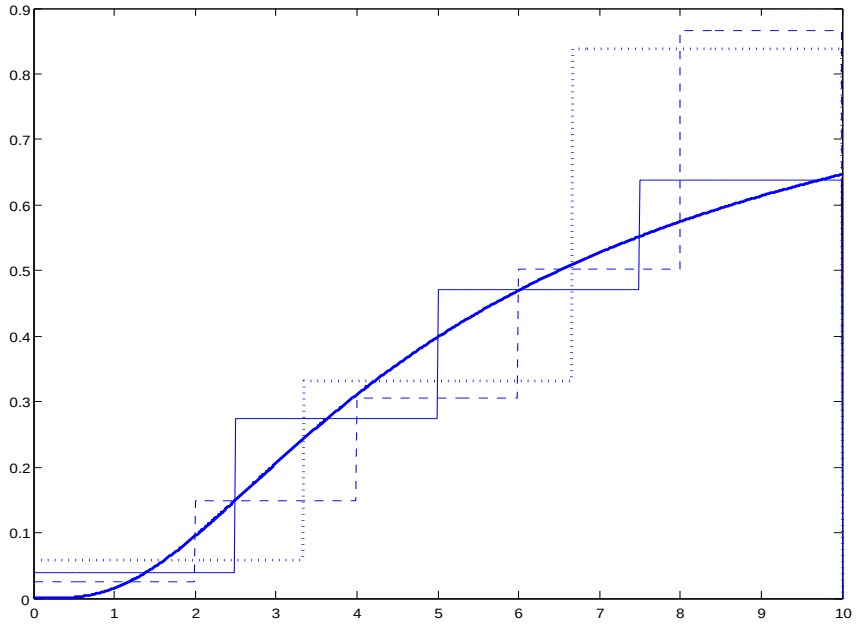


Figure 4.3: Model selection estimator in histogram basis for $n = 200$ (small dotted line), $n = 500$ (large dotted line) and $n = 1000$ (solid line), and true h (thick line) from the gamma distribution.

replications, namely,

$$\overline{\widehat{m}} = \frac{1}{L}\sum_{l=1}^{L} \widehat{m}^l \quad \text{and} \quad Var(\widehat{m}) = \frac{1}{L}\sum_{l=1}^{L} (\widehat{m}^l - \overline{\widehat{m}})^2$$

where $\widehat{m}^l$ is the selected model from the l^{th} sample. The results gathered in Table 4.3 indicate that the model selection algorithm really selects various values of $\widehat{m}$ for the different runs (see the variance of the chosen $\widehat{m}$'s), and thus adapts really to the data.

Besides, our estimator is quite fast-computing. (For example, the running time for the MSE2 computed with L=500 replications of a sample of size n=500 is a few seconds on a personnal computer.)

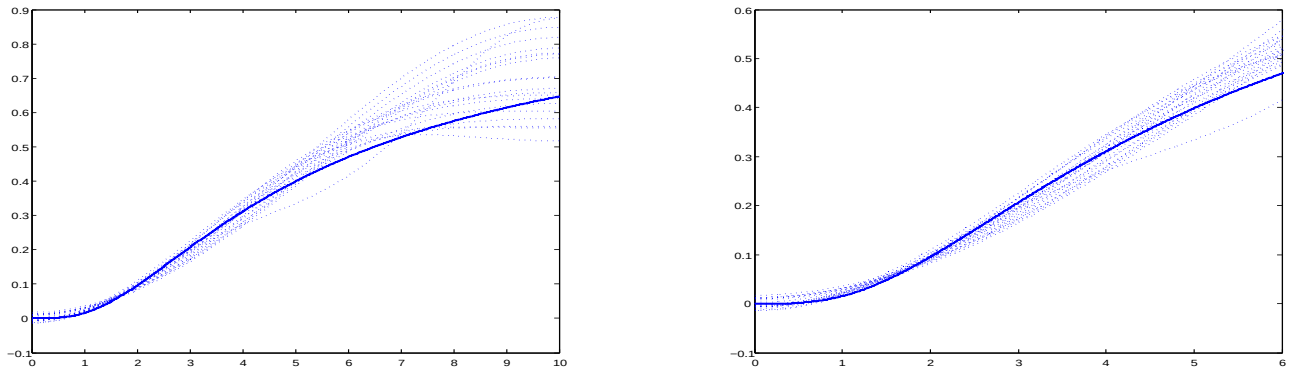


Figure 4.4: Beam of 20 estimators in trigonometric basis for $n = 500$ (small dotted line) from the gamma distribution and true h (thick line).

	Antoniadis and al.		Brunel and Comte		Reynaud-Bouret	
model	gamma	bimodal	gamma	bimodal	gamma	bimodal
n=200	0.0025	0.048	0.0023	0.1068	0.0032	0.150
n=500	0.0016	0.032	0.0013	0.0408	0.0012	0.051

Table 4.1: Results of MSE2 for the estimators of Antoniadis et al. (1999), Comte and Brunel (2005) and Reynaud-Bouret (2006), for bimodal and gamma models.

4.6 Proof of Theorem 4.3.1

The following Propositions are intermediate results to prove Theorem 4.3.1. Assume that $\mathbf{A_{frame}}$ holds.

Proposition 4.6.1 *Let $j = 1$ or 2. Under $\mathbf{A}^{(\mathbf{j})}_{\mathbf{mod}}$,*

$$\mathbb{E}\left[\|\widehat{h}_{\widehat{m}_j} - h\|^2_{\overline{F}_T} \mathbb{1}_{\Delta_1}\right] \leq C \inf_{m \in J_n} \left[\inf_{t \in S_m} \|t - h\|^2_{\overline{F}_T} + pen^{th}_j(m)\right] + \frac{C'}{n} \tag{4.18}$$

where C is a numerical constant and C' depends on $(K, \overline{F}_0, \|h\|_\infty)$.

Proposition 4.6.2 *For every model $S_m \in \mathcal{M}_n$,*

$$\|\widehat{h}_m - h\|_{\overline{F}_T} \leq \frac{2K\sqrt{N_n}}{\overline{F}_0} + \|h\|_{\overline{F}_T} \ a.s.$$

model	gamma	bimodal
n=200	0.0019	0.172
n=500	0.0009	0.080

Table 4.2: Results of MSE2 for our estimator, for bimodal and gamma models.

	gamma		bimodal	
	$\overline{\widehat{m}}$	$Var(\widehat{m})$	$\overline{\widehat{m}}$	$Var(\widehat{m})$
n=200	2.26	0.51	3.33	2.50
n=500	2.50	0.65	5.00	2.85
n=1000	2.89	0.65	5.49	2.89

Table 4.3: Mean and empirical variance of $\widehat{m}$ for bimodal and gamma models.

Proposition 4.6.3 *Assume that* $\mathbf{A_{mod}^{(1)}}$ *or* $\mathbf{A_{mod}^{(2)}}$ *holds, then*

$$P[\Delta_1^c] \leq 2N_n^2 \exp\left(-C_1 \overline{F}_0^2 \frac{n}{N_n}\right)$$

where C_1 *is a numerical constant.*

4.6.1 Proof of Theorem 4.3.1

Under $\mathbf{A_{mod}^{(j)}}$, $N_n \leq n/(\log n)^2$ so according to Propositions 4.6.2 and 4.6.3,

$$\begin{aligned}
&\mathbb{E}\left[\|\widehat{h}_{\widehat{m}_j} - h\|^2_{\overline{F}_T} \mathbb{1}_{\Delta_1^c}\right] \leq 2N_n^2 \left(\frac{2K\sqrt{N_n}}{\overline{F}_0} + \|h\|_{\overline{F}_T}\right)^2 \exp\left(-C_1 \overline{F}_0^2 \frac{n}{N_n}\right) \\
&\leq Cn^3 \exp\left(-C_1 \overline{F}_0^2 (\log n)^2\right) \\
&= Cn^3 \left(\exp(-C_1 \overline{F}_0^2 \log n)\right)^{\log n} = Cn^3 \left[n^{-C_1 \overline{F}_0^2}\right]^{\log n} = Cn^{3-C_2 \overline{F}_0^2 \log n} \leq \frac{C'}{n} \qquad (4.19)
\end{aligned}$$

which, together with Proposition 4.6.1, ends the proof of Theorem 4.3.1. □

.

4.6.2 Proof of Proposition 4.6.1

Let j=1 or 2. To simplify notations, we denote $pen(m) = pen_j^{th}(m)$ and $\widehat{m} = \widehat{m}_j$. Let S_m be a model in the collection $\mathcal{M}_n$ and h_m be any function in S_m. On the set Δ_2^{th}, by definition of $\widehat{m}_j$,

$$\gamma_n(\widehat{h}_{\widehat{m}}) + pen(\widehat{m}) \leq \gamma_n(h_m) + pen(m).$$

Thus,

$$\|\widehat{h}_{\widehat{m}}\|_n^2 - \|h_m\|_n^2 \leq pen(m) - pen(\widehat{m}) + \frac{2}{n}\sum_{i=1}^n \delta_i(\widehat{h}_{\widehat{m}} - h_m)(T_i).$$

Besides, $\|\widehat{h}_{\widehat{m}} - h_m\|_n^2 = \|\widehat{h}_{\widehat{m}}\|_n^2 + \|h_m\|_n^2 - 2\langle \widehat{h}_{\widehat{m}}, h_m\rangle_n$, so

$$\begin{aligned}
\|\widehat{h}_{\widehat{m}} - h_m\|_n^2 &\leq pen(m) - pen(\widehat{m}) + \frac{2}{n}\sum_{i=1}^n \delta_i(\widehat{h}_{\widehat{m}} - h_m)(T_i) - 2\langle \widehat{h}_{\widehat{m}}, h_m\rangle_n + 2\|h_m\|_n^2 \\
&= pen(m) - pen(\widehat{m}) - 2\langle \widehat{h}_{\widehat{m}} - h_m, h_m\rangle_n + \frac{2}{n}\sum_{i=1}^n \delta_i(\widehat{h}_{\widehat{m}} - h_m)(T_i) \\
&= pen(m) - pen(\widehat{m}) - 2\langle \widehat{h}_{\widehat{m}} - h_m, h_m - h\rangle_n + 2\nu_n(\widehat{h}_{\widehat{m}} - h_m)
\end{aligned}$$

where

$$\nu_n(t) = \frac{1}{n}\sum_{i=1}^n \delta_i t(T_i) - \langle t, h\rangle_n.$$

Let $S_m + S_{m'} = \{t + t', t \in S_m, t' \in S_{m'}\}$. The application ν_n is linear hence

$$\|\widehat{h}_{\widehat{m}} - h_m\|_n^2 \leq pen(m) - pen(\widehat{m}) - 2\langle \widehat{h}_{\widehat{m}} - h_m, h_m - h\rangle_n + 2\|\widehat{h}_{\widehat{m}} - h_m\|_{\overline{F}_T} \sup_{t \in S_m + S_{\widehat{m}}, \|t\|_{\overline{F}_T}=1} \nu_n(t).$$

Moreover, for every numbers b and c, $2bc \leq 2b^2 + (1/2)c^2$, and with Cauchy-Schwartz inequality,

$$\begin{aligned}
\|\widehat{h}_{\widehat{m}} - h_m\|_n^2 &\leq pen(m) - pen(\widehat{m}) + 2\|\widehat{h}_{\widehat{m}} - h_m\|_n\|h_m - h\|_n + \frac{1}{2}\|\widehat{h}_{\widehat{m}} - h_m\|_{\overline{F}_T}^2 \\
&\quad + 2\sup_{t \in S_m + S_{\widehat{m}}, \|t\|_{\overline{F}_T}=1} (\nu_n(t))^2.
\end{aligned}$$

For every $p(m, m')$ function of (m, m'),

$$\begin{aligned}
\|\widehat{h}_{\widehat{m}} - h_m\|_n^2 &\leq pen(m) - pen(\widehat{m}) + 2p(m, \widehat{m}) + 2\|\widehat{h}_{\widehat{m}} - h_m\|_n\|h_m - h\|_n + \frac{1}{2}\|\widehat{h}_{\widehat{m}} - h_m)\|_{\overline{F}_T}^2 \\
&\quad + 2\sup_{t \in S_m + S_{\widehat{m}}, \|t\|_{\overline{F}_T}=1} [(\nu_n(t))^2 - p(m, \widehat{m})] \\
&\leq pen(m) - pen(\widehat{m}) + 2p(m, \widehat{m}) + \frac{1}{4}\|\widehat{h}_{\widehat{m}} - h_m\|_n^2 + 4\|h - h_m\|_n^2 \\
&\quad + \frac{1}{2}\|\widehat{h}_{\widehat{m}} - h_m\|_{\overline{F}_T}^2 + 2\sup_{t \in S_m + S_{\widehat{m}}, \|t\|_{\overline{F}_T}=1} [(\nu_n(t))^2 - p(m, \widehat{m})]
\end{aligned}$$

since $2bc \leq 4b^2 + (1/4)c^2$ for every b, c. Thus

$$\begin{aligned}\frac{3}{4}\|\widehat{h}_{\widehat{m}} - h_m\|_n^2 &\leq pen(m) - pen(\widehat{m}) + 2p(m,\widehat{m}) + \frac{1}{2}\|\widehat{h}_{\widehat{m}} - h_m\|_{\overline{F}_T}^2 + 4\|h - h_m\|_{\overline{F}_T}^2 \\ &\quad +2 \sup_{t \in S_m + S_{\widehat{m}}, \|t\|_{\overline{F}_T}=1} [(\nu_n(t))^2 - p(m,\widehat{m})].\end{aligned}$$

On the set $\Delta_1 \cap \Delta_2^{th} = \Delta_1$ (see Proposition 4.4.1), $\|\widehat{h}_{\widehat{m}} - h_m\|_n^2 \geq (3/4)\|\widehat{h}_{\widehat{m}} - h_m\|_{\overline{F}_T}^2$ hence

$$\begin{aligned}\left(\left(\frac{3}{4}\right)^2 - \frac{1}{2}\right)\|\widehat{h}_{\widehat{m}} - h_m\|_{\overline{F}_T}^2 &\leq 4\|h - h_m\|_n^2 + pen(m) - pen(\widehat{m}) + 2p(m,\widehat{m}) \\ &\quad +2 \sup_{t \in S_m + S_{\widehat{m}}, \|t\|_{\overline{F}_T}=1} [(\nu_n(t))^2 - p(m,\widehat{m})].\end{aligned}$$

Moreover, we note that $\|\widehat{h}_{\widehat{m}} - h_m\|_{\overline{F}_T}^2 \geq (1/2)\|\widehat{h}_{\widehat{m}} - h\|_{\overline{F}_T}^2 - \|h - h_m\|_{\overline{F}_T}^2$ and $\mathbb{E}\left[\|h - h_m\|_n^2\right] = \|h - h_m\|_{\overline{F}_T}^2$ so

$$\begin{aligned}\mathbb{E}\left[\|\widehat{h}_{\widehat{m}} - h\|_{\overline{F}_T}^2 \mathbb{1}_{\Delta_1}\right] &\leq C_2 \left\{\|h - h_m\|_{\overline{F}_T}^2 + \mathbb{E}[pen(m) - pen(\widehat{m}) + 2p(m,\widehat{m})]\right. \\ &\quad \left. +\mathbb{E}\left[\sup_{t \in S_m + S_{\widehat{m}}, \|t\|_{\overline{F}_T}=1} \left((\nu_n(t))^2 - p(m,\widehat{m})\right)\right]\right\} \qquad (4.20)\end{aligned}$$

where C_2 is a numerical constant.

On the one hand, consider $p(m,m') = (pen(m) + pen(m'))/2$, then

$$pen(m) - pen(\widehat{m}) + 2p(m,\widehat{m}) = 2pen(m) \qquad (4.21)$$

On the other hand, $\nu_n(t)$ is a centered process since

$$\mathbb{E}[\delta_i t(T_i)] = \int_A t(x)h(x)\overline{F}_T(x)dx = \mathbb{E}[\langle t, h\rangle_n]$$

(see (4.7)). Therefore, we insert the mean term $\int_A t(x)h(x)\overline{F}_T(x)dx$ to obtain the sum of two variance-type terms. More precisely, we define

$$\nu_{n,1}(t) = \frac{1}{n}\sum_{i=1}^n \delta_i t(T_i) - \int_A t(x)h(x)\overline{F}_T(x)dx \qquad \text{and}$$

$$\nu_{n,2}(t) = \frac{1}{n}\sum_{i=1}^n \int_A t(x)h(x)\mathbb{1}_{T_i \geq x}dx - \int_A t(x)h(x)\overline{F}_T(x)dx.$$

Then, as $(b+c)^2 \leq \frac{3}{2}b^2 + 3c^2$ for every b and c,

$$\mathbb{E}\left[\sup_{t\in S_m+S_{\widehat{m}},\|t\|_{\overline{F}_T}=1}((\nu_n(t))^2 - p(m,\widehat{m}))\right]$$
$$\leq \frac{3}{2}\mathbb{E}\left[\sup_{t\in S_m+S_{\widehat{m}},\|t\|_{\overline{F}_T}=1}((\nu_{n,1}(t))^2 - \frac{2}{3}p(m,\widehat{m}))_+\right] + 3\mathbb{E}\left[\sup_{t\in S_m+S_{\widehat{m}},\|t\|_{\overline{F}_T}=1}(\nu_{n,2}(t))^2\right]. \quad (4.22)$$

Moreover, the two terms above are upper-bounded as follows.

Lemma 4.6.1

$$\mathbb{E}\left[\sup_{t\in S_m+S_{\widehat{m}},\|t\|_{\overline{F}_T}=1}(\nu_{n,2}(t))^2\right] \leq \frac{1}{\overline{F}_0 n}\|h\|^2_{\overline{F}_T}.$$

Lemma 4.6.2 *Let $j = 1$ or 2, and $\widehat{m}_j$ defined by (4.12) with $B > 3$. Then under $\mathbf{A}^{(\mathbf{j})}_{\mathbf{mod}}$,*

$$\mathbb{E}\left[\left(\sup_{t\in S_m+S_{\widehat{m}_j},\|t\|_{\overline{F}_T}=1}(\nu_{n,1}(t))^2 - \frac{2}{3}p_j(m,\widehat{m}_j)\right)_+\right] \leq \frac{C'}{n}$$

for some constant C' depending on $(K, \|h\|_{\infty,A}, \overline{F}_0)$.

Finally inequalities (4.20), (4.21), (4.22) and Lemmas 4.6.1 and 4.6.2 ends the proof of Proposition 4.6.1. □

Proof of Lemma 4.6.1

For every m, $\widehat{m} \in J_n$, $S_m + S_{\widehat{m}} \subset S_n$, hence

$$\mathbb{E}\left[\sup_{t\in S_m+S_{\widehat{m}},\|t\|_{\overline{F}_T}=1}(\nu_{n,2}(t))^2\right] \leq \mathbb{E}\left[\sup_{t\in S_n,\|t\|_{\overline{F}_T}=1}(\nu_{n,2}(t))^2\right].$$

Besides, by $\mathbf{A}_{\mathbf{frame}}$,

$$\left\{t, \|t\|^2_{\overline{F}_T} \leq 1\right\} \subset \left\{t, \|t\|^2 \leq \overline{F}_0^{-1}\right\} \quad (4.23)$$

so

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in S_m+S_{\widehat{m}_i},\|t\|_{\overline{F}_T}=1}(\nu_{n,2}(t))^2\right] &\leq \mathbb{E}\left[\sup_{t\in S_n,\|t\|^2\leq 1/\overline{F}_0}(\nu_{n,2}(t))^2\right] \\ &= \mathbb{E}\left[\sup_{\sum_{k=1}^{N_n}a_k^2\leq 1/\overline{F}_0}\left(\sum_{k=1}^{N_n}a_k(\langle\phi_k^n,h\rangle_n - \langle\phi_k^n,h\rangle_{\overline{F}_T})\right)^2\right]\end{aligned}$$

where $(\phi_1^n, \ldots, \phi_{N_n}^n)$ is an $\|.\|$-orthonormal basis of S_n. With Cauchy-Schwartz Inequality, we obtain

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\in S_m+S_{\widehat{m}},\|t\|_{\overline{F}_T}=1}(\nu_{n,2}(t))^2\right] &\leq \frac{1}{\overline{F}_0}\sum_{k=1}^{N_n}\mathbb{E}\left[(\langle\phi_k^n,h\rangle_n-\langle\phi_k^n,h\rangle_{\overline{F}_T})^2\right]\\
&= \frac{1}{\overline{F}_0}\sum_{k=1}^{N_n}\frac{1}{n}Var\left[\int_A h(x)\phi_k^n(x)\mathbb{1}_{\{T_1\geq x\}}dx\right]\\
&\leq \frac{1}{\overline{F}_0 n}\mathbb{E}\left[\sum_{k=1}^{N_n}\langle\phi_k^n,h(.)\mathbb{1}_{\{T_1\geq .\}}\rangle^2\right]\\
&= \frac{1}{\overline{F}_0 n}\mathbb{E}\left[\|(h(.)\mathbb{1}_{\{T_1\geq .\}})_{S_n}\|^2\right]
\end{aligned}$$

where $(h(.)\mathbb{1}_{\{T_1\geq .\}})_{S_n}$ denotes the L^2-orthogonal projection of $h(.)\mathbb{1}_{\{T_1\geq .\}}$ on S_n. Thus

$$\mathbb{E}\left[\sup_{t\in S_m+S_{\widehat{m}},\|t\|_{\overline{F}_T}=1}(\nu_{n,2}(t))^2\right]\leq\frac{1}{\overline{F}_0 n}\mathbb{E}\left[\|h(.)\mathbb{1}_{\{T_1\geq .\}}\|^2\right]=\frac{1}{\overline{F}_0 n}\|h\|^2_{\overline{F}_T}. \quad \square$$

Proof of Lemma 4.6.2

For j=1 or 2, we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\sup_{t\in S_m+S_{\widehat{m}_j},\|t\|_{\overline{F}_T}=1}(\nu_{n,1}(t))^2-\frac{2}{3}p_j(m,\widehat{m}_j)\right)_+\right]\\
&\leq\sum_{m'\in J_n}\mathbb{E}\left[\left(\sup_{t\in S_m+S_{m'},\|t\|_{\overline{F}_T}=1}(\nu_{n,1}(t))^2-\frac{2}{3}p_j(m,m')\right)_+\right].
\end{aligned}$$

Besides, for every models $S_m, S_{m'}$, we upper bound the term $\mathbb{E}[(\sup_{t\in S_m+S_{m'},\|t\|_{\overline{F}_T}=1}(\nu_{n,1}(t))^2-p_j(m,m'))_+]$ with Talagrand Inequality (see Introduction, Theorem 1.2.3).

• Consider $j=1$. Under $\mathbf{A}^{(1)}_{\mathbf{mod}}$, $S_m\subset S_{m'}$ or $S_{m'}\subset S_m$. Thus $S_m+S_{m'}$ is equal either to S_m or to $S_{m'}$. Let $D_{m+m'}=\max(D_m,D_{m'})$ denote the dimension of $S_m+S_{m'}$, and $(\phi_1^{m+m'},\ldots,\phi_{D_{m+m'}}^{m+m'})$ be the orthonormal basis of $S_m+S_{m'}$ defined as $\phi_k^{m+m'}=\phi_k^m$ if $S_m+S_{m'}=S_m$, and $\phi_k^{m'}$ if $S_m+S_{m'}=S_{m'}$.

Let us compute the term $\mathbb{H}$. With (4.23),

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in S_m+S_{m'},\|t\|_{\overline{F}_T}=1}(\nu_{n,1}(t))^2\right] &\leq \mathbb{E}\left[\sup_{t\in S_m+S_{m'},\|t\|\leq 1/\overline{F}_0}(\nu_{n,1}(t))^2)_+\right]\\ &\leq \frac{1}{\overline{F}_0}\sum_{k=1}^{D_{m+m'}}\frac{1}{n}Var\left[\delta_1\phi_k^{m+m'}(T_1)\right]\\ &\leq \frac{1}{\overline{F}_0}\sum_{k=1}^{D_{m+m'}}\frac{1}{n}\mathbb{E}\left[(\phi_k^{m+m'})^2(T_1)\right]. \qquad (4.24)\end{aligned}$$

Besides, according to (4.5) in Assumption $\mathbf{A}^{(\mathbf{1})}_{\mathbf{mod}}$,

$$\mathbb{E}\left[\sup_{t\in S_m+S_{m'},\|t\|_{\overline{F}_T}=1}(\nu_{n,1}(t))^2\right]\leq\frac{1}{n\overline{F}_0}\sup_{x\in A}\left|\sum_{k=1}^{D_{m+m'}}(\phi_k^{m+m'}(x))^2\right|\leq\frac{K^2(D_m+D_{m'})}{\overline{F}_0 n}=\mathbb{H}^2.$$

Moreover, we assume that $B>3$, so $(2/3)p_1(m,m')=\theta\mathbb{H}^2$ for some $\theta>1$. Let us compute the terms c and v involved in Talagrand Inequality.

$$\begin{aligned}\sup_{t\in S_m+S_{m'},\|t\|_{\overline{F}_T}=1}\|\delta_1 t(T_1)\|_\infty &\leq \sup_{t\in S_m+S_{m'},\|t\|^2\leq 1/\overline{F}_0}\left(\sup_{x\in A}|t(x)|\right)\\ &= \sup_{\sum_{k=1}^{D_{m+m'}}a_k^2\leq 1/\overline{F}_0}\left(\sup_{x\in A}\left|\sum_{k=1}^{D_{m+m'}}a_k\phi_k^{m+m'}(x)\right|\right).\end{aligned}$$

With Cauchy-Schwartz Inequality,

$$\sup_{t\in S_m+S_{m'},\|t\|_{\overline{F}_T}=1}\|\delta_1 t(T_1)\|_\infty\leq\frac{1}{\sqrt{\overline{F}_0}}\sup_{x\in A}\left|\sum_{k=1}^{D_{m+m'}}(\phi_k^{m+m'}(x))^2\right|\leq\frac{K}{\sqrt{\overline{F}_0}}\sqrt{D_m+D_{m'}}=b.$$

Moreover, (4.7) entails

$$\begin{aligned}\sup_{t\in S_m+S_{m'},\|t\|_{\overline{F}_T}=1}Var(\delta_1 t(T_1)) &\leq \sup_{t\in S_m+S_{m'},\|t\|_{\overline{F}_T}=1}\mathbb{E}[\delta_1 t^2(T_1)]\\ &= \sup_{t\in S_m+S_{m'},\|t\|_{\overline{F}_T}=1}\int_A t^2(x)h(x)\overline{F}_T(x)dx\leq\|h\|_{\infty,A}=v.\end{aligned}$$

Then Talagrand Inequality (Theorem 1.2.3) leads to

$$\begin{aligned}
&\mathbb{E}\left[\left(\sup_{t\in S_m+S_{m'},\|t\|_{\overline{F}_T}=1}(\nu_{n,1}(t))^2-\frac{2}{3}p_1(m,m')\right)_+\right]\\
\leq\ &\overline{C}\frac{\|h\|_{\infty,A}}{n}\exp\left(-\kappa\frac{K^2(D_m+D_{m'})}{\overline{F}_0\|h\|_{\infty,A}}\right)+\overline{C}'\frac{K^2(D_m+D_{m'})}{\overline{F}_0^2n^2}\exp(-\kappa'\sqrt{n}).
\end{aligned}$$

Thus, with (4.6) in $\mathbf{A}^{(\mathbf{1})}_{\mathbf{mod}}$,

$$\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left[\left(\sup_{t\in S_m+S_{m'},\|t\|_{\overline{F}_T}=1}(\nu_{n,1}(t))^2-\frac{2}{3}p_1(m,m')\right)_+\right]\leq\frac{C'}{n}$$

which concludes the proof of Lemma 4.6.2 for $j=1$.

• Consider $j=2$. Let $(\psi_1,\ldots,\psi_{D_{m+m'}})$ be a $\|.\|_{\overline{F}_T}$-orthonormal basis of $S_m+S_{m'}$. Similarly to (4.24),

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\in S_m+S_{m'},\|t\|_{\overline{F}_T}=1}(\nu_{n,1}(t))^2\right]\ &\leq\ \frac{1}{n}\sum_{k=1}^{D_{m+m'}}Var(\delta_1\psi_k^2(Y_1))\leq\frac{1}{n}\sum_{k=1}^{D_{m+m'}}\mathbb{E}[\delta_1\psi_k^2(Y_1)]\\
&=\ \frac{1}{n}\sum_{k=1}^{D_{m+m'}}\int_A\psi_k^2(x)h(x)\overline{F}_T(x)dx\\
&\leq\ \|h\|_{\infty,A}\frac{D_m+D_{m'}}{n}=\mathbb{H}^2.
\end{aligned}$$

Besides, according to $\mathbf{A}^{(\mathbf{2})}_{\mathbf{mod}}$,

$$\sup_{t\in S_m+S_{m'},\|t\|_{\overline{F}_T}=1}\|\delta_1t(T_1)\|_\infty\leq\sup_{t\in S_m+S_{m'},\|t\|^2\leq1/\overline{F}_0}\sup_{x\in A}|t(x)|\leq\frac{K}{\sqrt{\overline{F}_0}}\sqrt{N_n}=b$$

and the end of the proof is similar to the case $j=1$. □

4.6.3 Proof of Proposition 4.6.2

Let $m\leq N_n$,

$$\|\widehat{h}_m-h\|_{\overline{F}_T}\leq\|\widehat{h}_m\|_{\overline{F}_T}+\|h\|_{\overline{F}_T}\leq\|\widehat{h}_m\|+\|h\|_{\overline{F}_T}.$$

On $\left(\Delta_2^{th}\right)^c$, $\|\widehat{A}_m\|=0$. On Δ_2^{th},

$$\|\widehat{A}_m\|\leq\max\left(Sp(\widehat{G}_m^{-1})\right)\|\widehat{V}_m\|=\left[\min\left(Sp(\widehat{G}_m)\right)\right]^{-1}\|\widehat{V}_m\|\leq\frac{4}{3\overline{F}_0}\|\widehat{V}_m\|$$

Hence

$$\begin{aligned}
\|\widehat{h}_m - h\|_{\overline{F}_T} &\leq \frac{4}{3\overline{F}_0}\left[\sum_{k=1}^{D_m}\left(\frac{1}{n}\sum_{i=1}^{n}\phi_k^m(T_i)\delta_i\right)^2\right]^{1/2} + \|h\|_{\overline{F}_T} \\
&\leq \frac{4}{3\overline{F}_0}\left[\sum_{k=1}^{D_m}\frac{1}{n}\sum_{i=1}^{n}(\phi_k^m)^2(T_i)\right]^{1/2} + \|h\|_{\overline{F}_T} \\
&\leq \frac{4}{3\overline{F}_0}\sup_{x\in A}\left|\sum_{k=1}^{D_m}(\phi_k^m(x))^2\right|^{1/2} + \|h\|_{\overline{F}_T} \\
&\leq \frac{4K\sqrt{N_n}}{3\overline{F}_0} + \|h\|_{\overline{F}_T}. \quad \square
\end{aligned}$$

4.6.4 Proof of Proposition 4.6.3

The proof of Proposition 4.6.3 is inspired from Baraud (2002). By definition of Δ_1,

$$\Delta_1^c = \left\{|\|t\|_n^2 - \|t\|_{\overline{F}_T}^2| > \frac{1}{4}\|t\|_{\overline{F}_T}^2, \forall t \in S_n\right\} = \left\{\sup_{t\in S_n, \|t\|_{\overline{F}_T}\leq 1}\left|\eta_n(t^2)\right| > \frac{1}{4}\right\}$$

where $\eta_n(t) = (1/n)\sum_{i=1}^n(\int_A t(x)\mathbb{1}_{\{T_i\geq x\}}dx - \int_A t(x)\overline{F}_T(x)dx)$. Besides, for every $t \in S_n$,

$$\overline{F}_0\|t\|^2 \leq \|t\|_{\overline{F}_T}^2.$$

Thus,

$$\Delta_1^c \subset \left\{\sup_{t\in S_n, \|t\|^2\leq 1/\overline{F}_0}\left|\eta_n(t^2)\right| > \frac{1}{4}\right\}.$$

Let $(\psi_1, \ldots, \psi_{N_n})$ be an orthonormal base of the global space S_n for the norm $\|.\|$, then

$$\Delta_1^c \subset \left\{\sup_{\sum a_k^2 = 1/\overline{F}_0}\sum_{k,k'=1}^{N_n}|a_k||a_{k'}||S_{k,k'}| > \frac{1}{4}\right\}$$

where

$$S_{k,k'} = \frac{1}{n}\sum_{i=1}^{n}\left(\int_A \psi_k(x)\psi_{k'}(x)\mathbb{1}_{\{T_i\geq x\}}dx - \int_A \psi_k(x)\psi_{k'}(x)\overline{F}_T(x)dx\right).$$

On the one hand, let k, k' be fixed. Let

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(\int_A \psi_k\psi_{k'}\mathbb{1}_{\{T_i\geq x\}}dx\right)^2\right] = \mathbb{E}\left[\left(\int_A \psi_k\psi_{k'}\mathbb{1}_{\{T_1\geq x\}}dx\right)^2\right] = v_{k,k'}$$

and for every $l \geq 2$,

$$\begin{aligned}
\mathbb{E}\left[\left(\int_A \psi_k(x)\psi_{k'}(x)\mathbb{1}_{\{T_1\geq x\}}dx\right)_+^l\right] &\leq \mathbb{E}\left[\left(\int_A \psi_k(x)\psi_{k'}(x)\mathbb{1}_{\{T_1\geq x\}}dx\right)^2\left(\int_A |\psi_k(x)\psi_{k'}(x)|dx\right)^{l-2}\right] \\
&\leq v_{k,k'}\left(\int_A \psi_k^2(x)dx\int_A \psi_{k'}^2(x)dx\right)^{(l/2)-1} \\
&= v_{k,k'}c_{k,k'}^{l-2}
\end{aligned}$$

with $c_{k,k'} = 1$ for every k, k'. Thus Bernstein Inequality (Theorem 1.2.4) provides the following upper bound:

$$P\left[|S_{k,k'}| \geq \sqrt{2v_{k,k'}x} + c_{k,k'}x\right] \leq 2\exp(-nx), \quad \forall x > 0.$$

On the other hand,

$$\begin{aligned}
&\left\{|S_{k,k'}| < \sqrt{2v_{k,k'}x} + c_{k,k'}x, \forall k, k' = 1, \dots N_n\right\} \\
&\subset \left\{\sum_{k,k'=1}^{N_n}|a_k||S_{k,k'}||a_{k'}| < \sqrt{2x}\sum_{k,k'=1}^{N_n}|a_k|\sqrt{v_{k,k'}}|a_{k'}| + x\sum_{k,k'=1}^{N_n}|a_k|c_{k,k'}|a_{k'}|, \forall (a_k)_{k=1,\dots N_n}\right\} \\
&\subset \left\{\sup_{\sum a_k^2=1}\sum_{k,k'=1}^{N_n}|a_k||S_{k,k'}||a_{k'}| < \sqrt{2x}\sup_{\sum a_k^2=1}\sum_{k,k'=1}^{N_n}|a_k|\sqrt{v_{k,k'}}|a_{k'}| + x\sup_{\sum a_k^2=1}\sum_{k,k'=1}^{N_n}|a_k|c_{k,k'}|a_{k'}|\right\} \\
&= \left\{\sup_{t\in S_n, \|t\|^2\leq 1}|\eta_n(t^2)| \leq \sqrt{2x}\rho(V) + x\rho(C)\right\} \\
&= \left\{\sup_{t\in S_n, \|t\|^2\leq 1/\overline{F}_0}|\eta_n(t^2)| \leq \sqrt{2x}\frac{\rho(V)}{\overline{F}_0} + x\frac{\rho(C)}{\overline{F}_0}\right\}
\end{aligned}$$

where

$$\rho(V) = \sup_{\sum a_k^2=1}\sum_{k,k'=1}^{N_n}|a_k|\sqrt{v_{k,k'}}|a_{k'}|$$

$$\rho(C) = \sup_{\sum a_k^2=1}\sum_{k,k'=1}^{N_n}|a_k|c_{k,k'}|a_{k'}|.$$

Thus for every $x > 0$,

$$\begin{aligned}
P\left[\sup_{t\in S_n, \|t\|^2\leq 1/\overline{F}_0}|\eta_n(t^2)| > \sqrt{2x}\frac{\rho(V)}{\overline{F}_0} + x\frac{\rho(C)}{\overline{F}_0}\right] &\leq \sum_{k,k'=1}^{N_n} P\left[|S_{k,k'}| > \sqrt{2v_{k,k'}x} + c_{k,k'}x\right] \\
&\leq 2N_n^2\exp(-nx).
\end{aligned}$$

In order to upper bound $P[\Delta_1^c]$, we choose x such that $\sqrt{2x}\rho(V)/\overline{F}_0 \leq 1/8$ and $x\rho(C)/\overline{F}_0 \leq 1/8$. Let $L(\psi) = \max(\rho(C)/\overline{F}_0, 16(\rho(V)/\overline{F}_0)^2)$ then,

$$P[\Delta_1^c] \leq 2N_n^2 \exp\left(-\frac{n}{8L(\psi)}\right).$$

Let upper bound $L(\psi)$. Applying two times Cauchy-Schwartz Inequality, we obtain

$$\begin{aligned}(\rho(V))^2 &= \sup_{\sum a_k^2=1} \left[\sum_{k=1}^{N_n} |a_k| \left(\sum_{k'=1}^{N_n} |a_{k'}|\sqrt{v_{k,k'}}\right)\right]^2 \leq \sup_{\sum a_k^2=1} \left(\sum_{k=1}^{N_n} a_k^2\right) \left(\sum_{k=1}^{N_n} \left[\sum_{k'=1}^{N_n} ||a_{k'}|\sqrt{v_{k,k'}}\right]^2\right) \\ &= \sup_{\sum a_k^2=1} \sum_{k=1}^{N_n} \left(\sum_{k'=1}^{N_n} |a_{k'}|\sqrt{v_{k,k'}}\right)^2 \leq \sum_{k=1}^{N_n} \left(\sum_{k'=1}^{N_n} v_{k,k'}\right).\end{aligned}$$

We replace $v_{k,k'}$ by its expression.

$$(\rho(V))^2 \leq \sum_{k=1}^{N_n} \mathbb{E}\left[\sum_{k'=1}^{N_n} \langle \psi_{k'}, \psi_k 1\!\!1_{\{T_1 \geq .\}}\rangle^2\right]$$

Besides, $\sqrt{\sum_{k'=1}^{N_n} \langle \psi_{k'}, \psi_k 1\!\!1_{\{T_1 \geq .\}}\rangle^2}$ is equal to the norm of the $\|.\|$-projection of $\psi_k 1\!\!1_{\{T_1 \geq .\}}$ on S_n, so

$$\sum_{k'=1}^{N_n} \langle \psi_{k'}, \psi_k 1\!\!1_{\{T_1 \geq x\}}\rangle^2 \leq \|\psi_k 1\!\!1_{\{T_1 \geq .\}}\|^2 \leq \|\psi_k\|^2 = 1.$$

Hence $(\rho(V))^2 \leq N_n$. Moreover,

$$\rho(C) = \sup_{\sum a_k^2=1} \left(\sum_{k,k'=1}^{N_n} |a_k||a_{k'}|\right) = \sup_{\sum a_k^2=1} \left(\sum_{k=1}^{N_n} |a_k|\right)^2 \leq \sup_{\sum a_k^2=1} N_n \left(\sum_{k=1}^{N_n} a_k^2\right) = N_n.$$

Finally $L(\psi) \leq \max(N_n/\overline{F}_0, 16N_n/\overline{F}_0^2) = 16N_n/\overline{F}_0^2$ and

$$P[\Delta_1^c] \leq \exp\left(-C_1 \overline{F}_0^2 \frac{n}{N_n}\right). \quad \square$$

4.6.5 Comment about the constant in the penalty

Provided that the set Δ_2^{th} is replaced by

$$(\Delta_2^{th})' = \left\{\min(Sp(\widehat{G}_m) \geq (1-\alpha)\overline{F}_0\right\},$$

and the inequalities of the kind $2bc \leq 2b^2 + (1/2)c^2$ by $2bc \leq (1/\beta)b^2 + \beta c^2$ with α, β small enough, Theorem 4.3.1 holds for any constant $B > 1$.

4.7 Proof of Theorem 4.4.1

The following Propositions are intermediate results to prove Theorem 4.4.1. Assume that $\mathbf{A_{frame}}$ holds.

Proposition 4.7.1

1. *Under* $\mathbf{A^{(1)}_{mod}}$,

$$\mathbb{E}\left[\|\widehat{h}_{\widehat{m}_1} - h\|^2_{\overline{F}_T} \mathbb{1}_{\Delta_1\cap\Delta_2\cap\Delta_3}\right] \leq C \inf_{m\in\mathcal{M}_n}\left[\inf_{t\in S_m}\|t-h\|^2_{\overline{F}_T} + pen_1^{th}(m)\right] + \frac{C'}{n}$$

where C is a numerical constant and C' depends on $(K, \overline{F}_0, \|h\|_\infty)$.

2. *Under* $\mathbf{A^{(2)}_{mod}}$,

$$\mathbb{E}\left[\|\widehat{h}_{\widehat{m}_2} - h\|^2_{\overline{F}_T} \mathbb{1}_{\Delta_1\cap\Delta_2\cap\Delta_3\cap\Delta_4}\right] \leq C \inf_{m\in\mathcal{M}_n}\left[\inf_{t\in S_m}\|t-h\|^2_{\overline{F}_T} + pen_2^{th}(m)\right] + \frac{C'}{n}$$

where C is a numerical constant and C' depends on $(K, \overline{F}_0, \|h\|_\infty)$.

Proposition 4.7.2 *There exists a numerical constant C_2 such that, provided that $\alpha_n \leq \overline{F}_0/2$,*

$$P[\Delta_3^c] \leq 2\exp(-C_2 n\overline{F}_0).$$

Proposition 4.7.3 *Assume that condition (4.16) is satisfied, then*

$$P[\Delta_4^c \cap \Delta_1] \leq 4D\exp\left(-C\frac{n}{D}\right) \tag{4.25}$$

where C depends on $(\nu, \overline{F}_0, \|h\|_\infty, a)$.

4.7.1 Proof of Theorem 4.4.1

For every model m, similarly to Proposition 4.6.2,

$$\|\widehat{h}_m - h\|_{\overline{F}_T} \leq \frac{5K\sqrt{N_n}}{3\widehat{F}_0} + \|h\|_{\overline{F}_T} \leq \frac{5K\sqrt{N_n}}{3\alpha_n} + \|h\|_{\overline{F}_T} = \frac{5}{3}Kn + \|h\|_{\overline{F}_T}$$

since $\widehat{F}_0 \geq \alpha_n$ and $N_n \leq n$.

(1) Let $j = 1$.

$$\mathbb{E}\left[\|\widehat{h}_{\widehat{m}_1} - h\|^2_{\overline{F}_T} \mathbb{1}_{(\Delta_1\cap\Delta_2\cap\Delta_3)^c}\right] \leq \left(\frac{5}{3}Kn + \|h\|_{\overline{F}_T}\right)^2 P\left[(\Delta_1\cap\Delta_2\cap\Delta_3)^c\right].$$

Besides, according to Proposition 4.4.1, $\Delta_1 \cap \Delta_2 \cap \Delta_3 = \Delta_1 \cap \Delta_3$. Therefore, Propositions 4.6.3 and 4.7.2 entail

$$\begin{aligned} P\left[(\Delta_1 \cap \Delta_2 \cap \Delta_3)^c\right] &\leq P\left[\Delta_1^c\right] + P\left[\Delta_3^c\right] \\ &\leq 2N_n^2 \exp\left(-C_1 \overline{F}_0^2 \frac{n}{N_n}\right) + \exp\left(-C_2 \overline{F}_0 n\right). \end{aligned}$$

Thus similarly to (4.19),

$$\mathbb{E}\left[\|\widehat{h}_{m_1} - h\|_{\overline{F}_T}^2 \mathbb{1}_{(\Delta_1 \cap \Delta_2 \cap \Delta_3)^c}\right] \leq \frac{C'}{n}. \tag{4.26}$$

Proposition 4.7.1 and (4.26) conclude the proof of Theorem 4.4.1 for j=1.

(2) Let $j = 2$.

$$\begin{aligned} P\left[(\Delta_1 \cap \Delta_2 \cap \Delta_3 \cap \Delta_4)^c\right] &= P\left[(\Delta_1^c \cup \Delta_3^c \cup \Delta_4^c) \cap \Delta_1\right] + P\left[\Delta_1^c\right] \\ &\leq P\left[\Delta_1^c\right] + P\left[\Delta_3^c\right] + P\left[\Delta_4^c \cap \Delta_1\right] \end{aligned}$$

and Propositions 4.6.3, 4.7.2 and 4.7.3 allow to conclude similarly to the case $j = 1$. □

4.7.2 Proof of Proposition 4.7.1

We only expose the proof of (1) since the proof of (2) is very similar. The proof of Proposition 4.7.1 follows the same line as Proposition 4.6.1, let us point out the slight differences. Inequalities (4.20) and (4.22), as well as Lemma 4.6.1 hold. Hence, for every model m and every $h \in S_m$,

$$\mathbb{E}\left[\|\widehat{h}_{\widehat{m}_1} - h\|_{\overline{F}_T}^2 \mathbb{1}_{\Delta_1 \cap \Delta_2 \cap \Delta_3}\right] \leq C_1 \left\{\|h - h_m\|_{\overline{F}_T}^2 + \mathbb{E}\left[(pen_1(m) - pen_1(\widehat{m}_1) + 2p_1(m, \widehat{m}_1))\mathbb{1}_{\Delta_3}\right] + \right.$$
$$\left. + \|h\|_{\overline{F}_T}^2 \frac{1}{\overline{F}_0 n} + \sum_{m' \in \mathcal{M}_n} \mathbb{E}\left[\left(\sup_{t \in S_m + S_{m'}, \|t\|_{\overline{F}_T} = 1} (\nu_{n,1}(t))^2 - \frac{2}{3} p_1(m, m')\right)_+\right]\right\}$$

with

$$p_1(m, m') = \frac{2B}{5} \frac{K^2}{\overline{F}_0} \frac{D_m + D_{m'}}{n}.$$

The only difference with the proof of Proposition 4.6.1 is the upper bound of

$$\mathbb{E}\left[(pen_1(m) - pen_1(\widehat{m}_1) + 2p_1(m, \widehat{m}_1))\mathbb{1}_{\Delta_3}\right].$$

Indeed,

$$\begin{aligned}\mathbb{E}\left[(pen_1(m) - pen_1(\widehat{m}_1) + 2p_1(m,\widehat{m}_1))\mathbb{1}_{\Delta_3}\right] &= \mathbb{E}\left[\left(\frac{K^2B}{\widehat{F}_0}\frac{D_m - D_{\widehat{m}}}{n} + \frac{4K^2B}{5\overline{F}_0}\frac{D_m + D_{\widehat{m}}}{n}\right)\mathbb{1}_{\Delta_3}\right] \\ &\leq \mathbb{E}\left[\left(\frac{K^2B}{\widehat{F}_0}\frac{D_m - D_{\widehat{m}}}{n} + \frac{K^2B}{\widehat{F}_0}\frac{D_m + D_{\widehat{m}}}{n}\right)\mathbb{1}_{\Delta_3}\right] \\ &= \mathbb{E}\left[\frac{2K^2B}{\widehat{F}_0}\frac{D_m}{n}\mathbb{1}_{\Delta_3}\right] \leq \frac{8K^2B}{3\overline{F}_0}\frac{D_m}{n}. \quad \square\end{aligned}$$

4.7.3 Proof of Proposition 4.7.2

If $\alpha_n \leq \overline{F}_0/2$, then

$$P[\Delta_3^c] = P\left[\left|\frac{1}{n}\sum_{i=1}^{n}(\mathbb{1}_{\{T_i\geq 1\}} - \overline{F}_0)\right| \geq \frac{1}{4}\overline{F}_0\right] = P\left[\left|\frac{1}{n}\sum_{i=1}^{n}(\mathbb{1}_{\{T_i\geq 1\}} - \mathbb{E}[\mathbb{1}_{\{T_i\geq 1\}}])\right| \geq \frac{1}{4}\overline{F}_0\right]$$

We apply Bernstein Inequality (Theorem 1.2.4 with the parameters $c = 1$ and $v = \overline{F}_0$, then $P[\Delta_3^c] \leq 2\exp(-C_2 n\overline{F}_0)$ where C_1 is a numerical constant.

4.7.4 Proof of Proposition 4.7.3

Let x_0 and $\widehat{x}_0$ be in A such that

$$\nu = \|h\|_{\infty,A} = h(x_0) \quad \text{and} \quad \widehat{\nu}_n = \|\widehat{h}_D\|_{\infty,A} = \widehat{h}_D(\widehat{x}_0).$$

Then

$$\widehat{\nu}_n - \nu \leq (\widehat{h}_D - h)(\widehat{x}_0) = (\widehat{h}_D - h_D)(\widehat{x}_0) + (h_D - h)(\widehat{x}_0) \leq \sqrt{\frac{D}{a}}\sup_{j=1,\dots,D}|\widehat{a}_j^D - a_j^D| + \|h - h_D\|_\infty.$$

Similarly,

$$\nu - \widehat{\nu}_n \leq (h - \widehat{h}_D)(x_0) \leq (h - h_D)(x_0) + (h_D - \widehat{h}_D)(x_0) \leq \|h - h_D\|_\infty + \sqrt{\frac{D}{a}}\sup_{j=1,\dots,D}|\widehat{a}_j^D - a_j^D|.$$

Hence $|\nu - \widehat{\nu}_n| \leq \|h - h_D\|_\infty + \sqrt{D/a}\sup_{j=1,\dots,D}|\widehat{a}_j^D - a_j^D|$, and according to (4.16),

$$\begin{aligned}P[\Delta_4^c] &= P[|\nu - \widehat{\nu}_n| > \frac{1}{4}] \\ &\leq P\left[\|h - h_D\|_\infty + \sqrt{\frac{D}{a}}\sup_{j=1,\dots,D}|\widehat{a}_j^D - a_j^D| > \frac{\nu}{4}\right] \\ &\leq P\left[\sqrt{\frac{D}{a}}\sup_{j=1,\dots,D}|\widehat{a}_j^D - a_j^D| > \frac{\nu}{8}\right] \leq \sum_{j=1}^{D} P\left[\sqrt{\frac{D}{a}}|\widehat{a}_j^D - a_j^D| > \frac{\nu}{8}\right].\end{aligned}$$

Besides, for every $j = 1, \ldots, D$,

$$\begin{aligned}
\sqrt{\frac{D}{a}}(\widehat{a}_j^D - a_j^D) &= \sqrt{\frac{D}{a}}\left[\frac{(1/n)\sum_{i=1}^n \delta_i \varphi_j^D(T_i)}{\|\varphi_j^D\|_n^2} - \frac{\int_A \varphi_j^D(x)h(x)\overline{F}_T(x)dx}{\|\varphi_j^D\|_{\overline{F}_T(x)}^2}\right] \\
&= \left(\frac{1}{\|\varphi_j^D\|_n^2}\sqrt{\frac{D}{a}}\right)\frac{1}{n}\sum_{i=1}^n\left[\delta_i\varphi_j^D(T_i) - \int_A \varphi_j^D(x)h(x)\overline{F}_T(x)dx\right] \\
&\quad + \sqrt{\frac{D}{a}}\int_A \varphi_j^D(x)h(x)\overline{F}_T(x)dx\left[\frac{1}{\|\varphi_j^D\|_n^2} - \frac{1}{\|\varphi_j^D\|_{\overline{F}_T}^2}\right].
\end{aligned}$$

Moroever, on the set Δ_1,

$$\|\varphi_j^D\|_n^2 \geq \frac{3}{4}\|\varphi_j^D\|_{\overline{F}_T}^2 = \frac{3}{4}\int_{a(j-1)/D}^{aj/D}\frac{D}{a}\overline{F}_T(x)dx \geq \frac{3\overline{F}_0}{4}$$

and

$$\left|\int_A \varphi_j^D(x)\lambda(x)\overline{F}_T(x)dx\right| \leq \|h\|_{\overline{F}_T}\|\varphi_j^D\|_{\overline{F}_T} \leq \|h\|_{\overline{F}_T}.$$

Hence

$$\begin{aligned}
&\sqrt{\frac{D}{a}}|\widehat{a}_j^D - a_j^D|\mathbb{1}_{\Delta_1} \\
&\leq \frac{4}{3\overline{F}_0}\sqrt{\frac{D}{a}}\left|\frac{1}{n}\sum_{i=1}^n \delta_i\varphi_j^D(T_i) - \int_A \varphi_j^D(x)h(x)\overline{F}_T(x)dx\right| + \sqrt{\frac{D}{a}}\|h\|_{\overline{F}_T}\left|\frac{\|\varphi_j^D\|_{\overline{F}_T}^2 - \|\varphi_j^D\|_n^2}{\|\varphi_j^D\|_{\overline{F}_T}^2\|\varphi_j^D\|_n^2}\right| \\
&\leq \frac{4}{3\overline{F}_0}\sqrt{\frac{D}{a}}\left|\frac{1}{n}\sum_{i=1}^n(\delta_i\varphi_j^D(T_i) - \mathbb{E}[\delta_i\varphi_j^D(T_i)])\right| + \sqrt{\frac{D}{a}}\|h\|_{\overline{F}_T}\frac{4^2}{3^2\overline{F}_0^2}\left|\|\varphi_j^D\|_{\overline{F}_T}^2 - \|\varphi_j^D\|_n^2\right|.
\end{aligned}$$

Thus

$$\begin{aligned}
P[\Delta_4^c \cap \Delta_1] &\leq \sum_{j=1}^D P\left[\frac{4}{3\overline{F}_0}\sqrt{\frac{D}{a}}\left|\frac{1}{n}\sum_{i=1}^n(\delta_i\varphi_j^D(T_i) - \mathbb{E}[\delta_i\varphi_j^D(T_i)])\right| \geq \frac{\nu}{16}\right] \\
&\quad + \sum_{j=1}^D P\left[\sqrt{\frac{D}{a}}\|h\|_{\overline{F}_T}\frac{4^2}{3^2\overline{F}_0^2}\left|\|\varphi_j^D\|_{\overline{F}_T}^2 - \|\varphi_j^D\|_n^2\right| \geq \frac{\nu}{16}\right] \\
&= \sum_{j=1}^D (P_{1,j} + P_{2,j}). \qquad (4.27)
\end{aligned}$$

$P_{1,j}$ and $P_{2,j}$ are upper bounded with Bernstein Inequality (Theorem 1.2.4). For $P_{1,j}$, the parameters b and v are the following.

$$\mathbb{E}[\delta_i^2(\varphi_j^D)^2(T_i)] = \int_A (\varphi_j^D)^2(x)h(x)\overline{F}_T(x)dx \leq \|h\|_{\infty,A} = v \quad \text{and} \quad \|\delta_i\varphi_j^D(T_i)\|_\infty \leq \sqrt{\frac{D}{a}} = c.$$

Hence, for every $j \in \{1, \dots, D\}$,

$$P_{1,j} \leq 2\exp\left(-C\frac{n}{D}\right) \tag{4.28}$$

where C depends on $(\nu, \|h\|_{\infty,A}, \overline{F}_0, a)$.

Let us upper bound $P_{2,j}$. For every $j \in \{1, \dots, D\}$,

$$P_{2,j} = P\left[\frac{4^2\|h\|_{\overline{F}_T}}{3^2\overline{F}_0^2}\sqrt{\frac{D}{a}}\left|\frac{1}{n}\sum_{i=1}^n\left(\int_A(\varphi_j^D)^2(x)\mathbb{1}_{\{T_i\geq x\}}dx - \mathbb{E}\left[\int_A(\varphi_j^D)^2(x)\mathbb{1}_{\{T_i\geq x\}}dx\right]\right)\right| \geq \frac{\nu}{16}\right]$$

and

$$\mathbb{E}\left[\left(\int_A(\varphi_j^D)^2(x)\mathbb{1}_{\{T_i\geq x\}}dx\right)^2\right] \leq 1 = v \quad \text{and} \quad \left\|\int_A(\varphi_j^D)^2(x)\mathbb{1}_{\{T_i\geq x\}}dx\right\|_\infty \leq 1 = c.$$

Thus, with Bernstein Inequality we obtain

$$P_{2,j} \leq 2\exp\left(-C'\frac{n}{D}\right) \tag{4.29}$$

where C' depends on $(\nu, \|h\|_{\overline{F}_T}, \overline{F}_0, a)$. Then (4.27), (4.28) and (4.29) conclude the proof of Proposition 4.7.3. □

4.8 Appendix

Lemma 4.8.1 *Let M be a symmetric matrix of dimension n, with real coefficients, then*

$$\min(Sp(M)) = \min_{\{U\in\mathbb{R}^n, U\neq 0\}} \frac{U^tMU}{U^tU}.$$

Proof of Lemma 4.8.1**.** M is a real symmetric matrix hence according to classical algebra results, there exist an orthogonal matrix P and a diagonal matrix D such that $M = P^tDP$. Moreover, $D = \text{diag}(h_1, \dots, h_n)$ where h_j is an eigenvalue of M for every $j \in \{1, \dots, n\}$. Let $j_0 = \arg\min_{j=1,\dots n} h_j$.

On the one hand, let $U \in \mathbb{R}^n$, and $V = PU = (v_1, \ldots, v_n)^t$, then

$$U^t MU = (PU)^t D(PU) = \sum_{j=1}^{n} h_j v_j^2 \geq \min(Sp(M)) \sum_{j=1}^{n} v_j^2 = \min(Sp(M)) V^t V.$$

Thus $\min(Sp(M)) \leq \min_{\{U \in \mathbb{R}^n, U \neq 0\}} \frac{U^t MU}{U^t U}$.

On the other hand, Let V_0 be the vector whose coordinates are all zero except the j_0^{th} which is equal to 1. Let $U_0 = P^{-1} V_0$, then $U_0^t U_0 = V_0^t V_0 = 1$ since P is orthogonal.

$$U_0^t MU_0 = V_0^t DV_0 = h_{j_0} = \min(Sp(M)) U_0^t U_0$$

hence $\min(Sp(M)) = \min_{\{U \in \mathbb{R}^n, U \neq 0\}} \frac{U^t MU}{U^t U}$. $\square$

Chapitre 5

Généralisation de la méthode de sélection de modèle ponctuelle : application à l'estimation ponctuelle du risque instantané à partir de données censurées à droite

Ce chapitre présente une généralisation de la méthode de sélection de modèle ponctuelle developpée au Chapitre 3, à d'autres cadre que celui de l'estimation de densité. Un résultat général est tout d'abord énoncé, en dégageant les principales étapes de la majoration du risque de l'estimateur ponctuel de densité : à partir d'une collection $\{\widehat{g}_m\}$ d'estimateurs non adaptatifs, et d'une collection de fonctions $\{g_m\}$ (généralement les projections de la fonction cible sur les modèles), nous développons une procédure de sélection de modèle ponctuelle. Elle conduit à une inégalité presque-oracle sous certaines conditions, la principale portant sur la majoration de probabilités de déviation des termes $(\widehat{g}_m - g_m)^2(x_0)$.

Ce résultat est ensuite appliqué aux estimateurs non adaptatifs $\{\widehat{h}_m\}$ du taux de risque instantané construits au Chapitre 4 à partir d'un contraste de type regression : le vecteur $\widehat{A}_m$ des coefficients de $\widehat{h}_m$ dans une base $(\phi_1, \ldots, \phi_{D_m})$ vérifie une relation de la forme

$$\widehat{G}_m \widehat{A}_m = \widehat{V}_m,$$

où $\widehat{G}_m$ est la matrice de Gram de $(\phi_1, \ldots, \phi_{D_m})$ pour une norme empirique dépendant des observations. Afin d'avoir une expression manipulable des coefficients de $\widehat{h}_m$ et de majorer la probabilité de déviation de $(\widehat{h}_m - h_m)^2(x_0)$, nous considérons une collection de modèles constituée de polynômes par morceaux et les matrices $\widehat{G}_m$ sont alors diagonales par blocs. Ainsi, l'application du résultat général de sélection de modèle ponctuelle fournit un estimateur adaptatif du taux de risque instantané pour le risque quadratique ponctuel.

5.1 Introduction

In Chapter 3, we have developed a model selection procedure which provides an adaptive density estimator for the pointwise risk. In this chapter, we generalise the procedure by enhancing the main steps which are not specifically related to density estimation framework and state a general result which can fit other frameworks. Thus, we propose a general pointwise model selection procedure which enables to reach adaptive minimax rate of convergence.

In the density estimation framework, we have built a collection of estimators $\{\widehat{g}_m, m \in J_n\}$ by projection on a collection of linear models $\mathcal{M}_n = \{S_m, m \in J_n\}$. Then for a given point x_0 in the support of g, a model $\widehat{m}$ was selected by minimising an empirical criterion:

$$\widehat{Crit}(m) = \sup_{j \in J_n, D_j \geq D_m} \left[(\widehat{g}_j - \widehat{g}_m)^2(x_0) - \frac{x_j D_j + x_m D_m}{n} \right]_+ + B x_m \frac{D_m}{n}$$

where the $\{D_m\}$'s are quantities related to the models $\{S_m\}$ (most of the time, D_m is the dimension of S_m), and the $\{x_m\}$'s are positive weights of order $\log D_m$.

The model selection estimator $\widehat{g}_{\widehat{m}}$ satisfies an inequality close to an oracle. Indeed, the pointwise risk $\mathbb{E}[(\widehat{g}_{\widehat{m}} - g)^2(x_0)]$ has the order of the minimum of the risks of the non adaptive estimators $\{\widehat{g}_m\}$ for $m \in J_n$ (see Theorem 3.3.1). The proof of this result is based on the control of the deviation between the empirical criterion $\widehat{Crit}(m)$ and its non empirical counterpart:

$$Crit(m) = \sup_{j \in J_n, D_j \geq D_m} (g_j - g_m)^2(x_0) + B x_m \frac{D_m}{n}$$

where g_m is the L^2-projection of g on S_m. This deviation can be expressed in terms of deviation between $\widehat{g}_m$ and g_m. More precisely, the key point of the proof is the control of the following term

$$P\left[(\widehat{g}_m - g_m)^2(x_0) \geq C_0 \left(x + x_m \frac{D_m}{n} \right) \right] \tag{5.1}$$

via Bernstein Inequality.

This result can be extended to estimate an application g in a different framework, and get an adaptive estimator for the pointwise risk. We assume that we have a collection $\{\widehat{g}_m, m \in J_n\}$ of estimators of g, and a collection $\{(D_m, x_m), m \in J_n\}$ of integers and positive numbers. Moreover, we suppose that there exists a collection of unobserved applications $\{g_m, m \in J_n\}$ such that the probability (5.1) is small enough. Then the estimator $\widehat{g}_{\widehat{m}}$, where $\widehat{m}$ is the minimiser of $\widehat{Crit}(m)$ over J_n, satisfies a nearly-oracle inequality.

The procedure is applied to the hazard rate estimation in presence of right censoring from the non adaptive estimators defined in Chapter 4. Consider a sample $(Y_1, \ldots, Y_n)$ of i.i.d. positive random variables of common density f_Y and common survival function $\overline{F}_Y$, and a

compact A on which $\overline{F}_Y$ is lower bounded by a positive number, the hazard rate of Y is

$$h(x) = \frac{f_Y(x)}{\overline{F}_Y(x)}, \quad \forall x \in A.$$

We assume that the data are right-censored: there exists a sample $(C_1, \dots, C_n)$ of i.i.d. positive variables independent of the $\{Y_i\}$'s and we only observe the sample

$$(T_i = \min(Y_i, C_i), \delta_i = \mathbb{1}_{\{Y_i \leq C_i\}})_{i=1,\dots,n}. \tag{5.2}$$

In Chapter 4, a collection of non adaptive estimators $\{\widehat{h}_m, m \in J_n\}$ is built from the sample (5.2) by minimisation of a regression-type contrast on piecewise polynomial models. In the present chapter, the deviation (5.1), where the $\{h_m\}$'s are the orthogonal projection on models for a well choosen norm, is upper bounded with Bernstein Inequality and the general result we prove that the pointwise model selection estimator satisfies a nearly-oracle inequality. Moreover, the control of the bias term on Hölder provides a rate of convergence which should be the adaptive minimax rate, even if no minimax study is available in this context for the pointwise risk.

In the litterature, the hazard rate estimators adapted to the pointwise risk are generally based on kernels. The first hazard rate kernel estimator introduced by Leadbetter and Watson (1964) has been widely studied in the litterature (see for example Tanner and Wong (1983)). Moreover, the performance of basic kernel estimator have been improved by developping variable bandwidth estimators. Patil (1993) proposes a cross validation method to select the bandwidth. Nielsen (2003) carries out a numerical comparison of nine bandwidth estimators, built directly as function of the data or from a pilot estimator, that is a hazard rate estimator based on a non-variable bandwidth kernel, and one of them is more precisely studied by Bagkavos and Patil (2009).

The papers is organised as follows. Section 5.2 focuses on the general procedure and states the main result. This result is applied in Section 5.3 to hazard rate estimation. Section 5.4 is devoted to the proof of the general result, and Section 5.5 to the proofs of Section 5.3. Section 5.6 presents properties of piecewise polynomials which are used in Section 5.3.

5.2 Generalisation of the pointwise model selection procedure

5.2.1 Procedure

Let g be a function defined on an interval A, and x_0 a fixed point in A. Assume that we have

1. A collection $\{\widehat{g}_m, m \in J_n\}$ of estimators of g.

2. A collection $\{g_m, m \in J_n\}$ of applications (which can be the projection of g on subspaces of $L^2(A)$).

3. A collection $\{D_m, m \in J_n\}$ of positive integers, and a collection $\{x_m, m \in J_n\}$ of positive numbers which depend on the parameters of the problem and such that

$$D_j \leq D_{j'} \Rightarrow x_j \leq x_{j'}. \tag{5.3}$$

For every $m \in J_n$, let

$$Crit(m) = \sup_{j \in J_n, D_j \geq D_m} (g_j - g_m)^2(x_0) + B x_m \frac{D_m}{n} \tag{5.4}$$

$$\widehat{Crit}(m) = \sup_{j \in J_n, D_j \geq D_m} \left[(\widehat{g}_j - \widehat{g}_m)^2(x_0) - \frac{x_j D_j + x_m D_m}{n} \right]_+ + B x_m \frac{D_m}{n} \tag{5.5}$$

where $B > 1/2$. Then we set

$$m_{opt} = \arg\min_{m \in \mathcal{M}_n} Crit(m), \quad \widehat{m} = \arg\min_{m \in \mathcal{M}_n} \widehat{Crit}(m).$$

We assume that the following assumption holds.

- (**H**) There exist a set Δ and a function $F : \mathbb{R} \times \mathbb{N} \times J_n \times \mathbb{R}_+^* \to \mathbb{R}^+$ such that, for every $m \in J_n$, $x > 0$, and $a > 0$,

$$P\left[\left\{ |(\widehat{g}_m - g_m)(x_0)| \geq a \sqrt{x + x_m \frac{D_m}{n}} \right\} \cap \Delta \right] \leq F(x, m, n, a).$$

and there exist a numerical constant $0 < B' < 1/\sqrt{2}$ and a constant B_0 depending on the parameters of the problem such that, for every $n \in \mathbb{N}^*$

$$\sum_{m \in \mathcal{M}_n} \int_0^\infty F(x, m, n, B') dx \leq \frac{B_0}{n}. \tag{5.6}$$

Remark

1. The constant B' for which inequality (5.6) holds has to be specified since in applications, these constants are involved in the definition of the weights (x_m), as we will establish in Section 5.3.

2. Proposition 5.2.1 would hold with any constant $B > 0$ in (5.4) and (5.5), but we assume that $B \geq 1/2$ by sake of simplicity.

5.2.2 Result

The model selection estimator $\widehat{g}_{\widehat{m}}$ satisfies the following result.

Proposition 5.2.1 *Under Assumption* (**H**)*, there exist a constant* κ *depending on* B *and* B' *and a constant* κ' *depending on* B *and* B_0 *such that*

$$\mathbb{E}[(\widehat{g}_{\widehat{m}} - g)^2(x_0)\mathbb{1}_{\Delta}] \leq \kappa \left(Crit(m_{opt}) + (g - g_{m_{opt}})^2(x_0) \right) + \frac{\kappa'}{n}.$$

Comment. Similarly to Chapter 3, $(g - g_{m_{opt}})^2(x_0)$ and $\sup_{j \geq m_{opt}}(g_j - g_{m_{opt}})^2(x_0)$ have same order, thus the result of Proposition 5.2.1 is nearly an oracle inequality. Indeed, $\mathbb{E}[(\widehat{g}_{\widehat{m}} - g)^2(x_0)\mathbb{1}_{\Delta}]$ has the order of $\inf_{m \in J_n} Crit(m)$, which has the order of

$$\inf_{m \in J_n} \left\{ (g - g_m)^2(x_0) + B x_m \frac{D_m}{n} \right\}.$$

5.3 Pointwise adaptive estimation of the hazard rate in presence of right censoring

The framework and the non adaptive estimation procedure are identical to Chapter 4. We only recall the definition here and the reader is refered to Chapter 4 for more details.

5.3.1 Framework and notations

We consider the sample defined in (5.2) and a compact interval A on which $\overline{F}_T = \overline{F}_C \overline{F}_Y$ is lower bounded. Let $x_0 \in A$ be fixed; For sake of simplicity A is assumed equal to $[0, 1]$. We consider the following assumption.

> $\mathbf{A_{frame}}$: We assume that $\overline{F}_T$ is lower bounded on $A = [0, 1]$ by $\overline{F}_0 > 0$, and h is upper bounded by $\|h\|_{\infty,A} = \sup_{x \in A} h(x) < +\infty$. The quantities $\|h\|_{\infty,A}$ and $\overline{F}_0$ are supposed to be known.

Remark As in Chapter 4, the parameters $\|h\|_{\infty,A}$ and $\overline{F}_0$ could be replaced by estimators.

We define the following scalar products and norms on $L^2(A)$. For every $s, t \in L^2(A)$,

$$\langle s, t \rangle_{\overline{F}_T} = \int_A s(x)t(x)\overline{F}_T(x)dx, \quad \|t\|^2_{\overline{F}_T} = \int_A t^2(x)\overline{F}_T(x)dx$$

$$\langle s, t \rangle_n = \frac{1}{n}\sum_{i=1}^n \int_A s(x)t(x)\mathbb{1}_{T_i \geq x}dx, \quad \|t\|^2_n = \frac{1}{n}\sum_{i=1}^n \int_A t^2(x)\mathbb{1}_{T_i \geq x}dx.$$

Let M be a square matrix, we denote by $Sp(M)$ the spectrum of M i.e. the set of its eigenvalues.

Let β and L be positive numbers, and r the greatest integer smaller than β, we define the Hölder space $\mathcal{H}(\beta, L)$ on A:

$$\mathcal{H}(\beta, L) = \{f : A \to \mathbb{R}, |f^{(r)}(x) - f^{(r)}(y)| \leq L|x - y|^{\beta - r}, \forall x, y \in A\}.$$

For every $x \in \mathbb{R}$, we denote by $\lfloor x \rfloor$ the integer part of x, that is the greatest integer smaller than or equal to x.

5.3.2 Collection of estimators

We build a collection of estimators of h by minimisation of a regression-type contrast following the procedure of Chapter 4. We consider the collection of piecewise polynomials of degree smaller than or equal to a fixed positive number r. More precisely, let $(\phi^0, \ldots, \phi^r)$ be a $\|.\|$-orthonormal basis of polynomials on $[0, 1]$ which satisfies $deg(\phi^j) = j$ for every $j = 0, \ldots, r$ (an example of such family is given in Section 5.6.1). For every $m \in \mathbb{N}^*$ and $k \in \{1, \ldots, m\}$, let

$$\phi^j_{k,m}(x) = \sqrt{D_m}\phi^j(D_m x - (k - 1)) \tag{5.7}$$

with $D_m = m$. For every j, $\phi^j_{k,m}$ is supported on $I_{k,m} = [(k - 1)/D_m, k/D_m[$ and

$$S_{k,m} = Vect\{\phi^j_{k,m}, j = 0, \ldots, r\}$$

is the set of polynomials of degree smaller than or equal to r on $I_{k,m}$. Now for every $m \in \mathbb{N}^*$, we define

$$S_m = \oplus_{k=1}^{D_m} S_{k,m}$$

and consider the collection of models

$$\mathcal{M}_n = \{S_m, m = 1, \ldots, N_n\}$$

where $N_n \leq n/\log^2 n$. We note that for every constant $c > 0$, there exists a constant $C > 0$ such that

$$\sum_{m=1}^{N_n} \exp(-cD_m) \leq C. \tag{5.8}$$

To simplify notations, we denote by

$$J_m = ((0, 1, m), \ldots, (r, 1, m), (0, 2, m), \ldots, (j, k, m), \ldots, (r, D_m, m))$$

and $\phi_\lambda = \phi^j_{k,m}$ for every $\lambda = (j, k, m) \in J_m$.

Let us define the collection of estimators built on the collection $\mathcal{M}_n$. For a detailed heuristic, the reader is referred to Chapter 4, Section 4.3.1. For every $m \in \{1, \dots, N_n\}$, let

$$\widehat{G}_m = (\langle \phi_\lambda, \phi_{\lambda'} \rangle_n)_{\lambda,\lambda' \in J_m}, \quad \widehat{V}_m = \left(\frac{1}{n} \sum_{i=1}^{n} \delta_i \phi_\lambda(T_i) \right)_{\lambda \in J_m}$$

and

$$\Delta_m = \left\{ \min(Sp(\widehat{G}_m)) \geq \frac{3}{4} \overline{F}_0 \right\}. \tag{5.9}$$

The matrix $\widehat{G}_m$ is invertible on Δ_m (see Chapter 4), and we set

$$\widehat{h}_m = \sum_{\lambda \in J_m} \widehat{a}_\lambda \phi_\lambda = \sum_{k=1}^{D_m} \sum_{j=0}^{r} \widehat{a}_{k,m}^j \phi_{k,m}^j$$

where

$$\widehat{A}_m = (\widehat{a}_\lambda)_{\lambda \in J_m} = \begin{cases} \widehat{G}_m^{-1} \widehat{V}_m & \text{on } \Delta_m \\ 0 & \text{otherwise .} \end{cases}$$

Besides, for every $m \in \{1, \dots, N_n\}$, let h_m be the $\|.\|_{\overline{F}_T}$-projection of h on S_m. Then

$$h_m = \sum_{\lambda \in J_m} a_\lambda \phi_\lambda = \sum_{k=1}^{D_m} \sum_{j=0}^{r} a_{k,m}^j \phi_{k,m}^j$$

where

$$A_m = (a_\lambda)_{\lambda \in J_m} = G_m^{-1} V_m$$

and

$$G_m = \left(\langle \phi_\lambda, \phi_{\lambda'} \rangle_{\overline{F}_T} \right)_{\lambda,\lambda' \in J_m}, \quad V_m = \left(\langle \phi_\lambda, h \rangle_{\overline{F}_T} \right)_{\lambda \in J_m}.$$

G_m is the Gram matrix of the $\|.\|$-orthonormal basis (ϕ_λ) for the scalar product $\langle ., . \rangle_{\overline{F}_T}$, and the norms $\|.\|$ and $\|.\|_{\overline{F}_T}$ are equivalent so G_m is invertible. Note that

$$\mathbb{E}[\widehat{G}_m] = G_m \quad \text{and} \quad \mathbb{E}[\widehat{V}_m] = V_m.$$

On can note that, contrary to Chapter 4 where a general collection $\mathcal{M}_n$ can be considered, we restrict ourselves to piecewise polynomial models. Let us justify this choice. In order to upper bound the term

$$P\left[(\widehat{h}_m - h)^2(x_0) \geq a_0^2 \left(x + x_m \frac{D_m}{n} \right) \right]$$

with Bernstein Inequality (Theorem 1.2.4 in Introduction), we need to upper bound the variance term

$$\mathbb{E}[(\widehat{h}_m - h)^2(x_0)]. \tag{5.10}$$

Besides, we note that if the model S_m consists of histograms, the Gram matrix $\widehat{G}_m$ and G_m are diagonal and (5.10) has a simple expression. Nevertheless, histogram models can not enable us to reach minimax rate for regularity greater than 1. To preserve an explicit expression of (5.10), we consider models such that the matrix $\widehat{G}_m$ and G_m are block-diagonal with block dimension independent of n, which is guaranteed for our collection, since the models satisfy the strong localisation property (see Introduction, Section 1.2.5). Indeed, for every $k \neq k'$, and every $\lambda \in K_{k,m}$, $\lambda' \in K_{k',m}$,

$$\langle \phi_\lambda, \phi_{\lambda'} \rangle_n = 0 \quad \text{and} \quad \langle \phi_\lambda, \phi_{\lambda'} \rangle_{\overline{F}_T} = 0.$$

More precisely,

$$G_m \begin{pmatrix} G_m^{(1)} & & & \\ & . & (0) & \\ & & . & \\ & (0) & & G_m^{(D_m)} \end{pmatrix} \quad \text{with} \quad G_m^{(k)} = \left(\langle \phi_{k,m}^j, \phi_{k,m}^l \rangle_{\overline{F}_T} \right)_{j,l=0,\dots,r}$$

$$\widehat{G}_m \begin{pmatrix} \widehat{G}_m^{(1)} & & & \\ & . & (0) & \\ & & . & \\ & (0) & & \widehat{G}_m^{(D_m)} \end{pmatrix} \quad \text{with} \quad \widehat{G}_m^{(k)} = \left(\langle \phi_{k,m}^j, \phi_{k,m}^l \rangle_n \right)_{j,l=0,\dots,r}$$

Similarly, we decompose V_m and $\widehat{V}_m$ in blocks.

$$V_m = \begin{pmatrix} V_m^{(1)} \\ . \\ V_m^{(D_m)} \end{pmatrix} \quad \text{where} \quad V_m^{(l)} = \begin{pmatrix} \langle \phi_{l,m}^0, h \rangle_{\overline{F}_T} \\ . \\ \langle \phi_{l,m}^r, h \rangle_{\overline{F}_T} \end{pmatrix}$$

$$\widehat{V}_m = \begin{pmatrix} \widehat{V}_m^{(1)} \\ . \\ \widehat{V}_m^{(D_m)} \end{pmatrix} \quad \text{where} \quad \widehat{V}_m^{(l)} = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n \delta_i \phi_{l,m}^0(T_i) \\ . \\ \frac{1}{n} \sum_{i=1}^n \delta_i \phi_{l,m}^r(T_i) \end{pmatrix}$$

Let $k_0 \in \{1, \dots, m\}$ be such that $x_0 \in I_{k_0,m} = [(k_0 - 1)/D_m, k_0/D_m[$, then

$$h_m(x_0) = \sum_{j=0}^{r} a^j_{k_0,m} \phi^j_{k_0,m}(x_0)$$
$$\widehat{h}_m(x_0) = \sum_{j=0}^{r} \widehat{a}^j_{k_0,m} \phi^j_{k_0,m}(x_0).$$

The expansions of $h_m(x_0)$ and $\widehat{h}_m(x_0)$ involve a fixed number of coefficients as n grows, which is a key point to control the deviation between $\widehat{h}_m(x_0)$ and $h_m(x_0)$. The following Proposition controls the bias term $(h_m - h)^2(x_0)$ on Hölder spaces.

Proposition 5.3.1 *Let (L, β) be two positive numbers. There exist a constant L' which depends on (L, r, β) such that for every $h \in \mathcal{H}(\beta, L)$ and for every $m \in \{1, \ldots, N_n\}$*

$$\|h - h_m\|_\infty \leq L' D_m^{-\beta}$$

We note that h_m is the $\|.\|_{\overline{F}_T}$-projection of h on S_m, hence Proposition 5.6.1 (Section 5.6) which states a result about projection on piecewise polynomials leads to Proposition 5.3.1.

5.3.3 Pointwise model selection procedure

For every $m \in \{1, \ldots, N_n\}$, let

$$Crit(m) = \sup_{j \geq m} (h_j - h_m)^2(x_0) + B x_m \frac{D_m}{n}$$

$$\widehat{Crit}(m) = \sup_{j \geq m} \left[(\widehat{h}_j - \widehat{h}_m)^2(x_0) - \frac{x_j D_j + x_m D_m}{n} \right]_+ + B x_m \frac{D_m}{n}$$

where

$$x_m = 4B'(r+1)\mathcal{A}_1^2 \log D_m \max\left\{1, 9 \log D_m \frac{D_m}{n}\right\} \tag{5.11}$$

$$\mathcal{A}_1^2 = \left(\frac{4}{3\overline{F}_0}\right)^2 \left(\sum_{j=0}^{r} \|\phi^j\|_\infty^2\right).$$

and B and B' are numerical constants with $B \geq 1/2$ and $B' \geq 2$. Let

$$m_{opt} = \arg \min_{m=1,\ldots,N_n} Crit(m) \quad \text{and} \quad \widehat{m} = \arg \min_{m=1,\ldots,N_n} \widehat{Crit}(m).$$

5.3.4 Results

The model selection estimator $\widehat{h}_{\widehat{m}}$ satisfies the following theorem.

Theorem 5.3.1 *Assume that* $\mathbf{A_{frame}}$ *holds. There exist a numerical constant* κ *and a constant* κ' *which depends on* $\overline{F}_0$, $\|h\|_\infty$ *and on the basis* $\{\phi_j\}$ *such that*

$$\mathbb{E}[(\widehat{h}_{\widehat{m}} - h)^2(x_0)] \leq \kappa \left(Crit(m_{opt}) + (h - h_{m_{opt}})^2(x_0) \right) + \frac{\kappa'}{n}. \tag{5.12}$$

Theorem 5.3.1 states a nearly-oracle inequality for the pointwise risk of h (see Comment after Proposition 5.2.1). Moreover with Proposition 5.3.1, the right hand side in (5.12) is upper bounded on Hölder spaces.

Corollary 5.3.1 *Assume that* $\mathbf{A_{frame}}$ *holds and* $h \in \mathcal{H}(\beta, L)$ *for some positive numbers* β, L, *then there exists a constant* B_2 *which depends on* $(\overline{F}_0, \|h\|_\infty, L, \beta, r)$ *and on the basis* $\{\phi_j\}$ *such that*

$$\mathbb{E}[(\widehat{h}_{\widehat{m}} - h)^2(x_0)] \leq B_2 \left(\frac{n}{\log n} \right)^{-2\beta/(2\beta+1)}.$$

Comment The minimax rate of convergence for the hazard rate is known for the integrated risk, but no result is proved for the pointwise risk. Nevertheless, several papers underline the similarities between density and hazard rate estimation (see e.g. Diehl and Stute (1988) or Antoniadis et al. (1999)). In particular, in each of these papers the authors develop a density or subdensity estimator and divide it by an empirical estimator of survival function to form an estimator of h. Then they prove that the hazard rate and the density estimators converge at the same rate since the survival function is estimated at rate $1/n$. Thus, we can presume that these similarities still hold for lower bounds. If we refer to the results of Butucea (2001) about pointwise density estimation, we can speculate that the pointwise minimax rate of convergence for h on the Hölder space $\mathcal{H}(\beta, L)$ is $n^{-2\beta/(2\beta+1)}$, and the adaptive minimax rate is $(n/\log n)^{-2\beta/(2\beta+1)}$.

5.4 Proof of Proposition 5.2.1

We denote by $P_1[.] = P[. \cap \Delta]$ and $\mathbb{E}_1[.] = \mathbb{E}[.\mathbb{1}_\Delta]$.

The proof of Proposition 5.2.1 is based on the following decomposition.

$$\mathbb{E}_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0)] \leq \mathbb{E}_1[((\widehat{g}_{\widehat{m}} - g)^2(x_0) - \mathcal{U}_{opt})_+] + \mathbb{E}_1[\mathcal{U}_{opt}] \tag{5.13}$$

where $\mathcal{U}_{opt}$ is chosen to satisfy

1. $\mathbb{E}_1[\mathcal{U}_{opt}]$ has same order as $Crit(m_{opt})$.

2.

$$\mathbb{E}_1[((\widehat{g}_{\widehat{m}} - g)^2(x_0) - \mathcal{U}_{opt})_+] = 2\int_0^{+\infty} P_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0) - \mathcal{U}_{opt} \geq 2x]dx \leq \frac{C}{n} \qquad (5.14)$$

Consider the following Lemma.

Lemma 5.4.1 *Let δ and x be some positive numbers*

$$(1)\quad P_1\left[\left\{(\widehat{g}_{\widehat{m}} - g)^2(x_0) \geq (1+\delta)\left(\sup_{j\in J_n, D_j \geq D_{m_{opt}}} (g_j - g)^2(x_0) + Crit(m_{opt})\right) + x\right\} \cap \{D_{\widehat{m}} > D_{m_{opt}}\right\}\Bigg]$$
$$\leq F\left(\frac{x}{4}, m_{opt}, n, C_\delta\right) + \sum_{j\in J_n, D_j \geq D_{m_{opt}}} F\left(\frac{x}{4}, j, n, C_\delta\right) + \sum_{m\in J_n} F\left(\frac{x}{2B}, m, n, C_\delta\right).$$

$$(2)\quad P_1[\{(\widehat{g}_{\widehat{m}} - g)^2(x_0) \geq 2(1+\delta)Crit(m_{opt}) + 4\frac{x_{m_{opt}} D_{m_{opt}}}{n}$$
$$+2(\widehat{g}_{m_{opt}} - g)^2(x_0) + 2x\} \cap \{D_{\widehat{m}} \leq D_{m_{opt}}\}] \leq F\left(\frac{x}{2}, m_{opt}, n, C_\delta\right) + \sum_{j\in J_n, D_j \geq D_{m_{opt}}} F\left(\frac{x}{2}, j, n, C_\delta\right).$$

where $C_\delta = (2(1+1/\delta))^{-1/2}$.

Let $\delta > 0$ be such that $C_\delta = (2(1+1/\delta))^{-1/2} = B'$. Let

$$\mathcal{U}_{opt} = 2(\widehat{g}_{m_{opt}} - g)^2(x_0) + 2(1+\delta)Crit(m_{opt}) + 4\frac{x_{m_{opt}} D_{m_{opt}}}{n}$$
$$+(1+\delta)\sup_{j\in J_n, D_j \geq D_{m_{opt}}} (g_j - g)^2(x_0)$$

On the one hand, according to (1) and (2) in Lemma 5.4.1,

$$P_1[\{(\widehat{g}_{\widehat{m}} - g)^2(x_0) \geq \mathcal{U}_{opt} + 2x\} \cap \Delta] \leq F\left(\frac{x}{2}, m_{opt}, n, B'\right)$$
$$+2\sum_{j\in J_n, D_j \geq D_{m_{opt}}} F\left(\frac{x}{2}, j, n, B'\right) + 2\sum_{m\in J_n} F\left(\frac{x}{B}, m, n, B'\right).$$

Moreover,

$$\mathbb{E}_1[((\widehat{g}_{\widehat{m}} - g)^2(x_0) - \mathcal{U}_{opt})_+] = 2\int_0^{+\infty} P_1[(\widehat{g}_{\widehat{m}} - g)^2(x_0) \geq \mathcal{U}_{opt} + 2x]dx$$

$$\leq 4\int_0^{+\infty} F\left(\frac{x}{2}, m_{opt}, n, B'\right)dx + \sum_{j\in J_n, D_j \geq D_{m_{opt}}} 4\int_0^{+\infty} F\left(\frac{x}{2}, j, n, B'\right)dx + 2\sum_{m\in J_n}\int_0^{+\infty} F\left(\frac{x}{B}, m, n, B'\right)dx$$

$$= 8\int_0^{+\infty} F(x, m_{opt}, n, B')dx + 8\sum_{j\in J_n, D_j \geq D_{m_{opt}}}\int_0^{+\infty} F(x, j, n, B')dx + 2B\sum_{m\in J_n}\int_0^{+\infty} F(x, m, n, B')dx$$

$$\leq \frac{2B_0(8+B)}{n}.$$

according to Assumption (5.6). On the other hand,

$$\begin{aligned}\mathbb{E}_1[\mathcal{U}_{opt}] &= 2\mathbb{E}_1[(\widehat{g}_{m_{opt}} - g)^2(x_0)] + \left(2(1+\delta)Crit(m_{opt}) + 4\frac{x_{m_{opt}}D_{m_{opt}}}{n}\right.\\ &\quad \left. +(1+\delta)\sup_{j\in J_n, D_j \geq D_{m_{opt}}}(g_j - g_{m_{opt}})^2(x_0)\right).\end{aligned}$$

Moreover,

$$\mathbb{E}_1[(\widehat{g}_{m_{opt}} - g)^2(x_0)] \leq 2\mathbb{E}_1[(\widehat{g}_{m_{opt}} - g_{m_{opt}})^2(x_0)] + 2(g_{m_{opt}} - g)^2(x_0)$$

and with Assumption (**H**),

$$\begin{aligned}&\mathbb{E}_1[(\widehat{g}_{m_{opt}} - g_{m_{opt}})^2(x_0)]\\ &\leq \int_0^{+\infty} P_1[(\widehat{g}_{m_{opt}} - g_{m_{opt}})^2(x_0) \geq x]dx\\ &= \int_{-2Bx_{m_{opt}}D_{m_{opt}}/n}^{+\infty} P_1\left[(\widehat{g}_{m_{opt}} - g_{m_{opt}})^2(x_0) \geq y + 2B\frac{x_{m_{opt}}D_{m_{opt}}}{n}\right]dy\\ &= 2B\frac{x_{m_{opt}}D_{m_{opt}}}{n} + \int_0^{+\infty} P_1\left[|(\widehat{g}_{m_{opt}} - g_{m_{opt}})(x_0)| \geq \sqrt{2B}\sqrt{\frac{y}{2B} + \frac{x_{m_{opt}}D_{m_{opt}}}{n}}\right]dy.\end{aligned}$$

$\sqrt{2B} \geq 1 > B'$, so

$$
\begin{aligned}
&\mathbb{E}_1[(\widehat{g}_{m_{opt}} - g_{m_{opt}})^2(x_0)] \\
&\leq 2B\frac{x_{m_{opt}}D_{m_{opt}}}{n} + \int_0^{+\infty} P_1[|(\widehat{g}_{m_{opt}} - g_{m_{opt}})(x_0)| \geq B'\sqrt{\frac{y}{2B} + \frac{x_{m_{opt}}D_{m_{opt}}}{n}}]dy \\
&\leq 2B\frac{x_{m_{opt}}D_{m_{opt}}}{n} + \int_0^{+\infty} F\left(\frac{y}{2B}, m_{opt}, n, B'\right) dy \\
&\leq 2B\frac{x_{m_{opt}}D_{m_{opt}}}{n} + \frac{2BB_0}{n}.
\end{aligned}
$$

Hence

$$
\begin{aligned}
\mathbb{E}_1[\mathcal{U}_{opt}] &\leq 4(g_{m_{opt}} - g)^2(x_0) + 4B\frac{x_{m_{opt}}D_{m_{opt}}}{n} + \frac{4BB_0}{n} + 2(1+\delta)Crit(m_{opt}) \\
&\quad +4\frac{x_{m_{opt}}D_{m_{opt}}}{n} + (1+\delta)\sup_{j\in J_n, D_j \geq D_{m_{opt}}} (g_j - g_{m_{opt}})^2(x_0) \\
&\leq A_2[Crit(m_{opt}) + (g_{m_{opt}} - g)^2(x_0)] + \frac{B_1}{n}
\end{aligned}
$$

where A_2 depends on B and B' (as δ depends on B'), and B_1 depends on B and B_0. Then (5.13) concludes the proof of Proposition 5.2.1. □

5.4.1 Proof of Lemma 5.4.1

The deviation between $\widehat{Crit}(m)$ and $Crit(m)$ can be expressed with deviation between $\widehat{g}_m$ and g_m. Therefore, it is upper bounded via the application F as stated by the following claim.

Claim 5.1 *For every $\delta > 0$, for every $x > 0$ and for every model m:*

$$
P_1\left[\widehat{Crit}(m) \geq (1+\delta)Crit(m) + x\right] \leq F\left(\frac{x}{2}, m, n, C_\delta\right) + \sum_{j\in J_n, D_j \geq D_m} F\left(\frac{x}{2}, j, n, C_\delta\right)
$$

where

$$
C_\delta = \frac{1}{\sqrt{2(1+1/\delta)}}.
$$

Indeed,

$$
\begin{aligned}
&P_1\left[\widehat{Crit}(m) \geq (1+\delta)Crit(m) + x\right] \\
\leq\ &P_1\left[\sup_{j\in J_n, D_j \geq D_m}\left((\widehat{g}_j - \widehat{g}_m)^2(x_0) - \frac{x_j D_j + x_m D_m}{n}\right)_+ \geq (1+\delta)\sup_{j\in J_n, D_j \geq D_m}(g_j - g_m)^2(x_0) + x\right]
\end{aligned}
$$

As $\sup_{j\in J_n, D_j\geq D_m}(g_j-g_m)^2(x_0)+x$ is positive, we omit the positive part $(.)_+$. Besides, for every sequences $(a_j)_{j\in J}$ and $(b_j)_{j\in J}$,

$$P_1[\cap_{j\in J}\{a_j<b_j\}]\leq P_1\left[\sup_{j\in J}a_j<\sup_{j\in J}b_j\right].$$

Consider the complementary of the above expression:

$$P_1\left[\sup_{j\in J}a_j\geq\sup_{j\in J}b_j\right]\leq P_1[\cup_{j\in J}\{a_j\geq b_j\}].$$

Hence

$$\begin{aligned}
&P_1[\widehat{Crit}(m)\geq(1+\delta)Crit(m)+x]\\
\leq\ & P_1\left[\sup_{j\in J_n,D_j\geq D_m}\left((\widehat{g}_j-\widehat{g}_m)^2(x_0)-\frac{x_jD_j+x_mD_m}{n}\right)\geq(1+\delta)\sup_{j\in J_n,D_j\geq D_m}(g_j-g_m)^2(x_0)+x\right]\\
\leq\ & P_1\left[\cup_{j\in J_n,D_j\geq D_m}\left\{\left((\widehat{g}_j-\widehat{g}_m)^2(x_0)-\frac{x_jD_j+x_mD_m}{n}\right)\geq(1+\delta)(g_j-g_m)^2(x_0)+x\right\}\right] \quad (5.15)
\end{aligned}$$

Moreover, for every j,m,

$$\begin{aligned}
&\left\{\left((\widehat{g}_j-\widehat{g}_m)^2(x_0)-\frac{x_jD_j+x_mD_m}{n}\right)\geq(1+\delta)(g_j-g_m)^2(x_0)+x\right\}\\
&=\left\{(\widehat{g}_j-\widehat{g}_m)^2(x_0)\geq(1+\delta)(g_j-g_m)^2(x_0)+(1+\frac{1}{\delta})\left(\sqrt{\frac{x+(x_jD_j+x_mD_m)/n}{1+1/\delta}}\right)^2\right\}.
\end{aligned}$$

For every positive numbers a,b, $(a+b)^2\leq a^2(1+1/\delta)+b^2(1+\delta)$, therefore

$$\begin{aligned}
&\left\{\left((\widehat{g}_j-\widehat{g}_m)^2(x_0)-\frac{x_jD_j+x_mD_m}{n}\right)\geq(1+\delta)(g_j-g_m)^2(x_0)+x\right\}\\
&\subset\left\{(\widehat{g}_j-\widehat{g}_m)^2(x_0)\geq\left(|(g_j-g_m)(x_0)|+\sqrt{\frac{x+(x_jD_j+x_mD_m)/n}{1+1/\delta}}\right)^2\right\}\\
&=\left\{|(\widehat{g}_j-\widehat{g}_m)(x_0)|\geq|(g_j-g_m)(x_0)|+\sqrt{\frac{x+(x_jD_j+x_mD_m)/n}{1+1/\delta}}\right\}.
\end{aligned}$$

Besides,

$$|(\widehat{g}_j-\widehat{g}_m)(x_0)|-|(g_j-g_m)(x_0)|\leq|(\widehat{g}_j-\widehat{g}_m)(x_0)-(g_j-g_m)(x_0)|.$$

Thus

$$\left\{\left((\widehat{g}_j-\widehat{g}_m)^2(x_0)-\frac{x_jD_j+x_mD_m}{n}\right)\geq(1+\delta)(g_j-g_m)^2(x_0)+x\right\}$$
$$\subset\left\{|(\widehat{g}_j-g_j)(x_0)-(\widehat{g}_m-g_m)(x_0)|\geq\sqrt{\frac{x+(x_jD_j+x_mD_m)/n}{1+1/\delta}}\right\}.$$

For every positive numbers a,b, $\sqrt{a+b}\geq(1/\sqrt{2})(\sqrt{a}+\sqrt{b})$, hence

$$\left\{\left((\widehat{g}_j-\widehat{g}_m)^2(x_0)-\frac{x_jD_j+x_mD_m}{n}\right)\geq(1+\delta)\sup_{j\in J_n,D_j\geq D_m}(g_j-g_m)^2(x_0)+x\right\}$$
$$\subset\left\{|(\widehat{g}_j-g_j)(x_0)-(\widehat{g}_m-g_m)(x_0)|\geq\sqrt{\frac{1}{2(1+1/\delta)}}\left(\sqrt{\frac{x}{2}+\frac{x_jD_j}{n}}+\sqrt{\frac{x}{2}+\frac{x_mD_m}{n}}\right)\right\}$$
$$\subset\left\{|(\widehat{g}_j-g_j)(x_0)|\geq\sqrt{\frac{1}{2(1+1/\delta)}}\sqrt{\frac{x}{2}+\frac{x_jD_j}{n}}\right\}$$
$$\cup\left\{|(\widehat{g}_m-g_m)(x_0)|\geq\sqrt{\frac{1}{2(1+1/\delta)}}\sqrt{\frac{x}{2}+\frac{x_mD_m}{n}}\right\}$$
$$=\mathcal{O}_j\cup\mathcal{O}_m. \tag{5.16}$$

Reporting this result in (5.15), we get

$$\begin{aligned}
&P_1[\widehat{Crit}(m)\geq(1+\delta)Crit(m)+x]\\
&\leq P_1[\mathcal{O}_m\cup\left(\cup_{j\in J_n,D_j\geq D_m}\mathcal{O}_j\right)]\\
&\leq P_1[\mathcal{O}_m]+\sum_{j\in J_n,D_j\geq D_m}P_1[\mathcal{O}_j]\\
&\leq F\left(\frac{x}{2},m,n,C_\delta\right)+\sum_{j\in J_n,D_j\geq D_m}F\left(\frac{x}{2},j,n,C_\delta\right)
\end{aligned}$$

which ends the proof of Claim 5.1. □

- Let us prove (1) in Lemma 5.4.1.

$$P_1\left[\left\{(\widehat{g}_{\widehat{m}}-g)^2(x_0)\geq(1+\delta)\left(\sup_{j\in J_n,D_j\geq D_{m_{opt}}}(g_j-g)^2(x_0)+Crit(m_{opt})\right)+x\right\}\cap\{D_{\widehat{m}}>D_{m_{opt}}\}\right]$$
$$\leq P_1\left[\left\{(\widehat{g}_{\widehat{m}}-g)^2(x_0)\geq(1+\delta)\sup_{j\in J_n,D_j\geq D_{m_{opt}}}(g_j-g)^2(x_0)+\widehat{Crit}(\widehat{m})+\frac{x}{2}\right\}\cap\{D_{\widehat{m}}>D_{m_{opt}}\}\right]$$
$$+P_1\left[\widehat{Crit}(\widehat{m})\geq(1+\delta)Crit(m_{opt})+\frac{x}{2}\right].$$

By definition of $\widehat{m}$, $\widehat{Crit}(\widehat{m}) = \inf_{m\in J_n} \widehat{Crit}(m) \leq \widehat{Crit}(m_{opt})$ thus with Claim 5.1,

$$\begin{aligned} P_1\left[\widehat{Crit}(\widehat{m}) \geq (1+\delta)Crit(m_{opt}) + \frac{x}{2}\right] &\leq P_1\left[\widehat{Crit}(m_{opt}) \geq (1+\delta)Crit(m_{opt}) + \frac{x}{2}\right] \\ &\leq F\left(\frac{x}{4}, m_{opt}, n, C_\delta\right) + \sum_{j\in J_n, D_j \geq D_{m_{opt}}} F\left(\frac{x}{4}, j, n, C_\delta\right). \end{aligned}$$

Besides for every model m, $Crit(m) \geq Bx_m D_m/n$ by definition of $Crit(m)$, and if $D_{\widehat{m}} > D_{m_{opt}}$, $\sup_{j\in J_n, D_j \geq D_{m_{opt}}} (g_j - g)^2(x_0) \geq (g_{\widehat{m}} - g)^2(x_0)$, thus

$$\begin{aligned} & P_1\left[\left\{(\widehat{g}_{\widehat{m}} - g)^2(x_0) \geq (1+\delta) \sup_{j\in J_n, D_j \geq D_{m_{opt}}} (g_j - g)^2(x_0) + \widehat{Crit}(\widehat{m}) + \frac{x}{2}\right\} \cap \{D_{\widehat{m}} > D_{m_{opt}}\}\right] \\ \leq\ & P_1\left[(\widehat{g}_{\widehat{m}} - g)^2(x_0) \geq (1+\delta)(g_{\widehat{m}} - g)^2(x_0) + Bx_{\widehat{m}}\frac{D_{\widehat{m}}}{n} + \frac{x}{2}\right] \\ \leq\ & \sum_{m\in J_n} P_1\left[(\widehat{g}_m - g)^2(x_0) \geq (1+\delta)(g_m - g)^2(x_0) + Bx_m\frac{D_m}{n} + \frac{x}{2}\right]. \end{aligned}$$

Moreover,

$$\begin{aligned} & P_1\left[(\widehat{g}_m - g)^2(x_0) \geq (1+\delta)(g_m - g)^2(x_0) + Bx_m\frac{D_m}{n} + \frac{x}{2}\right] \\ & \leq P_1\left[\left(1 + \frac{1}{\delta}\right)(\widehat{g}_m - g_m)^2(x_0) + (1+\delta)(g_m - g)^2(x_0) \geq (1+\delta)(g_m - g)^2(x_0) + Bx_m\frac{D_m}{n} + \frac{x}{2}\right] \\ & = P_1\left[|(\widehat{g}_m - g_m)(x_0)| \geq C_\delta\sqrt{2B}\sqrt{\frac{x}{2B} + x_m\frac{D_m}{n}}\right] \\ & \leq P_1\left[|(\widehat{g}_m - g_m)(x_0)| \geq C_\delta\sqrt{\frac{x}{2B} + x_m\frac{D_m}{n}}\right] \\ & \leq F\left(\frac{x}{2B}, m, n, C_\delta\right). \end{aligned}$$

• Let us prove (2) in Lemma 5.4.1. Assume that $D_{\widehat{m}} \leq D_{m_{opt}}$, then $x_{\widehat{m}} \leq x_{m_{opt}}$ by Assumption (5.3). Hence

$$\begin{aligned} \widehat{Crit}(\widehat{m}) &\geq \sup_{j\in J_n, D_j \geq D_{\widehat{m}}} \left[(\widehat{g}_j - \widehat{g}_{\widehat{m}})^2(x_0) - \frac{x_{\widehat{m}}D_{\widehat{m}} + x_j D_j}{n}\right] \\ &\geq (\widehat{g}_{m_{opt}} - \widehat{g}_{\widehat{m}})^2(x_0) - \frac{x_{\widehat{m}}D_{\widehat{m}} + x_{m_{opt}}D_{m_{opt}}}{n}. \end{aligned}$$

Moreover $(\widehat{g}_{\widehat{m}} - g)^2(x_0) \leq 2[(\widehat{g}_{\widehat{m}} - \widehat{g}_{m_{opt}})^2(x_0) + (\widehat{g}_{m_{opt}} - g)^2(x_0)]$, thus

$$\begin{aligned}\widehat{Crit}(\widehat{m}) &\geq \frac{1}{2}(\widehat{g}_{\widehat{m}} - g)^2(x_0) - (\widehat{g}_{m_{opt}} - g)^2(x_0) - \frac{x_{\widehat{m}} D_{\widehat{m}} + x_{m_{opt}} D_{m_{opt}}}{n} \\ &\geq \frac{1}{2}(\widehat{g}_{\widehat{m}} - g)^2(x_0) - (\widehat{g}_{m_{opt}} - g)^2(x_0) - 2\frac{x_{m_{opt}} D_{m_{opt}}}{n}.\end{aligned}$$

Therefore

$$\begin{aligned}P_1[\{\frac{1}{2}(\widehat{g}_{\widehat{m}} - g)^2(x_0) &\geq (1+\delta)Crit(m_{opt}) + 2\frac{x_{m_{opt}} D_{m_{opt}}}{n} \\ &\quad +(\widehat{g}_{m_{opt}} - g)^2(x_0) + x\} \cap \{\widehat{m} \leq m_{opt}\}] \\ &\leq P_1\left[\left\{\widehat{Crit}(\widehat{m}) \geq (1+\delta)Crit(m_{opt}) + x\right\} \cap \{\widehat{m} \leq m_{opt}\}\right] \\ &\leq P_1\left[\widehat{Crit}(m_{opt}) \geq (1+\delta)Crit(m_{opt}) + x\right] \\ &\leq F\left(\frac{x}{2}, m_{opt}, n, C_\delta\right) + \sum_{j \in J_n, D_j \geq D_{m_{opt}}} F\left(\frac{x}{2}, j, n, C_\delta\right). \quad \square\end{aligned}$$

5.5 Proofs of Section 5.3

5.5.1 Proof of Theorem 5.3.1

Let

$$\Delta = \cap_{m=1}^{N_n} \Delta_m$$

where Δ_m is defined in (5.9). We denote by $P_1[.] = P[. \cap \Delta]$ and $\mathbb{E}_1[.] = \mathbb{E}_1[.\mathbb{1}_\Delta]$.

Claim 5.2 *There exists a function $F(x, m, n, a)$ such that for every $m \in \{1, \ldots, N_n\}$*

$$P_1\left[|(\widehat{h}_m - h_m)(x_0)| \geq a\sqrt{x + x_m \frac{D_m}{n}}\right] \leq F(x, m, n, a)$$

and

$$\sum_{m=1}^{N_n} \int_0^{+\infty} F\left(x, m, n, \frac{1}{\sqrt{A'}}\right) dx \leq \frac{B_0}{n}.$$

for some constant B_0.

Let us prove Claim 5.2. According to Chapter 4, $\widehat{h}_m$ minimises the empirical contrast

$$\gamma_n(t) = \|t\|_n^2 - \frac{2}{n}\sum_{i=1}^{n} \delta_i t(T_i)$$

for $t \in S_m$, and h_m minimises

$$\gamma(t) = \|t\|_{\overline{F}_T} - 2\langle t, h\rangle_{\overline{F}_T}.$$

For every $m \in \{1, \ldots, N_n\}$, we define the application $\overline{h}_m$

$$\overline{h}_m = \begin{cases} \arg\min_{t \in S_m} \overline{\gamma}_n(t) & \text{on } \Delta_m \\ 0 \quad \text{otherwise} \end{cases}$$

where

$$\overline{\gamma}_n(t) = \|t\|_n^2 - 2\langle t, h\rangle_{\overline{F}_T}.$$

Lemma 5.5.1 *For every $m \in \{1, \ldots, N_n\}$,*

$$(\widehat{h}_m - \overline{h}_m)^2(x_0) \leq \mathcal{A}_1^2 D_m \left(\sum_{j=0}^{r} \left(\frac{1}{n}\sum_{i=1}^{n} \left(\delta_i \phi_{k_0,m}^j(T_i) - \mathbb{E}[\delta_i \phi_{k_0,m}^j(T_i)] \right) \right)^2 \right).$$

Lemma 5.5.2 *For every $m \in \{1, \ldots, N_n\}$,*

$$(\overline{h}_m - h_m)^2(x_0) \leq \mathcal{A}_2^2 \left(\sum_{j,l=0}^{r} \left(\langle \phi_{k_0,m}^l, \phi_{k_0,m}^j \rangle_n - \langle \phi_{k_0,m}^l, \phi_{k_0,m}^j \rangle_{\overline{F}_T} \right)^2 \right)$$

where

$$\mathcal{A}_2^2 = (r+1)\|h\|_\infty^2 \left(\sum_{j=0}^{r} \|\phi^j\|_\infty^2 \right).$$

Lemmas 5.5.1 and 5.5.2 lead to Claim 5.2. Indeed, let $a > 0$.

$$\begin{aligned} & P_1 \left[|(\widehat{h}_m - h_m)(x_0)| \geq a\sqrt{x + x_m \frac{D_m}{n}} \right] \\ & \leq P_1 \left[(\widehat{h}_m - \overline{h}_m)^2(x_0) \geq \frac{a^2}{2}\left(x + x_m \frac{D_m}{n} \right) \right] + P_1 \left[(\overline{h}_m - h_m)^2(x_0) \geq \frac{a^2}{2}\left(x + x_m \frac{D_m}{n} \right) \right] \\ & = P_{1,m} + P_{2,m}. \end{aligned}$$

On the one hand, according to Lemma 5.5.1,

$$
\begin{aligned}
P_{1,m} &\leq P\left[\sum_{j=0}^{r}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_i\phi_{k_0,m}^j(T_i)-\mathbb{E}[\delta_i\phi_{k_0,m}^j(T_i)]\right)^2 \geq \frac{a^2}{2\mathcal{A}_1^2}\left(\frac{x}{D_m}+\frac{x_m}{n}\right)\right] \\
&\leq \sum_{j=0}^{r} P\left[\left|\frac{1}{n}\sum_{i=1}^{n}\delta_i\phi_{k_0,m}^j(T_i)-\mathbb{E}[\delta_i\phi_{k_0,m}^j(T_i)]\right| \geq \frac{a}{\sqrt{2(r+1)}\mathcal{A}_1}\sqrt{\frac{x}{D_m}+\frac{x_m}{n}}\right].
\end{aligned}
$$

We apply Bernstein Inequality (Theorem 1.2.4) with the following parameters v and c.

$$
\mathbb{E}\left[\delta_i^2(\phi_{k_0,m}^j(T_1))^2\right] = \int_A (\phi_{k_0,m}^j(x))^2 h(x)\overline{F}_T(x)dx \leq \|h\|_{\infty,A}\|\phi_{k_0,m}\|_{\overline{F}_T}^2 \leq \|h\|_\infty = v.
$$

$$
\|\delta_1\phi_{k_0,m}^j(T_1)\|_\infty \leq \sqrt{D_m} = c.
$$

Thus

$$
P_{1,m} \leq 2(r+1)\exp\left(-\min\left(\frac{n\epsilon^2}{2v},\frac{n\epsilon}{c}\right)\right)
$$

where

$$
\epsilon = \frac{a}{\sqrt{2(r+1)}\mathcal{A}_1}\sqrt{\frac{x}{D_m}+\frac{x_m}{n}}.
$$

$$
\frac{n\epsilon^2}{v} = \frac{a^2}{2(r+1)\mathcal{A}_1^2}\left(\frac{nx}{D_m}+x_m\right) = \mathcal{C}_1(x,m,n,a)
$$

and

$$
\frac{n\epsilon}{c} \geq \frac{a}{2\sqrt{r+1}\mathcal{A}_1}\left(\sqrt{x}\frac{n}{D_m}+\sqrt{x_m\frac{n}{D_m}}\right) = \mathcal{C}_2(x,m,n,a).
$$

Hence

$$
P_{1,m} \leq 2(r+1)\exp\left(-\min(\mathcal{C}_1(x,m,n,a),\mathcal{C}_2(x,m,n,a))\right).
$$

On the other hand, according to Lemma 5.5.2,

$$
\begin{aligned}
P_{2,m} &\leq \sum_{j,l=0}^{r} P\left[\mathcal{A}_2^2(\langle\phi_{k_0,m}^l,\phi_{k_0,m}^j\rangle_n - \langle\phi_{k_0,m}^l,\phi_{k_0,m}^j\rangle_{\overline{F}_T})^2 \geq \frac{a^2}{2(r+1)^2}\left(x+x_m\frac{D_m}{n}\right)\right] \\
&= \sum_{j,l=0}^{r} P\left[\left|\frac{1}{n}\sum_{i=1}^{n}\left(\int_A \phi_{k_0,m}^l(x)\phi_{k_0,m}^j(x)\mathbb{1}_{T_i\geq x}dx - \mathbb{E}\left[\int_A \phi_{k_0,m}^l(x)\phi_{k_0,m}^j(x)\mathbb{1}_{T_i\geq x}dx\right]\right)\right| \right. \\
&\qquad \left. \geq \frac{a}{\sqrt{2}(r+1)\mathcal{A}_2}\sqrt{x+x_m\frac{D_m}{n}}\right]
\end{aligned}
$$

We upper bound each term of this sum with Bernstein Inequality. Let $j, l \in \{0, \dots, r\}$, let us compute the parameters v and c. According to Cauchy Schwartz Inequality, we have

$$\begin{aligned}
\mathbb{E}\left[\left(\int_A \phi^l_{k_0,m}(x)\phi^j_{k_0,m}(x)\mathbb{1}_{T_i \geq x}dx\right)^2\right] &\leq \mathbb{E}\left[\left(\int_A \left(\phi^j_{k_0,m}(x)\right)^2 dx\right) \times \left(\int_A (\phi^l_{k_0,m}(x))^2 \mathbb{1}_{T_i \geq x}dx\right)\right] \\
&= \|\phi^l_{k_0,m}\|^2_{F_T} \\
&\leq 1 = v.
\end{aligned}$$

Moreover

$$\begin{aligned}
\left\|\int_A \phi^j_{k_0,m}(x)\phi^l_{k_0,m}(x)\mathbb{1}_{T_i \geq x}dx\right\|_\infty &\leq \int_A \left|\phi^j_{k_0,m}(x)\phi^l_{k_0,m}(x)dx\right| \\
&\leq \sqrt{\left(\int_A (\phi^j_{k_0,m}(x))^2 dx\right)\left(\int_A (\phi^l_{k_0,m}(x))^2 dx\right)} \\
&= 1 = c.
\end{aligned}$$

Let

$$\epsilon = \frac{a}{\sqrt{2}(r+1)\mathcal{A}_2}\sqrt{x + x_m\frac{D_m}{n}}.$$

$$\frac{n\epsilon^2}{v} = \frac{a^2}{2(r+1)^2\mathcal{A}_2^2}\left(nx + x_m D_m\right) = \mathcal{C}_3(x, m, n, a),$$

$$\frac{n\epsilon}{c} \geq \frac{a}{2(r+1)\mathcal{A}_2}\left(n\sqrt{x} + \sqrt{x_m D_m n}\right) = \mathcal{C}_4(x, m, n, a).$$

Then

$$P_{2,m} \leq 2(r+1)^2 \exp\left(-\min(\mathcal{C}_3(x, m, n, a), \mathcal{C}_4(x, m, n, a))\right).$$

Finally,

$$P_1\left[|(\widehat{h}_m - h_m)(x_0)| \geq a\sqrt{x + x_m\frac{D_m}{n}}\right] \leq F(x, m, n, a)$$

where

$$\begin{aligned}
F(x, m, n, a) &= 2(r+1)\exp\left(-\min(\mathcal{C}_1(x, m, n, a), \mathcal{C}_2(x, m, n, a))\right) \\
&+2(r+1)^2\exp\left(-\min(\mathcal{C}_3(x, m, n, a), \mathcal{C}_4(x, m, n, a))\right).
\end{aligned}$$

For every constant $C > 0$, $\int_0^{+\infty} \exp(-Cx)dx = 1/C$ and $\int_0^{+\infty} \exp(-C\sqrt{x})dx = 2/C^2$. Therefore,

$$\sum_{m=1}^{N_n} \int_0^{+\infty} \exp(-\mathcal{C}_1(x, m, n, a,))dx \leq \frac{2(r+1)\mathcal{A}_1^2}{a^2} \frac{1}{n} \sum_{m=1}^{N_n} D_m \exp\left(-\frac{a^2}{2(r+1)\mathcal{A}_1^2} x_m\right).$$

Assume that

$$D_m \exp\left(-\frac{a^2}{2(r+1)\mathcal{A}_1^2} x_m\right) \leq D_m^{-2} \quad \Leftrightarrow \quad x_m \geq \frac{6(r+1)\mathcal{A}_1^2}{a^2} \log D_m, \tag{5.17}$$

then

$$\sum_{m=1}^{N_n} \int_0^{+\infty} \exp(-\mathcal{C}_1(x, m, n, a,))dx \leq \frac{C_1}{n}.$$

Similarly,

$$\begin{aligned}
&\sum_{m=1}^{N_n} \int_0^{+\infty} \exp(-\mathcal{C}_2(x, m, n, a))dx \\
&= 2\left(\frac{2\sqrt{r+1}\mathcal{A}_1}{a}\right)^2 \sum_{m=1}^{N_n} \frac{D_m^2}{n^2} \exp\left(-\frac{a}{2\sqrt{r+1}\mathcal{A}_1} \sqrt{x_m \frac{n}{D_m}}\right) \\
&\leq \left(\frac{8(r+1)\mathcal{A}_1^2}{a^2}\right) \frac{1}{n} \sum_{m=1}^{N_n} D_m \exp\left(-\frac{a}{2\sqrt{r+1}\mathcal{A}_1} \sqrt{x_m \frac{n}{D_m}}\right).
\end{aligned}$$

Assume that

$$D_m \exp\left(-\frac{a}{2\sqrt{r+1}\mathcal{A}_1} \sqrt{x_m \frac{n}{D_m}}\right) \leq D_m^{-2} \quad \Leftrightarrow \quad x_m \geq \frac{36(r+1)\mathcal{A}_1^2}{a^2} \log^2 D_m \frac{D_m}{n}, \tag{5.18}$$

then

$$\sum_{m=1}^{N_n} \int_0^{+\infty} \exp(-\mathcal{C}_2(x, m, n, a))dx \leq \frac{C_2}{n}.$$

Moreover,

$$\sum_{m=1}^{N_n} \int_0^{+\infty} \exp(-\mathcal{C}_3(x, m, n, a))dx = \frac{2(r+1)^2\mathcal{A}_2^2}{a^2} \frac{1}{n} \sum_{m=1}^{N_n} \exp\left(-\frac{a^2}{2(r+1)^2\mathcal{A}_2^2} x_m D_m\right)$$

According to (5.8), there exists a constant C_3 such that

$$\sum_{m=1}^{N_n} \int_0^{+\infty} \exp(-\mathcal{C}_3(x,m,n,a))dx \leq \frac{C_3}{n}$$

$$\begin{aligned}\sum_{m=1}^{N_n} \int_0^{+\infty} \exp(-\mathcal{C}_4(x,m,n,a))dx &= 2\left(\frac{2(r+1)\mathcal{A}_2}{a}\right)^2 \frac{1}{n^2} \sum_{m=1}^{N_n} \exp\left(-\frac{a}{2(r+1)\mathcal{A}_2}\sqrt{x_m n}\right) \\ &\leq \frac{C_4}{n}.\end{aligned}$$

For every $a \geq 1/\sqrt{B'}$, conditions (5.17) and (5.18) are satisfied by definition of the weights (x_m) (see (5.11)). Thus there exists a constant B_0 which depends on $(\|h\|_\infty, r, a, \sum_{j=0}^r \|\phi^j\|^2)$ such that

$$\sum_{m=1}^{N_n} \int_0^{+\infty} F(x,m,n,\frac{1}{\sqrt{A'}})dx \leq \frac{B_0}{n}.$$

which concludes the proof of Claim 5.2. □

It follows immediately from Proposition 5.2.1 that

$$\mathbb{E}\left[(\widehat{h}_{\widehat{m}} - h)^2(x_0)\mathbb{1}_\Delta\right] \leq C[Crit(m_{opt}) + (g - g_{m_{opt}})^2(x_0)] + \frac{C'}{n} \tag{5.19}$$

where C is a numerical constant depending on B and B', and C' depends on $\|h\|_\infty$, $\overline{F}_0$, B' and on the basis $\{\phi^j\}$.

Claim 5.3

$$\mathbb{E}[(\widehat{h}_{\widehat{m}} - h)^2(x_0)\mathbb{1}_{\Delta^c}] \leq \frac{C}{n}.$$

Indeed, according to Proposition 4.6.3 in Chapter 4,

$$P[\Delta^c] \leq 2N_n^2 \exp\left(-C_1 \overline{F}_0^2 \frac{n}{N_n}\right)$$

for some numerical constant C_1. Moreover, for every $m \in \{1, \ldots, N_n\}$.

$$\begin{aligned}\left(\widehat{h}_m(x_0)\right)^2 &= \left(\sum_{j=0}^{r}\widehat{a}^j_{k_0,m}\phi^j_{k_0,m}(x_0)\right)^2\\ &\leq \left(\sum_{j=0}^{r}(\widehat{a}^j_{k_0,m})^2\right)\left(\sum_{j=0}^{r}\|\phi^j_{k_0,m}\|^2_\infty\right)\\ &= \|(\widehat{G}^{(k_0)}_m)^{-1}\widehat{V}^{(k_0)}_m\|^2 D_m\left(\sum_{j=0}^{r}\|\phi^j\|^2_\infty\right)\\ &\leq [\min(Sp(\widehat{G}^{(k_0)}_m))]^{-2}\|\widehat{V}^{(k_0)}_m\|^2 D_m\left(\sum_{j=0}^{r}\|\phi^j\|^2_\infty\right).\end{aligned}$$

Besides

$$\begin{aligned}\|\widehat{V}^{(k_0)}_m\|^2 &= \sum_{j=0}^{r}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_i\phi^j_{k_0,m}(T_i)\right)^2\\ &\leq \sum_{j=0}^{r}\frac{1}{n}\sum_{i=1}^{n}\delta_i^2\left(\phi^j_{k_0,m}(T_i)\right)^2\\ &\leq \sum_{j=0}^{r}\|\phi^j_{k_0,m}\|^2_\infty\\ &\leq D_m\left(\sum_{j=0}^{r}\|\phi^j\|^2_\infty\right).\end{aligned}$$

Thus

$$\left(\widehat{h}_m(x_0)\right)^2 \leq \left(\frac{4}{3\overline{F}_0}\right)^2\left(\sum_{j=0}^{r}\|\phi^j\|^2_\infty\right)^2\|h\|^2_{\infty,A}D_m^2.$$

Therefore

$$(\widehat{h}_{\widehat{m}}-h)^2(x_0) \leq 2\|h\|^2_{\infty,A}\left\{\left(\frac{4}{3\overline{F}_0}\right)^2\left(\sum_{j=0}^{r}\|\phi^j\|^2_\infty\right)^2 N_n^2+1\right\}$$

and

$$
\begin{aligned}
\mathbb{E}\left[(\widehat{h}_{\widehat{m}}-h)^2(x_0)\mathbb{1}_{\Delta^c}\right] &\leq 4\|h\|_{\infty,A}^2\left\{\left(\frac{4}{3\overline{F}_0}\right)^2\left(\sum_{j=0}^{r}\|\phi^j\|_\infty^2\right)^2 N_n^2+1\right\}N_n^2\exp\left(-C_1\overline{F}_0^2\frac{n}{N_n}\right)\\
&\leq \left[4\|h\|_{\infty,A}\left(\frac{4}{3\overline{F}_0}\right)\left(\sum_{j=0}^{r}\|\phi^j\|_\infty^2\right)\right]^2 n^4\exp\left(-C_1\overline{F}_0^2\log^2 n\right)\\
&= \left[4\|h\|_{\infty,A}\left(\frac{4}{3\overline{F}_0}\right)\left(\sum_{j=0}^{r}\|\phi^j\|_\infty^2\right)\right]^2 n^4\left(\exp\left(-C_1\overline{F}_0^2\log n\right)\right)^{\log n}\\
&= \left[4\|h\|_{\infty,A}\left(\frac{4}{3\overline{F}_0}\right)\left(\sum_{j=0}^{r}\|\phi^j\|_\infty^2\right)\right]^2 n^{5-C_1\overline{F}_0^2\log n}\frac{1}{n}\\
&\leq \frac{C}{n} \qquad (5.20)
\end{aligned}
$$

which concludes the proof of Claim 5.3. □

Finally (5.19) and Claim 5.3 provide the result of Theorem 5.3.1. Now let us prove Lemmas 5.5.1 and 5.5.2, which rely on the block diagonal structure of the Gram matrix $\widehat{G}_m$.

Proof of Lemma 5.5.1

Let $m \in \{1,\dots,N_n\}$ and k_0 be such that $x_0 \in [(k_0-1)/D_m, k_0/D_m[$. The result is obvious on Δ_m^c as $\widehat{h}_m = \overline{h}_m = 0$, thus we assume that we are on Δ_m, defined by (5.9).

Let $\overline{A}_m = (\overline{a}_\lambda)_{\lambda\in I_m}$ be the coefficients of $\overline{h}_m$ in the basis $(\phi_\lambda)_{\lambda\in I_m}$.

$$
\overline{h}_m = \sum_{k=1}^{D_m}\sum_{j=0}^{r}\overline{a}_{k,m}^j\phi_{k,m}^j.
$$

$\overline{A}_m$ satisfies

$$
\frac{\partial}{\partial a_{\lambda_0}}\overline{\gamma}_n\Big(\sum_{\lambda\in I_m}\overline{a}_\lambda\phi_\lambda\Big)=0, \quad \forall\lambda_0\in I_m
$$

which is equivalent to

$$
\widehat{G}_m\overline{A}_m = V_m. \qquad (5.21)
$$

According to the expression of $\widehat{G}_m$, $\widehat{V}_m$ and V_m in Section 5.3.2

$$(\widehat{h}_m - \overline{h}_m)^2(x_0) = \left(\sum_{j=0}^{r} (\widehat{a}^j_{k_0,m} - \overline{a}^j_{k_0,m}) \phi^j_{k_0,m}(x_0) \right)^2 .$$

As m and k_0 are fixed, we simplify notations by setting

$$a_j = a^j_{k_0,m} \quad \text{and} \quad \overline{a}_j = \overline{a}^j_{k_0,m}.$$

$$\begin{aligned}
(\widehat{h}_m - \overline{h}_m)^2(x_0) &= \left(\sum_{j=0}^{r} (\widehat{a}_j - \overline{a}_j) \phi^j_{k_0,m}(x_0) \right)^2 \\
&\leq \left(\sum_{j=0}^{r} (\widehat{a}_j - \overline{a}_j)^2 \right) \left(\sum_{j=0}^{r} (\phi^j_{k_0,m}(x_0))^2 \right) \\
&= \|(\widehat{G}^{(k_0)}_m)^{-1} (\widehat{V}^{(k_0)}_m - V^{(k_0)}_m)\|^2 \left(\sum_{j=0}^{r} (\phi^j_{k_0,m}(x_0))^2 \right). \qquad (5.22)
\end{aligned}$$

On the one hand, for every $j = 0, \ldots, r$,

$$\|\phi^j_{k_0,m}\|_\infty \leq \sqrt{D_m} \|\phi^j\|_\infty \quad \Rightarrow \quad \sum_{j=0}^{r} (\phi^j_{k_0,m}(x_0))^2 \leq D_m \left(\sum_{j=0}^{r} \|\phi^j\|^2_\infty \right). \qquad (5.23)$$

On the other hand, the matrix $\widehat{G}^{(k_0)}_m$ is symmetric positive (as a matrix of a scalar product) so its eigenvalues are non negative. Thus

$$\|(\widehat{G}^{(k_0)}_m)^{-1} (\widehat{V}^{(k_0)}_m - V^{(k_0)}_m)\| \leq \rho \left(\widehat{G}^{(k_0)}_m)^{-1} \right) \|\widehat{V}^{(k_0)}_m - V^{(k_0)}_m\|$$

where $\rho((\widehat{G}^{(k_0)}_m)^{-1})$ denotes the spectral radius of the matrix $(\widehat{G}^{(k_0)}_m)^{-1}$. Moreover on Δ_m,

$$\rho((\widehat{G}^{(k_0)}_m)^{-1}) = \left[\min(Sp(\widehat{G}^{(k_0)}_m)) \right]^{-1} \leq \left[\min(Sp(\widehat{G}_m)) \right]^{-1} \leq \frac{4}{3\overline{F}_0}. \qquad (5.24)$$

Besides, since $\mathbb{E}[\delta_i \phi^j_{k_0,m}(T_i)] = \langle \phi^j_{k_0,m}, h \rangle_{\overline{F}_T}$,

$$\|\widehat{V}^{(k_0)}_m - V^{(k_0)}_m\|^2 = \sum_{j=0}^{r} \left(\frac{1}{n} \sum_{i=1}^{n} (\delta_i \phi^j_{k_0,m}(T_i) - \mathbb{E}[\delta_i \phi^j_{k_0,m}(T_i)]) \right)^2 \qquad (5.25)$$

and (5.22), (5.23), (5.24), (5.25) conclude the proof of Lemma 5.5.1. □

Proof of Lemma 5.5.2

We denote by $\overline{a}_j$ the coefficient $\overline{a}^j_{k_0,m}$. Then similarly to Lemma 5.5.1,

$$\begin{aligned}
(h_m - \overline{h}_m)^2(x_0) &= \left(\sum_{j=0}^{r}(a_j - \overline{a}_j)\phi^j_{k_0,m}(x_0)\right)^2 \\
&\leq \left(\sum_{j=0}^{r}(a_j - \overline{a}_j)^2)\right)\left(\sum_{j=0}^{r}(\phi^j_{k_0,m}(x_0))^2\right) \\
&= \|((G_m^{(k_0)})^{-1} - (\widehat{G}_m^{(k_0)})^{-1})V_m^{(k_0)}\|^2\left(\sum_{j=0}^{r}(\phi^j_{k_0,m}(x_0))^2\right) \\
&\leq \rho^2((G_m^{(k_0)})^{-1} - (\widehat{G}_m^{(k_0)})^{-1})\|V_m^{(k_0)}\|^2\left(D_m\sum_{j=0}^{r}\|\phi^j\|^2_\infty\right). \qquad (5.26)
\end{aligned}$$

On the one hand,

$$\|V_m^{(k_0)}\|^2 = \sum_{j=0}^{r}\langle\phi^j_{k_0,m}, h\rangle^2_{\overline{F}_T} \leq \left(\int_{(k_0-1)/D_m}^{k_0/D_m} h^2(x)dx\right)\left(\sum_{j=0}^{r}\|\phi^j_{k_0,m}\|^2_{\overline{F}_T}\right) \leq \|h\|^2_\infty\frac{r+1}{D_m}. \quad (5.27)$$

On the other hand, let $(\psi^j_{k_0,m})_{j=0,\dots,r}$ be a $\|.\|_{\overline{F}_T}$-orthonormal basis of $S_{k_0,m}$, and P be the transition matrix from $(\phi^j_{k_0,m})_{j=0,\dots,r}$ to $(\psi^j_{k_0,m})_{j=0,\dots,r}$. $P^{-1}G_m^{(k_0)}P$ is the Gram matrix of the basis $(\psi^j_{k_{00},m})_{j=0,\dots,r}$ for the scalar product $\langle.,.\rangle_{\overline{F}_T}$, so

$$P^{-1}G_m^{(k_0)}P = I.$$

We denote by $\widehat{H}_m^{(k_0)} = P^{-1}\widehat{G}_m^{(k_0)}P$, then

$$P^{-1}(\widehat{G}_m^{(k_0)})^{-1}P = (\widehat{H}_m^{(k_0)})^{-1}.$$

By classical algebra results,

$$\rho((G_m^{(k_0)})^{-1} - (\widehat{G}_m^{(k_0)})^{-1}) = \rho(P^{-1}(I - (\widehat{H}_m^{(k_0)})^{-1})P) = \rho(I - (\widehat{H}_m^{(k_0)})^{-1})$$

and

$$Sp((\widehat{H}_m^{(k_0)})^{-1}) = Sp((\widehat{G}_m^{(k_0)})^{-1}).$$

Therefore,

$$\begin{aligned}
\rho(I - (\widehat{H}_m^{(k_0)})^{-1}) &= \max\left\{|1-\mu|, \mu \in Sp((\widehat{H}_m^{(k_0)})^{-1})\right\} \\
&= \max\left\{\frac{|1-\nu|}{\nu}, \nu \in Sp(\widehat{H}_m^{(k_0)})\right\}
\end{aligned}$$

On Δ_m, $\nu \geq 3\overline{F}_0/4$ for every $\nu \in Sp(\widehat{H}_m^{(k_0)}) = Sp(\widehat{G}_m^{(k_0)})$. Hence

$$\begin{aligned}\rho(I - (\widehat{H}_m^{(k_0)})^{-1}) &\leq \frac{4}{3\overline{F}_0} \max\{|1-\nu|, \nu \in Sp(\widehat{H}_m^{(k_0)})\} \\ &= \frac{4}{3\overline{F}_0}\rho(I - \widehat{H}_m^{(k_0)}) \\ &= \frac{4}{3\overline{F}_0}\rho((G_m^{(k_0)} - \widehat{G}_m^{(k_0)})).\end{aligned}$$

Besides, for every $r \times r$ square matrix M,

$$\rho^2(M) = \sup_{\|U\|=1} \|MU\|^2 = \sup_{\sum_{j=0}^r u_j^2=1} \sum_{j=0}^r (\sum_{l=0}^r m_{j,l}u_j)^2 \leq \sum_{j,l=0}^r m_{j,l}^2.$$

Hence, according to the definition of $G_m^{(k_0)}$ and $\widehat{G}_m^{(k_0)}$,

$$\rho^2(G_m^{(k_0)} - \widehat{G}_m^{(k_0)}) \leq \sum_{j,l=0}^r \left(\langle \phi_{k_0,m}^l, \phi_{k_0,m}^j \rangle_n - \langle \phi_{k_0,m}^l, \phi_{k_0,m}^j \rangle_{\overline{F}_T}\right)^2 \tag{5.28}$$

and (5.26), (5.27), (5.28) conclude the proof of Lemma 5.5.2 □

5.5.2 Proof of Corollary 5.3.1

Assume that $h \in \mathcal{H}(\beta, L)$. According to Proposition 5.3.1, there exists a constant L' such that for every $m \in \{1, \ldots, N_n\}$,

$$\begin{aligned}\sup_{j \geq m}(h_j - h_m)^2(x_0) &\leq 2\sup_{j \geq m}(h_j - h)^2(x_0) + 2(h_m - h)^2(x_0) \\ &\leq 4(L')^2 D_m^{-2\beta}.\end{aligned}$$

Let

$$\begin{cases} G(m) = L' D_m^{-2\beta} + A x_m \frac{D_m}{n} & \forall m \in \{1, \ldots, N_n\} \\ m_1 = \arg\min_{m=1,\ldots,N_n} G(m). \end{cases}$$

For every m, $Crit(m) \leq G(m)$.

1. If $m_1 \geq m_{opt}$

$$
\begin{aligned}
Crit(m_{opt}) + (g - g_{m_{opt}})^2(x_0) &\leq Crit(m_{opt}) + 2(g_{m_1} - g_{m_{opt}})^2(x_0) + 2(g_{m_1} - g)^2(x_0) \\
&\leq Crit(m_{opt}) + 2 \sup_{j \geq m_{opt}} (g_j - g_{m_{opt}})^2(x_0) + 2(g_{m_1} - g)^2(x_0) \\
&\leq 3Crit(m_{opt}) + 2(g_{m_1} - g)^2(x_0) \\
&\leq 3Crit(m_1) + 2(g_{m_1} - g)^2(x_0) \\
&\leq 3Crit(m_1) + 2L'D_{m_1}^{-2\beta} \\
&\leq 5G(m_1).
\end{aligned}
$$

2. If $m_1 < m_{opt}$

$$
\begin{aligned}
Crit(m_{opt}) + (g - g_{m_{opt}})^2(x_0) &\leq Crit(m_{opt}) + 2(g_{m_1} - g_{m_{opt}})^2(x_0) + 2(g_{m_1} - g)^2(x_0) \\
&\leq Crit(m_1) + 2 \sup_{j \geq m_1} (g_j - g_{m_1})^2(x_0) + 2L'D_{m_1}^{-2\beta} \\
&\leq 3Crit(m_1) + 2L'D_{m_1}^{-2\beta} \\
&\leq 5G(m_1).
\end{aligned}
$$

Moreover let $m_2 \in \{1, \ldots, N_n\}$ be such that

$$
\left(\frac{n}{\log n}\right)^{1/(2\beta+1)} \leq D_{m_2} \leq 2\left(\frac{n}{\log n}\right)^{1/(2\beta+1)}.
$$

Then

$$
G(m_1) \leq G(m_2) \leq C_0 \left(\frac{n}{\log n}\right)^{-2\beta/(2\beta+1)}.
$$

Thus, with Theorem 5.3.1,

$$
\mathbb{E}\left[(\widehat{h}_{\widehat{m}} - h)^2(x_0)\right] \leq 5\kappa G(m_1) + \frac{\kappa'}{n} \leq C_1 \left(\frac{n}{\log n}\right)^{-2\beta/(2\beta+1)} \tag{5.29}
$$

which concludes the proof of Corollary 5.3.1. □

5.6 Appendix

5.6.1 Orthonormal basis of polynomials

Let $\mathcal{P}_r(I)$ denotes the set of polynomials with degree smaller than or equal to r, defined on the interval I. As a well known result, there exists a collections of polynomials $(P_0, \ldots, P_r)$, called Legendre polynomials, which form an L^2-orthogonal basis of $\mathcal{P}_r([-1, 1])$. Besides, for every j,

$$
\int_{-1}^{1} P_j^2(x)dx = \frac{2}{2j+1}.
$$

Define $\phi^j(x) = \sqrt{2j+1}P_j(2x-1)$, then $(\phi^0, \ldots, \phi^r)$ is a $\|.\|$-orthonormal basis of $\mathcal{P}_r([0,1])$. Moreover, for every $j \in \mathbb{N}$ and every $x \in [-1,1]$, $P_j(x) \leq 1$. Hence

$$\|\phi^j\|_\infty \leq \sqrt{2j+1}. \tag{5.30}$$

5.6.2 Projection on sets of piecewise polynomials

Let $f : [0,1] \to \mathbb{R}$ be a function which belongs to the Hölder space $\mathcal{H}(s+\beta, L)$, where $s \in \mathbb{N}$ and $\beta \in [0,1[$. Let μ be a non negative function defined on $[0,1]$ such that $\|\mu\|_\infty = \sup_{x\in[0,1]} \mu(x) < +\infty$. For every t,$v \in L^2([0,1])$, we denote

$$\langle t, v\rangle_\mu = \int_0^1 t(x)v(x)\mu(x)dx \quad \text{and} \quad \|t\|_\mu^2 = \int_0^1 t^2(x)\mu(x)dx.$$

Let r be an integer larger than or equal to s and S be the set of the polynomials on $[0,1]$ with degree smaller than or equal to r. Let $\{\varphi_0, \ldots, \varphi_r\}$ be a $\|.\|_\mu$-orthogonal basis of S such that $deg(\varphi_j) = j$.

Let $m \in \mathbb{N}^*$, $D_m = 2^m$ and S_m be the set of piecewise polynomials of degree r and step $1/D_m$ on $[0,1]$:

$$S_m = Vect\{\varphi_{j,k}, j = 0, \ldots, r, k = 1, \ldots, D_m\} \tag{5.31}$$

where $\varphi_{j,k}(x) = \sqrt{D_m}\varphi_j(D_m x - (k-1))\mathbb{1}_{I_{k,m}}$ and $I_{k,m} = [(k-1)/D_m, k/D_m[$. Moreover, the family $(\varphi_{j,k})$ is $\langle ., .\rangle_\mu$-orthogonal. Finally, let $f_m = \arg\min_{t\in S} \|f-t\|_\mu^2$, then,

$$f_m = \sum_{j=0,\ldots,r} \sum_{k=1,\ldots,D} a_{j,k}\varphi_{j,k}$$

where $a_{j,k} = \langle f, \varphi_{j,k}\rangle_\mu$. Then the following result holds.

Proposition 5.6.1 *There exists a constant L'' which depends on $(L, r, s, \|\mu\|_\infty)$ and on the basis $\{\varphi_j\}$ such that,*

$$\|f - f_m\|_\infty \leq L'' D_m^{-(r+\beta)} \quad \forall f \in \mathcal{H}(s+\beta, L)$$

where f_m is the $\|.\|_\mu$-projection of f on the space S_m defined in (5.31).

Proof of Proposition 5.6.1
We first prove the following lemma.

Lemma 5.6.1 *There exists a constant L' which depends on $(L, r, s, \|\mu\|_\infty)$ and on the basis (φ_j) such that for every $m \in \mathbb{N}^*$, for every $k \in \{1, \ldots, D_m\}$ and for every $x, y \in I_{k,m}$,*

$$|f_m^{(s)}(x) - f_m^{(s)}(y)| \leq L'|x-y|^\beta.$$

Let $k \in \{1, \dots, D_m\}$ and $x, y \in I_{k,m}$ be fixed. For every $z \in I_{k,m}$, $f_m(z) = \sum_{j=0}^{r} a_{j,k}\varphi_{j,k}(z)$. For every $j \in \{0, \dots, r\}$, $deg(\varphi_{j,k}) = j$ so

$$f_m^{(s)}(z) = \sum_{j=s}^{r} a_{j,k}\varphi_{j,k}^{(s)}(z) = \sum_{j=s}^{r} a_{j,k} D_m^s \sqrt{D_m}\varphi_j^{(s)}(D_m z - (k-1)).$$

$$\begin{aligned}
|f_m^{(s)}(x) - f_m^{(s)}(y)| &\leq \sum_{j=s}^{r} |a_{j,k}| D_m^{s+1/2} |\varphi_j^{(s)}(D_m x - (k-1)) - \varphi_j^{(s)}(D_m y - (k-1))| \\
&\leq \sum_{j=s}^{r} |a_{j,k}| D_m^{s+1/2} \|\varphi_j^{(s+1)}\|_\infty |(D_m x - (k-1)) - (D_m y - (k-1))| \\
&= \sum_{j=s+1}^{r} |a_{j,k}| D_m^{s+3/2} \|\varphi_j^{(s+1)}\|_\infty |x - y|
\end{aligned}$$

since $\varphi_s^{(s+1)} = 0$.

Let us upper bound the $(a_{j,k})$'s. Let $z_0 \in I_{k,m}$ be fixed, for every $z \in I_{k,m}$ there exists $z_1 \in [z, z_0]$ or $[z_0, z]$ such that

$$\begin{aligned}
f(z) &= f(z_0) + (z - z_0)f'(z_0) + \cdots + (z - z_0)^{s-1}\frac{f^{(s-1)}(z_0)}{(s-1)!} + (z - z_0)^s \frac{f^{(s)}(z_1)}{s!} \\
&= f(z_0) + (z - z_0)f'(z_0) + \cdots + (z - z_0)^{s-1}\frac{f^{(s-1)}(z_0)}{(s-1)!} + (z - z_0)^s \frac{f^{(s)}(z_0)}{s!} \\
&\quad + (z - z_0)^s \frac{f^{(s)}(z_1) - f^{(s)}(z_0)}{s!} \\
&= P(z) + (z - z_0)^s \frac{f^{(s)}(z_1) - f^{(s)}(z_0)}{s!}
\end{aligned}$$

where $P(z)$ is polynomial of degree smaller than or equal to s. Then, the coefficients $(a_{j,k})$'s of the orthogonal projection of f on S_m satisfy, for every $j \geq s + 1$,

$$a_{j,k} = \langle \varphi_{j,k}, f \rangle_\mu = \langle \varphi_{j,k}, P \rangle_\mu + \int_{I_{k,m}} \varphi_{j,k}(z)(z - z_0)^s \frac{f^{(s)}(z_1) - f^{(s)}(z_0)}{s!} \mu(z) dz.$$

Besides, $\varphi_{j,k}$ is orthogonal to every polynomial of degree smaller than j, hence $\langle \varphi_{j,k}, P \rangle = 0$.

With Cauchy Schwarz Inequality,

$$\begin{aligned} a_{j,k} &\leq \sqrt{\int_{I_{k,m}} \varphi_{j,k}^2(z)\mu(z)dz}\sqrt{\int_{I_{k,m}} (z-z_0)^{2s}\frac{|f^{(s)}(z_1)-f^{(s)}(z_0)|^2}{(s!)^2}\mu(z)dz} \\ &\leq \frac{1}{s!}\sqrt{\int_{I_{k,m}} |z-z_0|^{2s}L^2|z_1-z_0|^{2\beta}\mu(z)dz} \\ &\leq \frac{L}{s!}D_m^{-(\beta+s)}\sqrt{\int_{I_{k,m}}\mu(z)dz} \\ &= \frac{L}{s!}D_m^{-(s+1/2+\beta)}\sqrt{\|\mu\|_\infty} \end{aligned}$$

as the length of the interval $I_{k,m}$ is equal to $1/D_m$. Therefore

$$\begin{aligned} |f_m^{(s)}(x)-f_m^{(s)}(y)| &\leq \frac{L\sqrt{\|\mu\|_\infty}}{s!}\sum_{j=s+1}^{r}\|\varphi_j^{(s+1)}\|_\infty D_m^{(s+3/2)-(s+1/2+\beta)}|x-y| \\ &= \frac{L\sqrt{\|\mu\|_\infty}}{s!}\sum_{j=s+1}^{r}\|\varphi_j^{(s+1)}\|_\infty D_m^{1-\beta}|x-y|^{1-\beta}|x-y|^{\beta} \\ &= \left(\frac{L\sqrt{\|\mu\|_\infty}}{s!}\sum_{j=s+1}^{r}\|\varphi_j^{(s+1)}\|_\infty\right)|x-y|^{\beta} \end{aligned}$$

since $|x-y|^{1-\beta} \leq D_m^{-(1-\beta)}$, which ends the proof of the Lemma. □

The proof of Proposition 5.6.1 is based on a result from DeVore and Lorentz (1993)(Theorem 10.8, Chapter 3).

Proposition 5.6.2 *Let g be a continuous application on an interval $[a,b]$. Let ν be a non negative application on $[a,b]$. Let P_r be the orthogonal projection of g on S for the norm $L^2([0,1],\nu(x)dx)$, then there exist $(r+1)$ distinct points $x_0,\ldots,x_r$ such that*

$$g(x_i) = P_r(x_i), \quad \forall i = 0,\ldots,r$$

Let $x_0 \in [0,1]$ and k be such that $x_0 \in I_{k,m}$. Let us denote by g and g_m the restrictions of f and f_m to $I_{k,m}$, then obviously g_m is the $\|.\|_\mu$-projection of g on $S_{k,m}$. Thus, by Proposition 5.6.2, there exists a set of $r+1$ points $(z_0^0,\ldots,z_r^0)$ on which g and g_m are equal. By applying iteratively the Mean value Theorem to g, we obtain that for every $j \in \{0,\ldots,s\}$, there exist $(r+1-j)$ points $z_0^j,\ldots,z_{r-j}^j \in I_{k,m}$ such that

$$g^{(j)}(z_i^j) = g_m^{(j)}(z_i^j).$$

Then

$$\begin{aligned}(g-g_m)(x_0) &= \int_{z_0^0}^{x_0}(g-g_m)'(t_1)dt_1\\ &= \int_{z_0^0}^{x_0}\left((g-g_m)'(t_1)-(g'-g_m')(z_0^1)\right)dt_1\\ &= \int_{z_0^0}^{x_0}\left(\int_{z_0^1}^{t_1}(g''-g_m'')(t_2)dt_2\right)dt_1\\ &= \int_{z_0^0}^{x_0}\left(\int_{z_0^1}^{t_1}\left(\int_{z_0^2}^{t_2}(g^{(3)}-g_m^{(3)})(t_3)dt_3\right)dt_2\right)dt_1\end{aligned}$$

By iteration, we get

$$(g-g_m)(x_0) = \int_{z_0^0}^{x_0}\left(\int_{z_0^1}^{t_1}\left(\int_{z_0^2}^{t_2}\dots\int_{z_0^{s-1}}^{t_{s-1}}(g^{(s)}-g_m^{(s)})(t_s)dt_s\dots\right)dt_2\right)dt_1$$

Moreover, $g^{(s)}(z_0^s)=g_m^{(s)}(z_0^s)$, so

$$(g-g_m)(x_0)=\int_{z_0^0}^{x_0}\left(\int_{z_0^1}^{t_1}\left(\int_{z_0^2}^{t_2}\dots\int_{z_0^{s-1}}^{t_{s-1}}\left[(g^{(s)}(t_s)-g^{(s)}(z_s^0))-(g_m^{(s)}(t_s)-g_m^{(s)}(z_s^0))\right]dt_s\dots\right)dt_2\right)dt_1$$

$g\in\mathcal{H}(s+\beta,L)$ thus

$$|g^{(s)}(t_s)-g^{(s)}(z_s^0)|\le L|t_s-z_s^0|^\beta\le LD_m^{-\beta},$$

Besides Lemma 5.6.1 ensures that

$$|g_m^{(s)}(t_s)-g_m^{(s)}(z_s^0)|\le L''D_m^{-\beta}.$$

Therefore

$$(g-g_m)(x_0)=\int_{z_0^0}^{x_0}\left(\int_{z_0^1}^{t_1}\left(\int_{z_0^2}^{t_2}\dots\int_{z_0^{s-1}}^{t_{s-1}}(L+L')D_m^{-\beta}dt_s\dots\right)dt_2\right)dt_1$$

The s integrals in the above expression are computed on intervals of length $1/D_m$, hence

$$|(f-f_m)(x_0)|=|(g-g_m)(x_0)|\le(L+L')D_m^{-\beta}\times\left(\frac{1}{D_m}\right)^{-s}=L''D_m^{-(s+\beta)}.\quad\square$$

Chapitre 6

Estimation de la fonction de distribution conditionnelle à partir de données censurées par intervalle, cas I

Ce chapitre est situé dans le contexte de la censure par intervalle, cas I : soit Y une variable positive appelée temps de survie dépendant d'une covariable $X \in \mathbb{R}$, et T une variable positive appelée temps de mesure telle que Y et T sont indépendants conditionnellement à X. Nous proposons une procédure d'estimation de la fonction de distribution F de Y conditionnellement à X :

$$F(x,u) = P[Y \leq u | X = x]$$

à partir de l'observation de

$$\left(X, T, \delta = \mathbb{1}_{\{Y \leq T\}}\right),$$

basée sur un constraste de régression. Un simple calcul montre que $\mathbb{E}[\delta|X,T] = F(X,T)$, c'est à dire que F est la fonction de régression de δ sur le couple (X,T). Nous estimons donc F par minimisation d'un constraste de régression, suivie d'une étape de sélection de modèle.

En s'appuyant sur la norme empirique associée à l'échantillon $\{(X_i, T_i)\}$, qui se dégage naturellement du contraste des moindres carrés, nous proposons une approche de l'étude du risque différente de la méthode classique, et basée sur une version de l'inégalité de Talagrand pour des variables non identiquement distribuées. Enfin, nous présentons une étude minimax qui atteste l'adaptativité de notre estimateur sur des espaces de Besov anisotropes.

6.1 Introduction

In some survival analysis studies, the observation of a positive variable of interest Y called lifetime, is restricted to the knowledge of whether or not Y exceeds a random measure time T. We only observe the time T and the "current status" of the system at time T, namely $\mathbb{1}_{\{Y \leq T\}}$. Such data arise naturally for example in infectious disease studies, when the time Y of infection is unobserved, and a test is carried out at time T. The lifetime Y may depend on observed covariates $\mathbf{X}$, and a general assumption in such models is that T depends on $\mathbf{X}$, and Y and T are independent given $\mathbf{X}$.

Current status data have been widely studied for the last two decades. Most results about nonparametric estimation of the survival function are based on NPMLE (Nonparametric Maximum Likelihood Estimator). Groeneboom and Wellner (1992) prove that the NPMLE is pointwise convergent at rate $n^{-1/3}$ which is the optimal rate, and van de Geer (1993) establishes a similar result for the L^2-risk. This unusual rate of convergence differs from the uncensored and right-censored cases, in which the distribution function can be estimated with the parametric rate of convergence $n^{-1/2}$. Besides, as far as the author knows, no minimax rate of convergence has been computed on classical regularity spaces. More recently, estimators developed from the NPMLE allow to take into account the knwon regularity of the function. Hudgens et al. (2007) build three estimators derived from the NPMLE, and compare their performances on simulated and real data. van der Laan and van der Vaart (2006) apply smoothing methods to the NPLME to estimate the survival function from current status data in presence of high dimensional covariates. Birgé (1999) proposes an easily computable histogram estimator which reaches the minimax rate of convergence. Nevertheless the procedures proposed in these papers are not adaptive on classical regularity spaces. Few results about adaptivity are available, and they do not include covariates: Ma and Kosorok (2006) introduce a NPMLE and a least square estimator on Sobolev classes, and select the regularity parameter with a penalised criterion. Brunel and Comte (2009) consider a least-square estimator on classical bases and introduce a model selection procedure with a more easily computable penalty function.

In this chapter we consider an i.i.d. sample $(X_i, Y_i)_{i=1,\dots,n}$, where the (X_i)'s are i.i.d. random variables with common density f_X, and the (Y_i)'s are positive variables called survival times. For every i, Y_i depends on X_i, and we denote by $F(x, y)$ the cumulative distribution function (c.d.f.) of Y_i given X_i, namely

$$F(x, y) = P[Y \leq y | X = x]$$

where $P[E_1|E_2]$ denotes the conditional probability of E_1 given E_2. We consider an i.i.d. sample $(T_i)_{i=1,\dots,n}$ of positive random variables such that for every $i \in \{1, \dots, n\}$, T_i and Y_i are independent given X_i, and we observe the sample

$$(T_i, \delta_i = \mathbb{1}_{Y_i \leq T_i}, X_i)_{i=1,\dots,n}. \tag{6.1}$$

We present an estimator of the conditional cumulative distribution function F from the sample described by (6.1). Moreover we compute the minimax rate of convergence over anisotropic Besov balls and prove that our estimator is minimax. The procedure, inspired from Brunel and Comte (2009), is based on the following heuristic. For every (x, u),

$$\mathbb{E}[\delta|(X,T) = (x,u)] = \mathbb{E}[\mathbb{1}_{Y \leq T}|(X,T) = (x,u)].$$

Given $X = x$, Y and T are independent, thus

$$\mathbb{E}[\delta|(X,T) = (x,u)] = \mathbb{E}[\mathbb{1}_{Y \leq u}|X = x] = P[Y \leq u|X = x] = F(x,u).$$

Thus F is the regression function of δ over (X,T), and the interval censoring issue turns into a regression function estimation problem where all the variables involved (X,T,δ) are observed. We consider a collection of linear subset of $L^2(\mathbb{R}^2)$, and build an estimator by minimisation of a least-square contrast on each subset. Then a model selection criterion provides an estimator whose rate of convergence is the one of the best estimator among the collection. More precisely, we get two oracle inequalities. Considering the risk associated to the empirical norm

$$\|\widehat{F} - F\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}(\widehat{F} - F)^2(X_i, T_i),$$

a nonclassical use of Talagrand Inequality states the adaptivity of our estimator under weak assumptions. Besides, the result can be extended to non random observation times (T_i)'s. Oracle inequality for the integrated risk requires classical assumptions about the collection of models and demands in particular that F is regular enough. Nevertheless, considering the integrated risk enables us to conduct a minimax study and prove that our estimator is optimal over anisotropic Besov balls $\mathcal{B}^{\beta}_{2,\infty}(L)$.

The chapter is organised as follows. General assumptions, estimation procedure and main result are presented in Section 6.2. In Section 6.3, we study the rate of convergence of the estimator over anisotropic Besov balls and prove that it is minimax. Section 6.4 is devoted to the proofs. Section 6.5 presents a version of Talagrand Inequality for non identically distribued variables and a linear algebra technical lemma.

6.2 Definition of the estimator, main assumptions and main result

6.2.1 Notations

For every i.i.d. random variables $\{V_i, W_i\}$, we denote by f_V the density of V_i and by $f_{V|W}(v,w)$ the conditional density of V_i at v given $W_i = w$ for every i.

We estimate $F(x,y)$ on a compact $A = A_1 \times A_2$ where A_1 is a compact interval of $\mathbb{R}$, and $A_2 = [0, a_2]$ for some positive a_2. Let $lg(A_i)$ denotes the length of the interval A_i.

For every t, $s \in L^2(A)$, let

$$\langle s,t\rangle_n = \frac{1}{n}\sum_{i=1}^{n} s(X_i,T_i)t(X_i,T_i)$$

and

$$\langle s,t\rangle_{f_{(X,T)}} = \int_{x\in A_1}\int_{u\in A_2} s(x,u)t(x,u)f_{(X,T)}(x,u)dudx.$$

6.2.2 Collection of models

In order to estimate F, we define a collection of finite-dimensional linear subsets of $L^2(A)$ called models. These models are constructed as tensor products of models on A_1 and A_2. Let $\mathcal{M}_n^{(1)} = \{S_{m_1}^{(1)}, m_1 \in I_n^{(1)}\}$ be a collection of linear subsets of $L^2(A_1)$ where $Dim(S_{m_1}^{(1)}) = D_{m_1}^{(1)} < +\infty$ and $(\phi_k^{m_1})_{k=1,\dots,D_{m_1}^{(1)}}$ is an orthonormal basis of $S_{m_1}^{(1)}$, for every $m_1 \in I_n^{(1)}$. Let $\mathcal{M}_n^{(2)} = \{S_{m_2}^{(2)}, m_2 \in I_n^{(2)}\}$ be a collection of linear subsets of $L^2(A_2)$ where $Dim(S_{m_2}^{(2)}) = D_{m_2}^{(2)} < +\infty$ and $(\psi_k^{m_2})_{k=1,\dots,D_{m_2}^{(2)}}$ is an orthonormal basis of $S_{m_2}^{(2)}$, for every $m_2 \in I_n^{(2)}$. Then for every $m = (m_1, m_2) \in I_n = I_n^{(1)} \times I_n^{(2)}$, we define

$$S_m = \left\{ t : A \to \mathbb{R}, \quad t(x,y) = \sum_{k=1,\dots,D_{m_1}^{(1)},\ l=1,\dots,D_{m_2}^{(2)}} a_{k,l}\phi_k^{m_1}(x)\psi_l^{m_2}(y) \right\}.$$

The family $\{\phi_k^{m_1}\phi_l^{m_2}, k = 1,\dots,D_{m_1}^{(1)},\ l = 1,\dots,D_{m_2}^{(2)}\}$ is an orthonormal basis of S_m, and the dimension of S_m is $D_m = D_{m_1}^{(1)}D_{m_2}^{(2)}$. We consider the collection $\mathcal{M}_n = \{S_m, m = (m_1,m_2) \in I_n\}$. We assume that the following assumption holds.

(**H**): Let $j = 1$ or 2. For every $b > 0$, there exists a constant B_j such that

$$\sum_{m_j \in I_n^{(j)}} \exp\left(-b\sqrt{D_{m_j}}\right) \le B_j, \quad \forall n \in \mathbb{N}^*.$$

Assumption (**H**) about collections $\mathcal{M}_n^{(j)}$ implies a similar result for $\mathcal{M}_n$. Indeed, let $b > 0$

$$\sum_{m\in I_n} \exp\left(-b\sqrt{D_m}\right) = \sum_{m_1\in I_n^{(1)}} \left(\sum_{m_2\in I_n^{(2)}} \exp\left(-b\sqrt{D_{m_1}^{(1)}}\sqrt{D_{m_2}^{(2)}}\right)\right).$$

Besides for every $x, y \geq 1$, $2xy \geq x + y$, thus

$$\begin{aligned}
\sum_{m \in I_n} \exp\left(-b\sqrt{D_m}\right) &\leq \sum_{m_1 \in I_n^{(1)}} \left(\sum_{m_2 \in I_n^{(2)}} \exp\left(-\frac{b}{2}(\sqrt{D_{m_1}^{(1)}} + \sqrt{D_{m_2}^{(2)}}\right)\right) \\
&= \left(\sum_{m_1 \in I_n^{(1)}} \exp\left(-\frac{b}{2}\sqrt{D_{m_1}^{(1)}}\right)\right) \left(\sum_{m_2 \in I_n^{(2)}} \exp\left(-\frac{b}{2}\sqrt{D_{m_2}^{(2)}}\right)\right) \\
&\leq B_1' B_2'
\end{aligned}$$

for some positive B_1', B_2'.

6.2.3 Regression contrast

Let $(x, u) \in \mathbb{R}^2$ be such that $f_{(X,T)}(x, u) > 0$.

$$\begin{aligned}
\mathbb{E}\left[\delta_1 | (X_1, T_1) = (x, u)\right] &= \mathbb{E}[\mathbb{1}_{Y_1 \leq T_1} | (X_1, T_1) = (x, u)] \\
&= \mathbb{E}\left[\mathbb{1}_{\{Y_1 \leq u\}} | (X_1, T_1) = (x, u)\right] \\
&= \int_{A_2} \mathbb{1}_{\{y \leq u\}} f_{Y|(X,T)}(y, x, u) dy \\
&= \int_{A_2} \mathbb{1}_{\{y \leq u\}} \frac{f_{(Y,X,T)}(y, x, u)}{f_{(X,T)}(x, u)} dy \\
&= \int_{A_2} \mathbb{1}_{\{y \leq u\}} \frac{f_{(Y,T)|X}(y, u, x) f_X(x)}{f_{(X,T)}(x, u)} dy.
\end{aligned}$$

Y_1 and T_1 are independent given X_1: for every $(x, y, u) \in A_1 \times A_2 \times A_2$,

$$f_{(Y,T)|X}(y, u, x) = f_{Y|X}(y, x) f_{T|X}(u, x).$$

Hence

$$\begin{aligned}
\mathbb{E}\left[\delta_1 | (X_1, T_1) = (x, u)\right] &= \int_{A_2} \mathbb{1}_{\{y \leq u\}} \frac{f_{Y|X}(y, x) f_{T|X}(u, x) f_X(x)}{f_{(X,T)}(x, u)} dy \\
&= \int_{A_2} \mathbb{1}_{\{y \leq u\}} \frac{f_{Y|X}(y, x) f_{(X,T)}(x, u)}{f_{(X,T)}(x, u)} dy \\
&= \int_{A_2} \mathbb{1}_{\{y \leq u\}} f_{Y|X}(y, x) dy \\
&= F(x, u).
\end{aligned}$$

It amounts to say that F is the regression function of δ_1 over (X_1, T_1). Thus we consider the least-square contrast, already use to compute the regression function estimator in Chapters 2 and 3, and which is classical in regression function estimation. For every $t \in L^2(A)$

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n} \left(t(X_i, T_i) - \delta_i\right)^2 .$$

$\gamma_n(t)$ measures how the $\{t(X_i, T_i)\}$'s approximate the $\{\delta_i\}$'s. As

$$\mathbb{E}[(\delta_i - t(X_i, T_i))^2] = \|F - t\|^2_{f_{(X,T)}} + \mathbb{E}[F(X_i, T_i) - F^2(X_i, T_i)]$$

and $\mathbb{E}[F(X_i, T_i) - F^2(X_i, T_i)]$ is independent of t, a minimiser of the mean of $(\delta_i - t(X_i, T_i))^2$ is a relevant estimator of F.

6.2.4 Minimum contrast estimators

Let S_m be a model in $\mathcal{M}_n$. We define

$$\widehat{F}_m = \arg\min_{t \in S_m} \gamma_n(t). \tag{6.2}$$

For sake of simplicity, we denote the index set $J_m = \{(k, l), k = 1, \ldots, D^{(1)}_{m_1}, l = 1, \ldots, D^{(2)}_{m_2}\}$ as a vector:

$$J_m = \left((1,1), \ldots, (1, D^{(2)}_{m_2}), (2,1), \ldots, (2, D^{(2)}_{m_2}), \ldots, (D^{(1)}_{m_1}, 1), \ldots, (D^{(1)}_{m_1}, D^{(2)}_{m_2})\right) \tag{6.3}$$

and $\widehat{F}_m(x, u) = \sum_{(k,l)\in J_m} \widehat{a}_{k,l}\phi^{m_1}_k(x)\psi^{m_2}_l(u)$. Then (6.2) is equivalent to

$$\widehat{G}_m \widehat{A}_m = \widehat{V}_m \tag{6.4}$$

where $\widehat{A}_m = [\widehat{a}_{k,l}]_{(k,l)\in J_m}$ is a column vector,

$$\widehat{G}_m = \left[\frac{1}{n}\sum_{i=1}^{n} \phi^{m_1}_k(X_i)\psi^{m_2}_l(T_i)\phi^{m_1}_{k'}(X_i)\psi^{m_2}_{l'}(T_i)\right]_{((k,l),(k',l'))\in J_m \times J_m}$$

is a $D_m \times D_m$-square matrix and

$$\widehat{V}_m = \left[\frac{1}{n}\sum_{i=1}^{n} \phi^{m_1}_k(X_i)\psi^{m_2}_l(T_i)\delta_i\right]_{(k,l)\in J_m}$$

is a column vector. The matrix $\widehat{G}_m$ is the Gram matrix related to $\{\phi^{m_1}_j \psi^{m_2}_l\}_{(k,l)\in J_m}$ for the scalar product $\langle ., . \rangle_n$.

We examinate the existence and unicity of $\widehat{F}_m$. Let $\widehat{S}_m$ be the subset of $\mathbb{R}^n$ defined by

$$\widehat{S}_m = \{(t(X_1, T_1), \dots, t(X_n, T_n)) \, , t \in S_m\}.$$

and $\widehat{Z}_m = \arg\min_{Z \in \widehat{S}_m} \frac{1}{n} \sum_{i=1}^n (Z_i - \delta_i)^2$. $\widehat{Z}_m$ is the projection of $(\delta_1, \dots, \delta_n)$ on $\widehat{S}_m$ for the canonical norm on $\mathbb{R}^n$, so $\widehat{Z}_m$ is uniquely defined. Moreover, by definition of $\widehat{S}_m$, there exists at least one function $G \in S_m$ such that $\widehat{Z}_m = (G(X_1, T_1), \dots, G(X_n, T_n))$, then G minimises $\gamma_n(t)$ on S_m. Moreover, if two such functions G exist, they are equal on the $\{(X_i, T_i)\}$'s, so $\|\widehat{F}_m - F\|_n^2$ remains the same. For that reason, the definition of $\arg\min_{t \in S_m} \gamma_n(t)$ is sensible for the risk $\mathbb{E}\left[\|\widehat{F}_m - F\|_n^2 | \{(X_i, T_i)\}_{i=1,\dots,n}\right]$

Let $\widehat{F}_m$ be any minimiser of γ_n on S_m. To prove the results of Section 6.3, we need our estimator to be bounded almost surely, thus we set for every $(x, u) \in A$

$$\tilde{F}_m(x,u) = \begin{cases} 0 & \text{if} \quad \widehat{F}_m(x,u) < 0 \\ 1 & \text{if} \quad \widehat{F}_m(x,u) > 1 \\ \widehat{F}_m(x,u) & \text{otherwise.} \end{cases}$$

Remark 11 *For every $(x, u) \in A$, $F(x, u) \in [0, 1]$. Hence almost surely,*

$$|\tilde{F}_m(x,u) - F(x,u)| \leq |\widehat{F}_m(x,u) - F(x,u)|, \quad \forall (x,u) \in A, \quad \forall m \in \mathbb{N}^*.$$

In particular,

$$\|\tilde{F}_m - F\|_n^2 \leq \|\widehat{F}_m - F\|_n^2 \quad a.s.$$

Thus, any upper bound of $\mathbb{E}\left[\|\widehat{F}_m - F\|_n^2 | \{(X_i, T_i)\}_{i=1,\dots,n}\right]$ is an upper bound of $\mathbb{E}\left[\|\tilde{F}_m - F\|_n^2 | \{(X_i, T_i)\}_{i=1,\dots,n}\right]$.

6.2.5 Bias-variance decomposition and model selection procedure

The estimation procedure from Section 6.2.4 provides a collection of estimators $\{\tilde{F}_m, m \in I_n\}$, among which one is automatically selected by a data driven procedure to be adaptive for the risk $\mathbb{E}\left[\|\widehat{F}_m - F\|_n^2 | \{(X_i, T_i)\}_{i=1,\dots,n}\right]$. For every $m \in I_n$, the risk $\mathbb{E}\left[\|\widehat{F}_m - F\|_n^2 | \{(X_i, T_i)\}_{i=1,\dots,n}\right]$ splits in bias and variance. With Pythagoras formula, almost surely,

$$\mathbb{E}\left[\|\widehat{F}_m - F\|_n^2 | \{(X_i, T_i)\}_{i=1,\dots,n}\right] = \|F - F_m\|_n^2 + \mathbb{E}\left[\|\widehat{F}_m - F_m\|_n^2 | \{(X_i, T_i)\}_{i=1,\dots,n}\right] \qquad (6.5)$$

where $F_m = \arg\min_{t \in S_m} \|F - t\|_n^2$. The best model among the collection, called the oracle, is the one which minimises the right-hand side in (6.5), but it is unknown since F and F_m are

unobserved. Hence we construct an estimator of this bias-variance sum and select the model $\widehat{m}$ which minimises it. More precisely, let $(x_1, \dots, x_n) \in A_1^n$, $(u_1, \dots, u_n) \in A_2^n$,

$$\mathcal{A} = \{X_1 = x_1, \dots, X_n = x_n, T_1 = u_1, \dots, T_n = u_n\}$$

and for every $s, t \in L^2(A)$,

$$\langle s, t \rangle_0 = \frac{1}{n} \sum_{i=1}^{n} t(x_i, u_i) s(x_i, u_i) \quad \text{and} \quad \|t\|_0^2 = \frac{1}{n} \sum_{i=1}^{n} t^2(x_i, u_i). \tag{6.6}$$

Then

$$\mathbb{E}\left[\|\widehat{F}_m - F\|_n^2 | \mathcal{A}\right] = \|F - F_m^0\|_0^2 + \mathbb{E}\left[\|\widehat{F}_m - F_m^0\|_0^2 | \mathcal{A}\right]$$

where $F_m^0 = \arg\min_{t \in S_m} \|F - t\|_0$. Let $(\varphi_\lambda)_{\lambda \in I_m}$ be a $\|.\|_0$-orthogonal basis of S_m such that $\|\varphi\|_0 = 0$ or 1 (see Lemma 6.5.1 states the existence of such a basis), then

$$F_m^0 = \sum_{\lambda \in I_m} \langle \varphi_\lambda, F \rangle_0 \varphi_\lambda \quad \text{and} \quad \widehat{F}_m = \sum_{\lambda \in I_m} \left(\frac{1}{n} \sum_{i=1}^{n} \varphi_\lambda(x_i, u_i) \delta_i \right) \varphi_\lambda.$$

Hence

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{F}_m - F_m^0\|_0^2 | \mathcal{A}\right] &= \sum_{\lambda \in I_m} \mathbb{E}\left[\left(\frac{1}{n} \sum_{i=1}^{n} (\varphi_\lambda(x_i, u_i)\delta_i - \langle \varphi_\lambda, F \rangle_0) \right)^2 | \mathcal{A} \right] \\
&= \sum_{\lambda \in I_m} \mathbb{E}\left[\left(\frac{1}{n} \sum_{i=1}^{n} \varphi_\lambda(x_i, u_i) \left(\mathbb{1}_{\{Y_i \le u_i\}} - \mathbb{E}[\mathbb{1}_{\{Y_i \le u_i\}}] \right) \right)^2 | \mathcal{A} \right] \\
&= \sum_{\lambda \in I_m} \frac{1}{n^2} \sum_{i,l=1}^{n} \varphi_\lambda(x_i, u_i) \varphi_\lambda(x_l, u_l) \mathbb{E}\left[\left(\mathbb{1}_{\{Y_i \le u_i\}} - \mathbb{E}[\mathbb{1}_{\{Y_i \le u_i\}}] \right) \left(\mathbb{1}_{\{Y_l \le u_l\}} - \mathbb{E}[\mathbb{1}_{\{Y_l \le u_l\}}] \right) | \mathcal{A} \right].
\end{aligned}$$

Given $\mathcal{A}$, the (Y_i)'s are independent, and so are the $\left(\mathbb{1}_{\{Y_i \le u_i\}}\right)$'s. Therefore for every $i \neq l$,

$$\mathbb{E}\left[\left(\mathbb{1}_{\{Y_i \le u_i\}} - \mathbb{E}[\mathbb{1}_{\{Y_i \le u_i\}}] \right) \left(\mathbb{1}_{\{Y_l \le u_l\}} - \mathbb{E}[\mathbb{1}_{\{Y_l \le u_l\}}] \right) | \mathcal{A} \right] = 0$$

and

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{F}_m - F_m^0\|_0^2 | \mathcal{A}\right] &= \sum_{\lambda \in I_m} \frac{1}{n^2} \sum_{i=1}^{n} \varphi_\lambda^2(x_i, u_i) \mathbb{E}\left[\left(\mathbb{1}_{\{Y_i \le u_i\}} - \mathbb{E}[\mathbb{1}_{\{Y_i \le u_i\}}] \right)^2 | \mathcal{A} \right] \\
&= \sum_{\lambda \in I_m} \frac{1}{n^2} \sum_{i=1}^{n} \varphi_\lambda^2(x_i, u_i) F(x_i, u_i)(1 - F(x_i, u_i)).
\end{aligned}$$

For every i, $F(x_i,u_i)\in[0,1]$ so $F(x_i,u_i)(1-F(x_i,u_i))\in[0,1/4]$, hence

$$\mathbb{E}\left[\|\widehat{F}_m - F_m^0\|_0^2|\mathcal{A}\right] \leq \frac{1}{4n}\sum_{\lambda\in I_m}\|\varphi_\lambda\|_0^2 \leq \frac{D_m}{4n}.$$

Thus the variance term is upper bounded by a term of order D_m/n. Moreover

$$\|F_m^0 - F\|_0^2 = \min_{t\in S_m}\frac{1}{n}\sum_{i=1}^n (t(x_i,u_i)-F(x_i,u_i))^2 = \min_{t\in S_m}\frac{1}{n}\sum_{i=1}^n (t(x_i,u_i)-\mathbb{E}[\delta_i|(X_i,T_i)=(x_i,u_i)])^2$$

which is naturally estimated on $\mathcal{A}$ by

$$\min_{t\in S_m}\frac{1}{n}\sum_{i=1}^n (t(x_i,u_i)-\delta_i)^2 = \gamma_n(\widehat{F}_m)$$

Finally, we select the following model:

$$\widehat{m} = \arg\min_{m\in I_n}\left[\gamma_n(\widehat{F}_m) + pen(m)\right]$$

where $pen(m)=\theta D_m/n$ for some numerical constant $\theta>1$. Our estimator of F is $\widehat{F}_{\widehat{m}}$.

Remark 12 *The condition on θ could be weakened to $\theta>1/4$ with slight changes in the proofs, but we assume that $\theta>1$ for sake of simplicity.*

6.2.6 Risk for the empirical norm

In Sections 6.2.6 and 6.3.1, two similar results are stated: the model selection estimator is proved to converge at the same rate as the best estimator among the collection, on the one hand for risk associated to the empirical norm $\|.\|_n$ (Theorem 6.2.1) and on the other hand for the risk associated to the integrated norm $\|.\|_{f_{(X,T)}}$ (Corollary 6.3.1). These two results are presented separately to underline the fact that very few assumptions are required to upper bound the $\|.\|_n^2$-risk, and that stronger assumptions about the collection of models arise to obtain the equivalence of the norms $\|.\|_n$ and $\|.\|_{f_{(X,T)}}$ on the models $\{S_m, m\in I_n\}$.

Theorem 6.2.1 *Assume that Assumption* (**H**) *holds, there exist numerical constants C_1 and C_2 such that almost surely,*

$$\mathbb{E}\left[\|\widehat{F}_{\widehat{m}} - F\|_n^2|\{(X_i,T_i)\}_{i=1,\dots,n}\right] \leq C_1\inf_{m\in I_n}\left\{\inf_{t\in S_m}\|F-t\|_n^2 + pen(m)\right\} + \frac{C_2}{n}. \tag{6.7}$$

Comments

1. For every model $m \in I_n$, $\{\inf_{t\in S_m} \|F-t\|_n^2 + pen(m)\}$ has the same order as $\|\widehat{F}_m - F\|_n^2$ (see Section 6.2.5). Thus Theorem 6.2.1 indicates that almost surely, the model selection estimator $\widehat{F}_{\widehat{m}}$ converges as fast as the best model among the collection, up to a multiplicative constant.

2. Taking expectation in both sides of (6.7), we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{F}_{\widehat{m}} - F\|_n^2\right] &\leq C_1 \inf_{m\in I_n}\left\{\mathbb{E}\left[\inf_{t\in S_m}\|F-t\|_n^2\right] + pen(m)\right\} + \frac{C_2}{n} \\
&\leq C_1 \inf_{m\in I_n}\left\{\inf_{t\in S_m}\mathbb{E}\left[\|F-t\|_n^2\right] + pen(m)\right\} + \frac{C_2}{n} \\
&= C_1 \inf_{m\in I_n}\left\{\inf_{t\in S_m}\|F-t\|_{f_{(X,T)}}^2 + pen(m)\right\} + \frac{C_2}{n}.
\end{aligned}$$

3. It is clear that the same result holds with non random observation times $(T_1,\ldots,T_n)$.

4. By Remark 11,

$$\mathbb{E}\left[\|\tilde{F}_{\widehat{m}} - F\|_n^2 | \{(X_i,T_i)\}_{i=1,\ldots,n}\right] \leq C_1 \inf_{m\in I_n}\left\{\inf_{t\in S_m}\|F-t\|_n^2 + pen(m)\right\} + \frac{C_2}{n}.$$

6.3 Minimax rate of convergence on anisotropic Besov balls $\mathcal{B}_{2,\infty}^{\beta}(A,L)$

In this section we study the minimax rate of convergence for the distribution function F on anisotropic Besov balls, and prove that our estimator reaches it.

6.3.1 Risk for the integrated norm

In order to prove that our estimator is minimax over Besov balls, we need to study the rate of convergence for the non empirical risk $\mathbb{E}\left[\|\tilde{F}_{\widehat{m}} - F\|_{f_{(X,T)}}^2\right]$. We state a result similar to Theorem 6.2.1 provided additional conditions.

($\mathbf{A_1}$): There exist $h_0 > 0$ and $h_1 < +\infty$ such that

$$h_0 \leq f_{(X,T)}(x,u) \leq h_1, \quad \forall (x,u) \in A.$$

($\mathbf{A_2}$): For $j = 1$ and 2, there exists a model $S_n^{(j)} \in \mathcal{M}_n^{(j)}$, such that for every $m_j \in I_n^{(j)}$, $S_{m_j}^{(j)} \subset S_n^{(j)}$. Let $N_n^{(j)} = Dim(S_n^{(j)})$. Besides, there exists a polynomial $P^{(j)}$ such that

$$Card(\mathcal{M}_n^{(j)}) \leq P^{(j)}(n), \quad \forall n \in \mathbb{N}. \tag{6.8}$$

($\mathbf{A_3}$): There exists a positive constant K_1 such that, for every $m_1 \in I_n^{(1)}$,

$$\sup_{x \in A_1} \sum_{k=1}^{D_{m_1}^{(1)}} (\phi_k^{m_1}(x))^2 \leq K_1 D_{m_1}^{(1)}$$

and a positive constant K_2 such that, for every $m_2 \in I_n^{(2)}$,

$$\sup_{u \in A_2} \sum_{l=1}^{D_{m_2}^{(2)}} (\psi_l^{m_2}(u))^2 \leq K_2 D_{m_2}^{(2)}.$$

Moreover $N_n^{(1)} N_n^{(2)} \leq \sqrt{n}/\log n$.

The following Proposition states a similar result for the two variable models.

Proposition 6.3.1 *Assume that* ($\mathbf{A_1}$), ($\mathbf{A_2}$) *and* ($\mathbf{A_3}$) *hold.*

1. *For every* $n \in \mathbb{N}$, $Card(\mathcal{M}_n) \leq P(n) = P^{(1)}(n)P^{(2)}(n)$.
2. *For every* $m \in I_n$,

$$\sup_{(x,u) \in A} \left(\sum_{k=1,\ldots,D_{m_1}^{(1)},\ l=1,\ldots,D_{m_2}^{(2)}} (\phi_k^{m_1}(x)\psi_l^{m_2}(u))^2 \right) \leq K D_m \tag{6.9}$$

   where $K = K_1 K_2$.

3. *For every* $m \in I_n$, $S_m \subset S_n$ *where*

$$S_n = \left\{ t : A \to \mathbb{R}, t(x,u) = \sum_{k=1,\ldots,N_n^{(1)},\ l=1,\ldots,N_n^{(2)}} a_{k,l} \phi_k^n(x) \psi_l^n(u) \right\}.$$

Under these additional conditions, Theorem 6.2.1 leads to the following result.

Corollary 6.3.1 *Assume that* ($\mathbf{H}$), ($\mathbf{A_1}$), ($\mathbf{A_2}$) *and* ($\mathbf{A_3}$) *hold then*

$$\mathbb{E}\left[\|\tilde{F}_{\widehat{m}} - F\|^2_{f_{(X,T)}} \right] \leq C_3 \inf_{m \in I_n} \left\{ \inf_{t \in S_m} \|F - t\|^2_{f_{(X,T)}} + pen(m) \right\} + \frac{C_4}{n}$$

where C_3 *is a numerical constant and* C_4 *depends on* h_0 *and* K.

Comment Corollary 6.3.1 indicates that the rate of convergence of $\tilde{F}_{\widehat{m}}$ for the $\|.\|_{f_{(X,T)}}$-risk is the one of the best estimator among the collection $\{\tilde{F}_m, m \in I_n\}$ (see Comment (1) after Theorem 6.2.1)

6.3.2 Definition of anisotropic Besov spaces

We recall the definition of two-dimensional Besov spaces stated for example in Hochmuth (2002). Let $\Omega \subset \mathbb{R}^2$, and $f \in L^2(\Omega)$. For $j = 1$ or 2, $r \in \mathbb{N}^*$ and $h > 0$, let

$$\Delta_{h,j}^r(f)(x,y) = \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} f((x,y) + khe_j)$$

be the directional partial difference operator for every $(x,y) \in \Omega_{h,j}^r$ where

$$\Omega_{h,j}^r = \{(x,y) \in \Omega, (x,y) + rhe_j \in \Omega\}$$

and (e_1, e_2) is the canonical basis of $\mathbb{R}^2$. For $t > 0$, let

$$\omega_{r,j}(f,t,\Omega) = \sup_{|h| \leq t} \|\Delta_{h,j}^r(f)(x,y)\|_{L^2(\Omega_{h,j}^r)}$$

be the directional modulus of smoothness for the L^2-norm. Let $\beta = (\beta_1, \beta_2) \in (\mathbb{R}_+^*)^2$ and $r_j = \lfloor \beta_j \rfloor + 1$ where $\lfloor \beta_j \rfloor$ denotes the integer part of β_j. We define the anisotropic Besov space of parameters $(\beta, 2, \infty)$ as

$$\mathcal{B}_{2,\infty}^{\beta}(\Omega) = \left\{ f \in L^2(\Omega), |f|_{\mathcal{B}_{2,\infty}^{\beta}(\Omega)} < +\infty \right\}$$

where

$$|f|_{\mathcal{B}_{2,\infty}^{\beta}(\Omega)} = \sup_{t>0} \left[t^{-\beta_1} \omega_{r_1,1}(f,t,\Omega) + t^{-\beta_2} \omega_{r_2,2}(f,t,\Omega) \right].$$

We consider the following norm on $\mathcal{B}_{2,\infty}^{\beta}(\Omega)$,

$$\|f\|_{\mathcal{B}_{2,\infty}^{\beta}(\Omega)} = |f|_{\mathcal{B}_{2,\infty}^{\beta}(\Omega)} + \|f\|.$$

6.3.3 Rate of convergence of $\tilde{F}_{\widehat{m}}$ on anisotropic Besov balls

For classical collections of models, the bias term $\inf_{t \in S_m} \|F - t\|_{f(X,T)}^2$ is upper bounded with the following lemma, proved in Lacour (2007) based on papers from Hochmuth (2002) and Nikol'skii (1975).

Lemma 6.3.1 *Assume that $F \in \mathcal{B}_{2,\infty}^{\beta}(A, L)$ where*

$$\mathcal{B}_{2,\infty}^{\beta}(A, L) = \left\{ F \in \mathcal{B}_{2,\infty}^{\beta}(A), \|F\|_{\mathcal{B}_{2,\infty}^{\beta}(A)} \leq L \right\}$$

for some $L > 0$ and $\beta = (\beta_1, \beta_2) \in (\mathbb{R}_+^*)^2$. For $j = 1$ and 2, and for $m_j \in \mathcal{M}_n^{(i)}$ assume that the space $S_m^{(j)}$ is one of the following.

- $S_{m_j}^{(j)}$ *is the set of piecewise polynomials with maximum degree* $s_j > \beta_j - 1$, *and step* $lg(A_j)/D_{m_j}^{(j)}$.
- $S_{m_j}^{(j)} = Vect\{\psi_{l,k}, l \leq m_j, k \in \mathbb{Z}\}$ *where* ψ *is a mother wavelet with regularity* $s_j > \beta_j - 1$, $\psi_{l,k}(x) = 2^{l/2}\psi(2^l x - k)$ *and* $D_{m_j}^{(j)} = 2^{m_j}$.
- $S_{m_j}^{(j)}$ *is the set of trigonometric polynomials with degree smaller than or equal to* $D_{m_j}^{(j)}$.

Then there exists a positive constant C_0 *such that*

$$\inf_{t \in S_m} \|F - t\| \leq C_0 \left((D_{m_1}^{(1)})^{-\beta_1} + (D_{m_2}^{(2)})^{-\beta_2} \right).$$

Plugging the result of Lemma 6.3.1 in Corollary 6.3.2 provides the rate of convergence of the risk of $\tilde{F}_{\widehat{m}}$.

Corollary 6.3.2 *Assume that* $F \in \mathcal{B}_{2,\infty}^{\beta_1,\beta_2}(A, L)$ *with* β_1, $\beta_2 > 1$ *Let* $\mathcal{M}_n^{(1)}$ *and* $\mathcal{M}_n^{(2)}$ *be collections set up from linear spaces described in Lemma* 6.3.1, *with*

$$N_n^{(j)} \leq \left(\frac{n}{\log^2 n} \right)^{1/4} \quad for \quad j = 1, 2. \tag{6.10}$$

Assume that the assumptions of Corollary 6.3.1 *hold, then there exists a positive constant* C_5 *such that*

$$\mathbb{E}\left[\|\tilde{F}_{\widehat{m}} - F\|^2 \right] \leq C_5 n^{-\overline{\beta}/(\overline{\beta}+1)}$$

where

$$\frac{2}{\overline{\beta}} = \frac{1}{\beta_1} + \frac{1}{\beta_2}.$$

Indeed, for every $m = (m_1, m_2)$,

$$\begin{aligned}
\mathbb{E}\left[\|\tilde{F}_{\widehat{m}} - F\|_{f(X,T)}^2 \right] &\leq C_3 \left\{ \inf_{t \in S_m} \|F - t\|_{f(X,T)}^2 + pen(m) \right\} + \frac{C_4}{n} \\
&\leq C_3 \left\{ h_1 \inf_{t \in S_m} \|F - t\|^2 + pen(m) \right\} + \frac{C_4}{n} \\
&\leq C_3 \left\{ 2h_1 C_0 \left((D_{m_1}^{(1)})^{-2\beta_1} + (D_{m_2}^{(2)})^{-2\beta_2} \right) + \theta \frac{D_{m_1} D_{m_2}}{n} \right\} + \frac{C_4}{n}.
\end{aligned}$$

Let $\overline{m}_1$ and $\overline{m}_2$ be such that

$$1 \leq D_{\overline{m}_1}^{(1)} n^{-\beta_2/(\beta_1+\beta_2+2\beta_1\beta_2)} \leq 2 \quad \text{and} \quad 1 \leq D_{\overline{m}_2}^{(2)} n^{-\beta_1/(\beta_1+\beta_2+2\beta_1\beta_2)} \leq 2.$$

Such models exist for n large enough according to (6.10), since

$$\frac{\beta_2}{\beta_1+\beta_2+2\beta_1\beta_2} < \frac{1}{4} \quad \text{and} \quad \frac{\beta_1}{\beta_1+\beta_2+2\beta_1\beta_2} < \frac{1}{4}. \tag{6.11}$$

Finally, there exists a constant C such that

$$2h_1C_0\left((D^{(1)}_{\overline{m}_1})^{-2\beta_1}+D^{(2)}_{\overline{m}_2})^{-2\beta_2}\right)+\theta\frac{D^{(1)}_{\overline{m}_1}D^{(2)}_{\overline{m}_2}}{n} \leq Cn^{-\overline{\beta}/(\overline{\beta}+1)}.$$

Moreover,

$$\mathbb{E}\left[\|\tilde{F}_{\widehat{m}}-F\|^2\right] \leq \frac{1}{h_0}\mathbb{E}\left[\|\tilde{F}_{\widehat{m}}-F\|^2_{f_{(X,T)}}\right]$$

which proves Corollary 6.3.2. □

Remark 13 *The condition β_1, $\beta_2 > 1$ in Corollary* 6.3.2 *can be generalised to*

$$(\beta_1,\beta_2) \in (\beta_1^*,+\infty)\times(\beta_2^*,+\infty)$$

for a known couple $(\beta_1,^*,\beta_2^*)$ with $\overline{\beta^*} \geq 1$, where $\overline{\beta}^*$ is the harmonic mean of β_1^* and β_2^*, by considering $N_n^{(1)}$ and $N_n^{(2)}$ such that

$$N_n^{(1)} \leq \tfrac{1}{(\log n)^{1/2}}n^{\beta_1^*/(\beta_1^*+\beta_2^*+2\beta_1^*\beta_2^*)} \quad \textit{and} \quad N_n^{(2)} \leq \tfrac{1}{(\log n)^{1/2}}n^{\beta_2^*/(\beta_1^*+\beta_2^*+2\beta_1^*\beta_2^*)}.$$

Then the results of Corollary 6.3.2 *hold, and the proof is similar except that* (6.11) *is replaced by*

$$\begin{array}{l}\frac{\beta_2}{\beta_1+\beta_2+2\beta_1\beta_2} \leq \frac{\beta_2^*}{\beta_1^*+\beta_2^*+2\beta_1^*\beta_2^*} \\ \frac{\beta_1}{\beta_1+\beta_2+2\beta_1\beta_2} \leq \frac{\beta_1^*}{\beta_1^*+\beta_2^*+2\beta_1^*\beta_2^*}.\end{array} \tag{6.12}$$

Thus, if we know a priori that F is more regular in one direction than in the other, we can take into account this information by an appropriate choice of $(N_n^{(1)},N_n^{(2)})$.

6.3.4 Lower bound

We recall the definition of the minimax rate of convergence.

Definition 6.3.1 *Let $\mathcal{F}$ be a set of conditional cumulative distribution functions on A. Let $(r_n)_{n\in\mathbb{N}}$ be a sequence of positive numbers, r_n is the minimax rate of convergence for F over $\mathcal{F}$ if there exist two constants c and C such that*

$$c \leq \inf_{\widehat{F}_n}\sup_{F\in\mathcal{F}}\left(r_n^{-1}\mathbb{E}[\|\widehat{F}_n-F\|^2]\right) \leq C$$

where the infimum is taken over all possible estimators $\widehat{F}_n$.

According to Corollary 6.3.2, provided that $\beta_1,\, \beta_2 > 1$,

$$\inf_{\widehat{F}_n} \sup_{F\in\mathcal{B}^{\beta}_{2,\infty}(A,L)} \left(n^{-\overline{\beta}/(\overline{\beta}+1)}\mathbb{E}[\|\widehat{F}_n - F\|^2]\right) \leq \sup_{F\in\mathcal{B}^{\beta}_{2,\infty}(A,L)} \left(n^{-\overline{\beta}/(\overline{\beta}+1)}\mathbb{E}[\|\widehat{F}_{\widehat{m}} - F\|^2]\right) \leq C.$$

Moreover, the following result holds.

Proposition 6.3.2 *Let $\beta = (\beta_1, \beta_2) \in (0, +\infty) \times (1, +\infty)$. Assume that $h_1 = \|f_{(X,T)}\|_\infty < +\infty$, there exists a constant c which depends on (β, L, h_1) such that*

$$\inf_{\widehat{F}_n} \sup_{F\in\mathcal{B}^{\beta}_{2,\infty}(A,L)} \mathbb{E}\left[n^{-\overline{\beta}/(\overline{\beta}+1)}\|\widehat{F}_n - F\|^2\right] \geq c.$$

Therefore, for every $\beta_1,\, \beta_2 > 1$, the minimax rate of convergence over $\mathcal{B}^{\beta}_{2,\infty}(A, L)$ is $n^{-\overline{\beta}/(\overline{\beta}+1)}$, and $\tilde{F}_{\widehat{m}}$ is minimax over every Besov ball $\mathcal{B}^{\beta}_{2,\infty}(A, L)$. Thus our estimator adapts to the unknown regularity β of the function F.

6.4 Proofs

6.4.1 Proof of Theorem 6.2.1

Let $m = (m_1, m_2) \in I_n$ and $F_m \in S_m$. By definition of $\widehat{m}$ and $\widehat{F}_m$, for every $F_m \in S_m$

$$\gamma_n(\widehat{F}_{\widehat{m}}) + pen(\widehat{m}) \leq \gamma_n(\widehat{F}_m) + pen(m) \leq \gamma_n(F_m) + pen(m). \tag{6.13}$$

Besides, for every $s, t \in S_n$,

$$\begin{aligned}
\gamma_n(t) - \gamma_n(s) &= \frac{1}{n}\sum_{i=1}^{n}\left[(t(X_i,T_i) - \delta_i)^2 - (s(X_i,T_i) - \delta_i)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[t^2(X_i,T_i) - s^2(X_i,T_i) - 2t(X_i,T_i)\delta_i + 2s(X_i,T_i)\delta_i)\right] \\
&= \|t\|_n^2 - \|s\|_n^2 - \frac{2}{n}\sum_{i=1}^{n}\left(t(X_i,T_i) - s(X_i,T_i)\right)\delta_i \\
&= \left(\|t - F\|_n^2 - \|F\|_n^2 + 2\langle t, F\rangle_n\right) - \left(\|s - F\|_n^2 - \|F\|_n^2 + 2\langle s, F\rangle_n\right) \\
&\qquad -\frac{2}{n}\sum_{i=1}^{n}\left(t(X_i,T_i) - s(X_i,T_i)\right)\delta_i \\
&= \|t - F\|_n^2 - \|s - F\|_n^2 - \frac{2}{n}\sum_{i=1}^{n}\left(t(X_i,T_i) - s(X_i,T_i)\right)\left(\delta_i - F(X_i,T_i)\right) \\
&= \|t - F\|_n^2 - \|s - F\|_n^2 - 2\nu_n(t - s)
\end{aligned} \tag{6.14}$$

where

$$\nu_n(t) = \frac{1}{n}\sum_{i=1}^{n} (\delta_i - F(X_i, T_i))\, t(X_i, T_i), \quad \forall t \in L^2(A).$$

Thus (6.13) implies

$$\begin{aligned}
\|\widehat{F}_{\widehat{m}} - F\|_n^2 &\leq \|F_m - F\|_n^2 + pen(m) - pen(\widehat{m}) + 2\nu_n(\widehat{F}_{\widehat{m}} - F_m) \\
&\leq \|F_m - F\|_n^2 + pen(m) - pen(\widehat{m}) \\
&\quad +2\|\widehat{F}_{\widehat{m}} - F_m\|_n \sup_{t \in S_m + S_{\widehat{m}},\ \|t\|_n \leq 1} \nu_n(t).
\end{aligned}$$

This last inequality is the main distinction with the more classical proof developed in similar contexts (see e.g. Chapter 2). In general we consider $\sup_{t \in S_m + S_{\widehat{m}},\ \|t\|_{f_{(X,T)}} \leq 1} \nu_n(t)$ instead of $\sup_{t \in S_m + S_{\widehat{m}},\ \|t\|_n \leq 1} \nu_n(t)$. Technically, this change leads us to consider a version of Talagrand Inequality for non identically distribued variables (Theorem 6.5.1) instead of the i.i.d. version (Theorem 1.2.3 in Introduction). Moreover, this proof requires weaker assumptions and generates smaller constants in the upper bounds than the classical one.

For every function $p(m, m')$ of m and m',

$$\begin{aligned}
\|\widehat{F}_{\widehat{m}} - F\|_n^2 &\leq \|F_m - F\|_n^2 + pen(m) - pen(\widehat{m}) + \frac{1}{4}\|\widehat{F}_{\widehat{m}} - F_m\|_n^2 \\
&\quad +4 \sup_{t \in S_m + S_{\widehat{m}},\ \|t\|_n \leq 1} (\nu_n(t))^2 \\
&= \|F_m - F\|_n^2 + pen(m) - pen(\widehat{m}) + 4p(m, \widehat{m}) + \frac{1}{4}\|\widehat{F}_{\widehat{m}} - F_m\|_n^2 \\
&\quad +4 \sup_{t \in S_m + S_{\widehat{m}},\ \|t\|_n \leq 1} \left((\nu_n(t))^2 - p(m, \widehat{m}) \right).
\end{aligned}$$

Now, consider

$$p(m, m') = \frac{1}{4}(pen(m) + pen(m')) = \frac{\theta}{4}\frac{D_m + D_{m'}}{n}.$$

$$\begin{aligned}
\|\widehat{F}_{\widehat{m}} - F\|_n^2 &\leq \|F_m - F\|_n^2 + 2pen(m) + \frac{1}{4}\left(2\|\widehat{F}_{\widehat{m}} - F\|_n^2 + 2\|F_m - F\|_n^2 \right) \\
&\quad +4 \sup_{t \in S_m + S_{\widehat{m}},\ \|t\|_n \leq 1} \left((\nu_n(t))^2 - p(m, \widehat{m}) \right).
\end{aligned}$$

Finally,

$$\begin{aligned}
\frac{1}{2}\|\widehat{F}_{\widehat{m}} - F\|_n^2 &\leq \frac{3}{2}\|F_m - F\|_n^2 + 2pen(m) \\
&\quad +4 \sum_{m' \in I_n} \sup_{t \in S_m + S_{m'},\ \|t\|_n \leq 1} \left[(\nu_n(t))^2 - p(m, m') \right]_+ . \qquad (6.15)
\end{aligned}$$

Lemma 6.4.1 *There exist numerical constants C_0 and κ_0 which only depend on the constant θ in the penalty such that, for every $m, m' \in I_n$, and every $(x_1, \ldots, x_n) \in A_1^n$ and $(u_1, \ldots, u_n) \in A_2^n$,*

$$\mathbb{E}\left[\left(\sup_{t \in S_m + S_{m'}, (1/n)\sum_{i=1}^n t^2(x_i,u_i) \leq 1} \left(\frac{1}{n}\sum_{i=1}^n (\mathbb{1}_{\{Y_i \leq u_i\}} - F(x_i,u_i))t(x_i,u_i)\right)^2 - p(m,m')\right)_+ \middle| \mathcal{A}\right]$$
$$\leq \frac{C_0}{n}\exp(-\kappa_0\sqrt{D_m + D_{m'}})$$

where $\mathcal{A}$ denotes the following event:

$$\mathcal{A} = \{X_1 = x_1, \ldots, X_n = x_n, T_1 = u_1, \ldots, T_n = u_n\}.$$

Proof of Lemma 6.4.1

The proof relies on Talagrand Inequality (Theorem 6.5.1). Let $(x_1, \ldots, x_n) \in A_1^n$, $(u_1, \ldots, u_n) \in A_2^n$ and m, $m' \in I_n$ be fixed. Let

$$\mu_n(t) = \frac{1}{n}\sum_{i=1}^n (\mathbb{1}_{\{Y_i \leq u_i\}} - F(x_i,u_i))t(x_i,u_i).$$

Then

$$Z = \sup_{t \in S_m + S_{m'}, \|t\|_0 \leq 1} (\mu_n(t))^2 = \sup_{f \in \mathcal{F}_{m,m'}} \left(\frac{1}{n}\sum_{i=1}^n f^{(i)}(\mathbb{1}_{Y_i \leq u_i})\right)^2$$

where $\mathcal{F}_{m,m'}$ is the following set of functions from $\mathbb{R}$ to $\mathbb{R}^n$:

$$\mathcal{F}_{m,m'} = \left\{f = (f^{(1)}, \ldots, f^{(n)}), f^{(i)}(x) = t(x_i,u_i)(x - F(x_i,u_i)),\ t \in S_m + S_{m'} \text{ and } \|t\|_0 \leq 1\right\}.$$

Let $(\varphi_\lambda)_{\lambda=1,\ldots,D_{m+m'}}$ be a $\|.\|_0$-orthogonal basis of $S_m + S_{m'}$ such that $\|\varphi_\lambda\|_0 = 0$ or 1, where $D_{m+m'}$ denotes the dimension of $S_m + S_{m'}$ (see Lemma 6.5.1). Let Γ be the set

$$\Gamma = \{\lambda \in \{1, \ldots, D_{m+m'}\}, \|\varphi_\lambda\|_0 \neq 0\}.$$

Let $t \in S_{m+m'}$, $t = \sum_{\lambda=1}^{D_{m+m'}} a_\lambda \varphi_\lambda$ then

$$\|t\|_0^2 = \sum_{\lambda,\lambda'=1}^{D_{m+m'}} a_\lambda a_{\lambda'} \langle \varphi_\lambda, \varphi_{\lambda'}\rangle_0 = \sum_{\lambda \in \Gamma} a_\lambda^2.$$

We compute the term $\mathbb{H}$ in Theorem 6.5.1.

$$\mathbb{E}[Z^2|\mathcal{A}] = \mathbb{E}\left[\sup_{t \in S_m + S_{m'},\ \|t\|_0 \leq 1} (\mu_n(t))^2 \middle| \mathcal{A}\right] = \mathbb{E}\left[\sup_{\sum_{\lambda \in \Gamma} a_\lambda^2 \leq 1} \left(\sum_{\lambda=1}^{D_{m+m'}} a_\lambda \mu_n(\varphi_\lambda)\right)^2 \middle| \mathcal{A}\right].$$

Besides, for every $\lambda \notin \Gamma$ and for every $i \in \{1, \dots, n\}$, $\varphi_\lambda(x_i, u_i) = 0$. Hence

$$\mu_n(\varphi_\lambda) = \frac{1}{n} \sum_{i=1}^{n} (\mathbb{1}_{\{Y_i \leq u_i\}} - F(x_i, u_i)) \varphi_\lambda(x_i, u_i) = 0, \quad \forall \lambda \notin \Gamma.$$

Therefore,

$$\begin{aligned}
\mathbb{E}[Z^2|\mathcal{A}] &\leq \mathbb{E}\left[\sup_{\sum_{\lambda\in\Gamma} a_\lambda^2 \leq 1} \left(\sum_{\lambda\in\Gamma} a_\lambda \mu_n(\varphi_\lambda) \right)^2 \middle| \mathcal{A} \right] \\
&\leq \mathbb{E}\left[\sup_{\sum_{\lambda\in\Gamma} a_\lambda^2 \leq 1} \left(\sum_{\lambda\in\Gamma} a_\lambda^2 \right) \left(\sum_{\lambda\in\Gamma} \mu_n(\varphi_\lambda)^2 \right) \middle| \mathcal{A} \right] \\
&= \sum_{\lambda\in\Gamma} \mathbb{E}\left[\left(\frac{1}{n} \sum_{i=1}^{n} (\mathbb{1}_{\{Y_i \leq u_i\}} - F(x_i, u_i)) \varphi_\lambda(x_i, u_i) \right)^2 \middle| \mathcal{A} \right] \\
&= \frac{1}{n^2} \sum_{i,l=1}^{n} \sum_{\lambda\in\Gamma} \varphi_\lambda(x_i, u_i) \varphi_\lambda(x_l, u_l) \mathbb{E}\left[\left(\mathbb{1}_{\{Y_i \leq u_i\}} - F(x_i, u_i) \right) \left(\mathbb{1}_{\{Y_l \leq u_l\}} - F(x_l, u_l) \right) \middle| \mathcal{A} \right].
\end{aligned}$$

For every i, $\mathbb{E}[\mathbb{1}_{\{Y_i \leq u_i\}}|\mathcal{A}] = F(x_i, u_i)$ (see Section 6.2.3), and for every $i \neq l$, $\mathbb{1}_{\{Y_i \leq u_i\}}$ and $\mathbb{1}_{\{Y_l \leq u_l\}}$ are independent, thus

$$\begin{aligned}
&\mathbb{E}\left[\left(\mathbb{1}_{\{Y_i \leq u_i\}} - F(x_i, u_i) \right) \left(\mathbb{1}_{\{Y_l \leq u_l\}} - F(x_l, u_l) \right) \middle| \mathcal{A} \right] \\
&= \mathbb{E}\left[\left(\mathbb{1}_{\{Y_i \leq u_i\}} - F(x_i, u_i) \right) | X_i = x_i, T_i = u_i \right] \mathbb{E}\left[\left(\mathbb{1}_{\{Y_l \leq u_l\}} - F(x_l, u_l) \right) | X_l = x_l, T_l = u_l \right] \\
&= 0.
\end{aligned}$$

Thus

$$\mathbb{E}[Z^2|\mathcal{A}] \leq \frac{1}{n} \sum_{\lambda\in\Gamma} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\left(\mathbb{1}_{\{Y_i \leq u_i\}} - F(x_i, u_i) \right)^2 | X_i = x_i, T_i = u_i \right] \varphi_\lambda^2(x_i, u_i) \right).$$

Moreover, for every $i = 1, \dots, n$,

$$\begin{aligned}
&\mathbb{E}\left[\left(\mathbb{1}_{\{Y_i \leq u_i\}} - F(x_i, u_i) \right)^2 | X_i = x_i, T_i = u_i \right] \\
&= (1 - F(x_i, u_i))^2 P[Y_i \leq u_i | X_i = x_i, T_i = u_i] + (F(x_i, u_i))^2 P[Y_i > u_i | X_i = x_i, T_i = u_i] \\
&= (1 - F(x_i, u_i))^2 F(x_i, u_i) + (F(x_i, u_i))^2 (1 - F(x_i, u_i)) \\
&= (1 - F(x_i, u_i)) F(x_i, u_i) \leq \frac{1}{4} \qquad (6.16)
\end{aligned}$$

since $F(x_i, u_i) \in [0, 1]$. Hence

$$\begin{aligned}
\mathbb{E}\left[Z^2|\mathcal{A}\right] &\leq \frac{1}{4n}\sum_{\lambda\in\Gamma}\left(\frac{1}{n}\sum_{i=1}^{n}\varphi_\lambda^2(x_i,u_i)\right)\\
&= \frac{1}{4n}\sum_{\lambda\in\Gamma}\|\varphi_\lambda\|_0^2\\
&= \frac{Card(\Gamma)}{4n}\\
&\leq \frac{D_m + D_{m'}}{4n} = \mathbb{H}^2.
\end{aligned}$$

Now we compute the terms b and v.

$$\begin{aligned}
&\sup_{t\in S_m+S_{m'},\|t\|_0\leq 1}\left(\sup_{i=1,\dots,n}\left\|(\mathbb{1}_{\{.\leq u_i\}} - F(x_i,u_i))t(x_i,u_i)\right\|_\infty\right)\\
&= \sup_{t\in S_m+S_{m'},\|t\|_0\leq 1}\left(\sup_{i=1,\dots,n}|t(x_i,u_i)|\|\mathbb{1}_{\{.\leq u_i\}} - F(x_i,u_i)\|_\infty\right).
\end{aligned}$$

$\mathbb{1}_{\{Y_i\leq u_i\}}$ and $F(x_i, u_i)$ are in $[0, 1]$, so $\|\mathbb{1}_{\{Y_i\leq u_i\}} - F(x_i, u_i)\|_\infty \leq 1$ a.s.. Moreover, let $t \in S_m + S_{m'}$ be such that $\|t\|_0 \leq 1$, for every $i \in \{1, \dots, n\}$

$$t^2(x_i, u_i) \leq \sum_{l=1}^{n} t^2(x_l, u_l) = n\|t\|_0^2 \leq n.$$

Thus

$$\sup_{t\in S_m+S_{m'},\|t\|_0\leq 1}\left(\sup_{i=1,\dots,n}\left\|\left(\mathbb{1}_{\{.\leq u_i\}} - F(x_i,u_i)\right)t(x_i,u_i)\right\|_\infty\right) \leq \sqrt{n} = b.$$

Besides

$$\begin{aligned}
&\sup_{t\in S_m+S_{m'},\|t\|_0\leq 1}\left(\frac{1}{n}\sum_{i=1}^{n} Var\left((\mathbb{1}_{\{Y_i\leq u_i\}} - F(x_i,u_i))t(x_i,u_i)|X_i = x_i, T_i = u_i\right)\right)\\
= &\sup_{t\in S_m+S_{m'},\|t\|_0\leq 1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\mathbb{1}_{\{Y_i\leq u_i\}} - F(x_i,u_i))^2|X_i = x_i, T_i = u_i\right]t^2(x_i,u_i)\right)
\end{aligned}$$

According to (6.16),

$$\begin{aligned}
&\sup_{t\in S_m+S_{m'},\|t\|_0\leq 1}\left(\frac{1}{n}\sum_{i=1}^{n} Var\left((\mathbb{1}_{\{Y_i\leq u_i\}}-F(x_i,u_i))t(x_i,u_i)|X_i=x_i,T_i=u_i\right)\right)\\
\leq\ & \frac{1}{4}\sup_{t\in S_m+S_{m'},\|t\|_0\leq 1}\left(\frac{1}{n}\sum_{i=1}^{n} t^2(x_i,u_i)\right)\\
=\ & \frac{1}{4}=v.
\end{aligned}$$

$p(m,m')=\theta\mathbb{H}^2$ with $\theta>1$, thus by Theorem 6.5.1 there exist numerical constants $\overline{C},\overline{C}',\overline{K},\overline{K}'$ which only depend on θ such that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{t\in S_m+S_{m'},\|t\|_0\leq 1}\left((\mu_n(t))^2-p(m,m')\right)_+|\mathcal{A}\right]\\
&\leq \overline{C}\frac{v}{n}\exp\left(-\overline{\kappa}\frac{n\mathbb{H}^2}{v}\right)+\overline{C}'\frac{b^2}{n^2}\exp\left(-\overline{\kappa}'\frac{n\mathbb{H}}{b}\right)\\
&=\frac{\overline{C}}{4n}\exp\left(-\frac{\overline{\kappa}}{4}(D_m+D_{m'})\right)+\frac{\overline{C}'}{n}\exp\left(-\frac{\overline{\kappa}'}{2}\sqrt{D_m+D_{m'}}\right)\\
&\leq\frac{C_0}{n}\exp(-\kappa_0\sqrt{D_m+D_{m'}}).
\end{aligned}$$

This concludes the proof of Lemma 6.4.1. □

According to Lemma 6.4.1, for every m, $m'\in I_n$,

$$\begin{aligned}
&\mathbb{E}\left[\left(\sup_{t\in S_m+S_{m'},(1/n)\sum_{i=1}^n t^2(X_i,Y_i)\leq 1}(\nu_n(t))^2-p(m,m')\right)_+\Bigg|\{(X_i,T_i)\}_{i=1,\dots,n}\right]\\
&\leq\frac{C_0}{n}\exp(-\kappa_0\sqrt{D_m+D_{m'}})
\end{aligned}$$

almost surely. Now, by Assumption (**H**), there exists a numerical constant B which depends on θ such that

$$\sum_{m'\in I_n}\mathbb{E}\left[\left(\sup_{t\in S_m+S_{m'},(1/n)\sum_{i=1}^n t^2(X_i,Y_i)\leq 1}(\nu_n(t))^2-p(m,m')\right)_+\Bigg|\{(X_i,T_i)\}_{i=1,\dots,n}\right]\leq\frac{C_0B}{n}$$

almost surely. Then by (6.15),

$$\mathbb{E}\left[\|\widehat{F}_{\widehat{m}}-F\|_n^2\,|\{(X_i,T_i)\}_{i=1,\dots,n}\right]\leq 2\|F-F_m\|_n^2+4pen(m)+\frac{4C_0B}{n}$$

which concludes the proof of Theorem 6.2.1. □

6.4.2 Proof of Corollary 6.3.1

The proof of is divided in two Propositions. Let

$$\Omega = \left\{ \left| \frac{\|t\|_n^2}{\|t\|_{f_{(X,T)}}^2} - 1 \right| \leq \frac{1}{2}, \forall t \in S_n \right\}.$$

Proposition 6.4.1

$$\mathbb{E}\left[\|\tilde{F}_{\widehat{m}} - F\|_{f_{(X,T)}}^2 \mathbb{1}_\Omega \right] \leq C_1' \inf_{m \in I_n} \left\{ \inf_{t \in S_m} \|F - t\|_{f_{(X,T)}}^2 + pen(m) \right\} + \frac{C_2'}{n}$$

where C_1' and C_2' are numerical constants.

Proposition 6.4.2 *Under the assumptions of Corollary* 6.3.1*,*

$$\mathbb{E}\left[\|\tilde{F}_{\widehat{m}} - F\|_{f_{(X,T)}}^2 \mathbb{1}_{\Omega^c} \right] \leq \frac{C_6}{n}$$

where C_6 depends on h_0 and K.

Proof of Proposition 6.4.1

First of all, the matrix $\widehat{G}_m$ is invertible on the set Ω. Indeed, let μ be an eigenvalue of $\widehat{G}_m$ and $U \in \mathbb{R}^{D_m}$ an eigenvector with norm 1, then

$$\mu = U^t \widehat{G}_m U = \left\| \sum_{\lambda \in J_m} u_\lambda \xi_\lambda \right\|_n^2$$

where $(\xi_\lambda)_{\lambda \in J_m}$ is a $\|.\|$-orthogonal basis of S_m. Hence, on Ω

$$\begin{aligned} \mu &\geq \frac{1}{2} \left\| \sum_{\lambda \in J_m} u_\lambda \xi_\lambda \right\|_{f_{(X,T)}}^2 \\ &\geq \frac{h_0}{2} \left\| \sum_{\lambda \in J_m} u_\lambda \xi_\lambda \right\|^2 \\ &= \frac{h_0}{2} \sum_{\lambda \in J_m} u_\lambda^2 = \frac{h_0}{2}. \end{aligned}$$

Moreover, let

$$F_n = \arg\min_{t \in S_n} \|F - t\|_{f_{(X,T)}}^2$$

be the projection of F on the global model S_n. Then $(\widehat{F}_{\widehat{m}} - F_n) \in S_n$ so

$$\|\widehat{F}_{\widehat{m}} - F\|^2_{f_{(X,T)}} = \|\widehat{F}_{\widehat{m}} - F_n\|^2_{f_{(X,T)}} + \|F_n - F\|^2_{f_{(X,T)}}.$$

By definition of Ω,

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{F}_{\widehat{m}} - F\|^2_{f_{(X,T)}} \mathbb{1}_\Omega\right] &\leq 2\mathbb{E}\left[\|\widehat{F}_{\widehat{m}} - F_n\|^2_n \mathbb{1}_\Omega\right] + \|F_n - F\|^2_{f_{(X,T)}} \\
&\leq 4\mathbb{E}\left[\|\widehat{F}_{\widehat{m}} - F\|^2_n \mathbb{1}_\Omega\right] + 4\mathbb{E}\left[\|F_n - F\|^2_n \mathbb{1}_\Omega\right] + \|F_n - F\|^2_{f_{(X,T)}} \\
&= 4\mathbb{E}\left[\|\widehat{F}_{\widehat{m}} - F\|^2_n \mathbb{1}_\Omega\right] + 7\|F_n - F\|^2_{f_{(X,T)}}.
\end{aligned}$$

Moreover, according to Comment 2. after Theorem 6.2.1,

$$\mathbb{E}\left[\|\widehat{F}_{\widehat{m}} - F\|^2_n\right] \leq C_1 \inf_{m \in I_n} \left\{ \inf_{t \in S_m} \|F - t\|^2_{f_{(X,T)}} + pen(m) \right\} + \frac{C_2}{n}.$$

For every $m \in J_n$, $S_m \subset S_n$ so

$$\inf_{t \in S_n} \|F - t\|^2_{f_{(X,T)}} \leq \inf_{t \in S_m} \|F - t\|^2_{f_{(X,T)}}$$

thus

$$\mathbb{E}\left[\|\widehat{F}_{\widehat{m}} - F\|^2_{f_{(X,T)}} \mathbb{1}_\Omega\right] \leq \inf_{m \in J_n} \left\{ (4C_1 + 7) \inf_{t \in S_m} \|F - t\|^2_{f_{(X,T)}} + 4C_1 pen(m) \right\} + \frac{4C_2}{n}.$$

Besides, according to Remark 11,

$$\mathbb{E}\left[\|\tilde{F}_{\widehat{m}} - F\|^2_{f_{(X,T)}} \mathbb{1}_\Omega\right] \leq \mathbb{E}\left[\|\widehat{F}_{\widehat{m}} - F\|^2_{f_{(X,T)}} \mathbb{1}_\Omega\right]$$

which ends the proof of Proposition 6.4.1. □

Proof of Proposition 6.4.2

The proof is based on the following Lemma.

Lemma 6.4.2 *Under the assumptions of Theorem* 6.2.1*,*

$$P[\Omega^c] \leq 2(N_n)^2 \exp\left(-\frac{3 - 2\sqrt{2}}{2} \frac{nh_0}{(N_n)^2 K^2} \right). \tag{6.17}$$

Indeed, assume that Lemma 6.4.2 holds. $\tilde{F}_{\widehat{m}}(x,u)$ and $F(x,u) \in [0,1]$ for every $(x,u) \in A$, so

$$\|\tilde{F}_{\widehat{m}} - F\|^2_{f_{(X,T)}} \leq 1$$

Hence, let $c_0 = h_0(3-2\sqrt{2})/(2K^2)$.

$$\begin{aligned}
\mathbb{E}\left[\|\tilde{F}_{\widehat{m}} - F\|^2_n \mathbb{1}_{\Omega^c}\right] &\leq 2(N_n)^2 \exp\left(-c_0 \frac{n}{(N_n)^2}\right) \\
&\leq 2n \exp(-c_0 \log^2 n) \\
&\leq \frac{1}{n} 2n^{-(c_0 \log n - 2)}
\end{aligned}$$

and $2n^{-(c_0 \log n - 2)}$ is upper bounded by a constant C_6 which depends on K and h_0. This concludes the proof of Proposition 6.4.2. □

Proof of Lemma 6.4.2

Let $\{\chi_{(k,l)}, (k,l) \in J_n\}$ be an $\|.\|_{f_{(X,T)}}$-orthonormal basis of the global space S_n where J_n is the set of index defined in (6.3) for $D^{(1)}_{m_1} = N^{(1)}_n$ and $D^{(2)}_{m_2} = N^{(2)}_n$. Assumption $(\mathbf{A_3})$ for the model $S_m = S_n$ implies that

$$\begin{aligned}
& \sup_{(x,u)\in A} \left(t(x,u)\right)^2 \leq K N_n \|t\|^2, \quad \forall t \in S_n \\
\Rightarrow\ & \sup_{(x,u)\in A} \left(t(x,u)\right)^2 \leq \tfrac{K}{h_0} N_n \|t\|^2_{f_{(X,T)}}, \quad \forall t \in S_n \\
\Leftrightarrow\ & \sup_{(x,u)\in A} \sum_{(k,l)\in J_n} \left(\chi_{(k,l)}(x,u)\right)^2 \leq \tfrac{K}{h_0} N_n
\end{aligned} \tag{6.18}$$

and the latest equivalence comes from Proposition 1.2.1 and the consecutive remark.

$$\begin{aligned}
P[\Omega^c] &= P\left[\exists t \in S_n, \left|\|t\|^2_n - \|t\|^2_{f_{(X,T)}}\right| > \frac{1}{2}\|t\|^2_{f_{(X,T)}}\right] \\
&= P\left[\sup_{t\in S_n, \|t\|^2_{f_{(X,T)}}=1} \left|\|t\|^2_n - \|t\|^2_{f_{(X,T)}}\right| > \frac{1}{2}\right].
\end{aligned}$$

As $\left(\chi_{(k,l)}\right)_{(k,l)\in J_n}$ is a $\|.\|_{f_{(X,T)}}$-orthonormal basis of S_n,

$$\begin{aligned}
P[\Omega^c] = P\Bigg[& \sup_{\sum_{(k,l)\in J_n} a^2_{k,l}=1} \Bigg| \sum_{(k,l)\in J_n,(k',l')\in J_n} a_{k,l} a_{k',l'} \\
& \left(\frac{1}{n}\sum_{i=1}^n \left(\chi_{(k,l)}(X_i,T_i)\chi_{(k,l)}(X_i,T_i) - \mathbb{1}_{\{(k,l)=(k',l')\}}\right)\right)\Bigg| > \frac{1}{2}\Bigg].
\end{aligned}$$

Let

$$\begin{aligned} S_{k,l,k',l'} &= \frac{1}{n}\sum_{i=1}^{n}\left(\chi_{(k,l)}(X_i,T_i)\chi_{(k,l)}(X_i,T_i) - 1\!\!1_{\{(k,l)=(k',l')\}}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\chi_{(k,l)}(X_i,T_i)\chi_{(k,l)}(X_i,T_i) - \mathbb{E}\left[\chi_{(k,l)}(X_i,T_i)\chi_{(k',l')}(X_i,T_i)\right]\right). \end{aligned}$$

Then

$$P[\Omega^c] \leq P\left[\sup_{\sum_{(k,l)\in J_n} a_{k,l}^2=1} \sum_{(k,l)\in J_n,(k',l')\in J_n} |a_{k,l}||a_{k',l'}||S_{k,l,k',l'}| > \frac{1}{2}\right].$$

Let C and V be the following $N_n \times N_n$-square matrix.

$$V = \left(\sqrt{v_{k,l,k',l'}}\right)_{(k,l)\in J_n,(k,l')\in J_n} \quad \text{where } v_{k,l,k',l'} = \mathbb{E}\left[\chi^2_{(k,l)}(X_i,T_i)\chi^2_{(k',l')}(X_i,T_i)\right],$$

$$C = \left(c_{k,l,k',l'}\right)_{(k,l)\in J_n,(k,l')\in J_n} \quad \text{where } c_{k,l,k',l'} = \left\|\chi_{(k,l)}(X_i,T_i)\chi_{(k',l')}(X_i,T_i)\right\|_\infty$$

and

$$\begin{aligned} \rho(V) &= \sup_{\sum_{(k,l)\in J_n} a_{k,l}^2=1} \sum_{(k,l)\in J_n,(k',l')\in J_n} |a_{k,l}||a_{k',l'}|\sqrt{v_{k,l,k',l'}}, \\ \rho(C) &= \sup_{\sum_{(k,l)\in J_n} a_{k,l}^2=1} \sum_{(k,l)\in J_n,(k',l')\in J_n} |a_{k,l}||a_{k',l'}|c_{k,l,k',l'}. \end{aligned}$$

Let

$$x = \frac{3-2\sqrt{2}}{2}\min\left(\frac{1}{\rho^2(V)}, \frac{1}{\rho(C)}\right), \tag{6.19}$$

then

$$\left.\begin{aligned} \sqrt{2x}\rho(V) \leq \sqrt{3-2\sqrt{2}} = \sqrt{2}-1 \\ x\rho(C) \leq \tfrac{3-2\sqrt{2}}{2} \end{aligned}\right\} \Rightarrow \quad \sqrt{2x}\rho(V) + x\rho(C) \leq \tfrac{1}{2}. \tag{6.20}$$

Thus

$$P[\Omega^c] \leq P\left[\sup_{\sum_{(k,l)\in J_n} a_{k,l}^2=1} \left(\sum_{(k,l)\in J_n,(k',l')\in J_n} |a_{k,l}||a_{k',l'}||S_{k,l,k',l'}| > \sqrt{2x}\rho(V) + x\rho(C)\right)\right].$$

Besides, for every $(a_{k,l})_{(k,l)\in J_n}$ such that $\sum_{(k,l)\in J_n} a_{k,l}^2 = 1$,

$$\sum_{(k,l)\in J_n,(k',l')\in J_n} |a_{k,l}||a_{k',l'}||S_{k,l,k',l'}| > \sqrt{2x}\rho(V) + x\rho(C)$$

$$\Rightarrow \sum_{(k,l)\in J_n,(k',l')\in J_n} |a_{k,l}||a_{k',l'}||S_{k,l,k',l'}| > \sup_{\sum_{(k,l)\in J_n} b_{k,l}^2=1} \left(\sum_{(k,l)\in J_n,(k',l')\in J_n} |b_{k,l}||b_{k',l'}|(\sqrt{2xv_{k,l,k',l'}} + xc_{k,l,k',l'}) \right)$$

$$\Rightarrow \sum_{(k,l)\in J_n,(k',l')\in J_n} |a_{k,l}||a_{k',l'}||S_{k,l,k',l'}| > \sum_{(k,l)\in J_n,(k',l')\in J_n} |a_{k,l}||a_{k',l'}|(\sqrt{2xv_{k,l,k',l'}} + xc_{k,l,k',l'})$$

$$\Rightarrow \quad |S_{k,l,k',l'}| > \sqrt{2xv_{k,l,k',l'}} + xc_{k,l,k',l'}, \quad \forall (k,l) \in J_n, (k',l') \in J_n.$$

Hence,

$$P[\Omega^c] \leq \sum_{(k,l)\in J_n,(k',l')\in J_n} P\left[|S_{k,l,k',l'}| > \sqrt{2xv_{k,l,k',l'}} + xc_{k,l,k',l'}\right].$$

According to Bernstein Inequality (Theorem 1.2.4 in Introduction), by definition of $v_{k,l,k',l'}$ and $c_{k,l,k',l'}$,

$$P[\Omega^c] \leq \sum_{(k,l)\in J_n,(k',l')\in J_n} 2\exp(-nx) = 2(N_n)^2\exp(-nx). \tag{6.21}$$

Now, we upper bound $\rho^2(V)$ and $\rho(C)$. With Cauchy Schwartz Inequality,

$$\begin{aligned}
\rho^2(V) &= \sup_{\sum_{(k,l)\in J_n} a_{k,l}^2=1} \left(\sum_{(k,l)\in J_n,(k',l')\in J_n} |a_{k,l}||a_{k',l'}|\sqrt{v_{k,l,k',l'}} \right)^2 \\
&\leq \sup_{\sum_{(k,l)\in J_n} a_{k,l}^2=1} \left(\sum_{(k,l)\in J_n} a_{k,l}^2 \right) \left(\sum_{(k,l)\in J_n} \left(\sum_{(k',l')\in J_n} |a_{k',l'}|\sqrt{v_{k,l,k',l'}} \right)^2 \right) \\
&= \sup_{\sum_{(k,l)\in J_n} a_{k,l}^2=1} \sum_{(k,l)\in J_n} \left(\sum_{(k',l')\in J_n} |a_{k',l'}|\sqrt{v_{k,l,k',l'}} \right)^2 \\
&\leq \sup_{\sum_{(k,l)\in J_n} a_{k,l}^2=1} \sum_{(k,l)\in J_n} \left(\left(\sum_{(k',l')\in J_n} a_{k',l'}^2 \right) \left(\sum_{(k',l')\in J_n} v_{k,l,k',l'} \right) \right) \\
&= \sum_{(k,l)\in J_n,(k',l')\in J_n} v_{k,l,k',l'}.
\end{aligned}$$

By definition of $v_{k,l,k',l'}$,

$$\begin{aligned}\rho^2(V) &\leq \mathbb{E}\left[\sum_{(k,l)\in J_n}\chi^2_{(k,l)}(X_i,T_i)\left(\sum_{(k',l')\in J_n}\chi^2_{(k',l')}(X_i,T_i)\right)\right]\\ &\leq \sup_{(x,u)\in A}\left(\sum_{(k',l')\in J_n}\chi_{(k',l')}(x,u)\right)\times\sum_{(k,l)\in J_n}\mathbb{E}\left[\chi^2_{(k,l)}(X_i,T_i)\right].\end{aligned}$$

For every $(k,l)\in J_n$, $\mathbb{E}\left[\chi^2_{(k,l)}(X_i,T_i)\right]=1$. Hence with (6.18)

$$\rho^2(V)\leq\frac{K(N_n)^2}{h_0}.$$

Similarly,

$$\begin{aligned}\rho(C) &\leq \sqrt{\sum_{(k,l)\in J_n,(k',l')\in J_n}c^2_{k,l,k',l'}}\\ &= \sqrt{\sum_{(k,l)\in J_n,(k',l')\in J_n}\left\|\chi_{(k,l)}(X_i,T_i)\chi_{(k',l')}(X_i,T_i)\right\|^2_\infty}\\ &\leq \sqrt{\sum_{(k,l)\in J_n,(k',l')\in J_n}\left\|\chi_{(k,l)}(X_i,T_i)\right\|^2_\infty\left\|\chi^2_{(k',l')}(X_i,T_i)\right\|^2_\infty}.\end{aligned}$$

With (6.18),

$$\left\|\chi_{(k,l)}(X_i,T_i))\right\|^2_\infty\leq\sup_{(x,u)\in A}\left|\chi_{(k,l)}(x,u)\right|^2\leq\frac{KN_n}{h_0}\left\|\chi_{(k,l)}\right\|^2_{f_{(X,T)}}=\frac{KN_n}{h_0}$$

and

$$\left\|\chi_{(k',l')}(X_i,T_i)\right\|^2_\infty\leq\frac{KN_n}{h_0}.$$

Hence

$$\rho(C)\leq\sqrt{\sum_{(k,l)\in J_n,(k',l')\in J_n}\left(\frac{KN_n}{h_0}\right)^2}=\sqrt{(N_n)^2\left(\frac{KN_n}{h_0}\right)^2}=\frac{K(N_n)^2}{h_0}.$$

Plugging the upper bounds of $\rho^2(V)$ and $\rho(C)$ in (6.19) implies

$$x\geq\frac{h_0}{K(N_n)^2}\frac{3-2\sqrt{2}}{2}$$

and (6.21) ends the proof of Lemma 6.4.2. □

6.4.3 Proof of Proposition 6.3.2

The proof is based on the following Theorem from Tsybakov (2004) (Chapter 2, Theorem 2.5). Let $\mathcal{B} = \mathcal{B}^{\beta}_{2,\infty}(A, L)$.

Theorem 6.4.1 *Assume that there exist $M \geq 2$ and $F_0, \ldots, F_M$ such that*

1. *$F_j \in \mathcal{B}$ for every $j \in \{0, \ldots, M\}$.*
2. *$\|F_j - F_l\|^2 \geq 2r$ for every $j \neq l \in \{0, \ldots, M\}$.*
3. *$P_j^{(n)} << P_0^{(n)}$ for every $j \in \{0, \ldots, M\}$, where $P_j^{(n)}$ denotes the distribution of $(X_i, T_i, \delta_i)_{i=1,\ldots,n}$ if $F = F_j$, and*

$$\frac{1}{M}\sum_{j=1}^{M} K(P_j^{(n)}, P_0^{(n)}) \leq \alpha \log M$$

with $0 < \alpha < 1/8$.

Then there exists a constant c such that

$$\inf_{\widehat{F}_n} \sup_{F \in \mathcal{B}} \mathbb{E}\left[r\|\widehat{F}_n - F\|^2\right] \geq c.$$

Up to rescalings and translations, we assume that $A = [0,1] \times [0,1]$. We construct a set of distribution functions $\{F_0, \ldots, F_M\}$ which satisfies (1), (2) and (3).

Construction of the (F_i)'s

Let

$$F_0(x,u) = \mathbb{1}_{[0,1]}(x)\left(a\mathbb{1}_{[0,+\infty[}(u) + au\mathbb{1}_{[0,1]}(u) + (1-a)\mathbb{1}_{[1,+\infty[}(u)\right)$$

with $a = \min(1/3, L/2)$. F_0 is a conditional distribution since for every $x \in [0,1]$,

- $F_0(x,u) = 0, \quad \forall u < 0.$
- $F_0(x,u) = 1, \quad \forall u \geq 1.$
- $F_0(x,.)$ is increasing on $[0,1]$.

Let ψ be a one-dimensional compactly supported wavelet. Up to a rescaling, the support of ψ is assumed to be $[0,1]$. For every $J = (j_1, j_2) \in \mathbb{N}^2$, $S = (s_1, s_2) \in \mathbb{Z}^2$, let

$$\psi_{J,S}(x,u) = 2^{(j_1+j_2)/2}\psi(2^{j_1}x - s_1)\psi(2^{j_2}u - s_2)$$

for a fixed J which will be determined further. For every $J \in \mathbb{N}$, there exists a subset R_J of $\mathbb{Z}$ such that

* $Supp(\psi_{J,S}) = I_{J,S} \subset]0,1[^2$ for every $S \in R_J$.
* The applications $\{\psi_{J,S}, S \in R_J\}$ have disjoint supports.
* $|R_J| = 2^{j_1+j_2}$.

Let b be a positive constant which will be determined later. For every $\epsilon \in \{0,1\}^{|R_J|}$, let

$$G_\epsilon = \sqrt{\frac{b}{n}} \sum_{S \in R_J} \epsilon_S \psi_{J,S}$$

and $F_\epsilon = F_0 + G_\epsilon$. For every $x \in [0,1]$,

- $F_\epsilon(x,u) = F_0(x,u) = 0, \quad \forall u < 0$
- $F_\epsilon(x,u) = F_0(x,u) = 1 \quad \forall u \geq 1$

Moreover, for every $(x,u) \in [0,1]^2$,

$$F_\epsilon(x,u) = a + \int_0^u \left(a + \sqrt{\frac{b}{n}} \sum_{S \in R_J} \epsilon_S \frac{\partial \psi_{J,S}}{\partial y}(x,y) \right) dy.$$

Assume that

$$\sqrt{\frac{b}{n}} 2^{j_1/2} 2^{3j_2/2} \|\psi\|_\infty \left\| \frac{\partial \psi(x,.)}{\partial y} \right\|_\infty \leq \frac{a}{2} \tag{6.22}$$

then for every $x \in [0,1]$ the application $F_\epsilon(x,.)$ is increasing on $[0,1]$, so F_ϵ is a conditional distribution function on $[0,1]^2$.

Condition which guarantees that $F_\epsilon \in \mathcal{B}$ for every ϵ

On the one hand, assume that ψ is regular enough, then according to Hochmuth (2002) (Theorem 3.5),

$$|G_\epsilon|_{\mathcal{B}^\beta_{2,\infty}([0,1]^2)} \leq (2^{j_1\beta_1} + 2^{j_2\beta_2}) \|G_\epsilon\|.$$

Moreover

$$\|G_\epsilon\| = \sqrt{\sum_{S \in R_J} \epsilon_S^2 \frac{b}{n}} \leq \sqrt{\frac{b}{n}} 2^{(j_1+j_2)/2}.$$

Thus

$$\|G_\epsilon\|_{\mathcal{B}^\beta_{2,\infty}([0,1]^2)} = |G_\epsilon|_{\mathcal{B}^\beta_{2,\infty}([0,1]^2)} + \|G_\epsilon\| \leq \sqrt{\frac{b}{n}} 2^{(j_1+j_2)/2}(2^{j_1\beta_1} + 2^{j_2\beta_2} + 1).$$

On the other hand, $|F_0|_{\mathcal{B}^\beta_{2,\infty}([0,1]^2)} = 0$. Indeed, let $r_i = \lfloor \beta_i \rfloor + 1$ for $i = 1$ and 2. Then $r_1 \geq 1$, $r_2 \geq 2$ and

$$|F_0|_{\mathcal{B}^\beta_{2,\infty}([0,1]^2)} = \sup_{t>0} \left[t^{-\beta_1} \omega_{r_1,1}(F_0, t, [0,1]^2)_2 + t^{-\beta_2} \omega_{r_2,2}(F_0, t, [0,1]^2)_2 \right].$$

Besides, for every $(x,u) \in [0,1]^2$ and $h > 0$ such that $(x+h,u) \in [0,1]^2$, $F_0(x+h,u) = F_0(x,u)$. So, as $r_1 \geq 1$,

$$\Delta^{r_1}_{h,1} F_0(x,u) = 0.$$

Hence

$$\omega_{r_1,1}(F_0, t, [0,1]^2)_2 = \sup_{|h|\leq t} \|\Delta^{r_1}_{h,1} F_0\|_2 = 0.$$

Moreover on $[0,1]^2$, $F_0(x,u) = a(1+u)$ if $u < 1$ and $F_0(x,1) = 1$. Thus, let $\tilde{F}_0(x,u) = a(1+u)$ for every $(x,u) \in [0,1]^2$, F_0 and $\tilde{F}_0$ are equal on $[0,1]^2$ except on a set of measure 0, so $\|\Delta^{r_2}_{h,2} F_0\| = \|\Delta^{r_2}_{h,2} \tilde{F}_0\|$. Besides, for all $(x,u) \in [0,1]^2$

$$\Delta^1_{h,2} \tilde{F}_0(x,u) = ah \quad \Rightarrow \quad \Delta^{r_2-1}_{h,2} \Delta^1_{h,2} \tilde{F}_0(x,u) = 0$$

as $r_2 - 1 \geq 1$. Then $\omega_{r_2,2}(F_0, t, [0,1]^2)_2 = 0$. Therefore $|F_0|_{\mathcal{B}^\beta_{2,\infty}([0,1]^2)} = 0$,

$$\|F_0\|_{\mathcal{B}^\beta_{2,\infty}([0,1]^2)} = \sqrt{\int_0^1 \int_0^1 a^2(1+u)^2 du dx} = \sqrt{\frac{7}{3}} a$$

and

$$\|F_\epsilon\|_{\mathcal{B}^\beta_{2,\infty}([0,1]^2)} \leq \sqrt{\frac{7}{3}} a + \sqrt{\frac{b}{n}} 2^{(j_1+j_2)/2}(2^{j_1\beta_1} + 2^{j_2\beta_2} + 1).$$

By definition, $a \leq L/2$ so $\|F_\epsilon\|_{\mathcal{B}^\beta_{2,\infty}([0,1]^2)} \leq L$ as soon as

$$\sqrt{\frac{b}{n}} 2^{(j_1+j_2)/2}(2^{j_1\beta_1} + 2^{j_2\beta_2} + 1) \leq L \left(1 - \sqrt{\frac{7}{12}}\right). \tag{6.23}$$

Expression of $\|F_\epsilon - F_{\epsilon'}\|^2$

$$\begin{aligned} \|F_\epsilon - F_{\epsilon'}\|^2 &= \frac{b}{n} \sum_{S \in R_J} \int_{I_{J,S}} (\epsilon_S - \epsilon'_S)^2 \psi^2_{J,S}(x,u) dx du \\ &= \frac{b}{n} \sum_{S \in R_J} 1\!\!1_{\{\epsilon_S \neq \epsilon'_S\}} = \frac{b}{n} \rho(\epsilon, \epsilon') \end{aligned} \tag{6.24}$$

Upper bound of $K(P_\epsilon^{(n)}, P_0^{(n)})$

For every $i \in \{1, \ldots, n\}$, the distribution of (X_i, T_i, δ_i) under F_ϵ is

$$p_\epsilon(x, u, d) = \left[(F_\epsilon(x, u))^d (1 - F_\epsilon(x, u))^{1-d}\right] f_{(X,T)}(x, u)$$

with respect to $\mathcal{L} \otimes \mathcal{L} \otimes \mu$ where $\mathcal{L}$ is the Lebesgue measure and μ is the counting measure on $\mathbb{N}$. Similarly, under F_0, (X_i, T_i, δ_i) has a distribution

$$p_0(x, u, d) = \left[(F_0(x, u))^d (1 - F_0(x, u))^{1-d}\right] f_{(X,T)}(x, u)$$

with respect to $\mathcal{L} \otimes \mathcal{L} \otimes \mu$. For every $\epsilon \in \{0, 1\}^{|R_J|}$, P_ϵ is absolutely continuous with respect to P_0. Indeed,

$$F_0(x, u) = 0 \quad \Rightarrow \quad (x, u) \notin [0, 1] \times [0, +\infty[\quad \Rightarrow \quad F_\epsilon(x, u) = 0,$$

$$F_0(x, u) = 1 \quad \Rightarrow \quad (x, u) \in [0, 1] \times [1, +\infty[\quad \Rightarrow \quad F_\epsilon(x, u) = 1,$$

thus, $p_0(x, u, d) = 0 \Rightarrow p_\epsilon(x, u, d) = 0$.

$$K(P_\epsilon, P_0) = \int_{\mathbb{R}^2} \left[\log\left(\frac{F_\epsilon(x, u)}{F_0(x, u)}\right) F_\epsilon(x, u) + \log\left(\frac{1 - F_\epsilon(x, u)}{1 - F_0(x, u)}\right)(1 - F_\epsilon(x, u))\right] f_{(X,T)}(x, u) dx du$$

Out of the intervals $\{I_{J,S}, S \in R_J\}$, F_ϵ and F_0 are equal. Hence

$$= \sum_{S \in R_J} \int_{I_{J,S}} \left[\log\left(1 + \frac{\theta_S}{a(1 + u)}\right)(a(1 + u) + \theta_S) + \log\left(1 - \frac{\theta_S}{1 - a(1 + u)}\right)(1 - a(1 + u) - \theta_S)\right] f_{(X,T)}(x, u) dx du$$

where

$$\theta_S = \epsilon_S \sqrt{\frac{b}{n}} \psi_{J,S}(x, u).$$

Besides for every $v > -1$, there exists w such that

$$\log(1 + v) = v - \frac{v^2}{2} \frac{1}{(1 + w)^2} \leq v.$$

By construction of F_ϵ, for every $S \in R_J$ and every $(x, u) \in I_{J,S}$,

$$\frac{F_\epsilon(x, u)}{F_0(x, u)} > 0 \quad \Rightarrow \quad \frac{\theta_S}{a(1 + u)} > -1$$

$$\frac{1 - F_\epsilon(x, u)}{1 - F_0(x, u)} > 0 \quad \Rightarrow \quad -\frac{\theta_S}{1 - a(1 + u)} > -1.$$

Therefore

$$K(P_\epsilon, P_0) \leq \sum_{S \in R_J} \int_{I_{J,S}} \left[\theta_S + \frac{\theta_S^2}{a(1+u)} - \theta_S + \frac{\theta_S^2}{1-a(1+u)} \right] f_{(X,T)}(x,u) dx du.$$

For every $u \in [0,1]$,

$$\frac{1}{a(1+u)} \leq \frac{1}{a} \quad \text{and} \quad \frac{1}{1-a(1+u)} \leq \frac{1}{1-2a}.$$

and by definition of a, $a > 0$ and $1 - 2a \geq 1/3 > 0$. Thus,

$$\begin{aligned} K(P_\epsilon, P_0) &\leq \left(\frac{1}{a} + \frac{1}{1-2a} \right) \sum_{S \in R_J} \int_{I_{J,S}} \theta_S^2 f_{(X,T)}(x,y) dx dy \\ &\leq \left(\frac{1}{a} + \frac{1}{1-2a} \right) \frac{b}{n} \|f_{(X,T)}\|_\infty |R_J| \\ &= a' \|f_{(X,T)}\|_\infty \frac{b 2^{j_1+j_2}}{n}. \end{aligned}$$

where $a' = 1/a + 1/(1-2a)$. Finally,

$$K(P_\epsilon^{(n)}, P_0^{(n)}) \leq a' \|f_{(X,T)}\|_\infty b 2^{j_1+j_2}.$$

Conclusion

Acoording to Lemma 2.7, Chapter 2 in Tsybakov (2004), there exists a family $(\epsilon^{(0)}, \ldots, \epsilon^{(M)}) \subset \{0,1\}^{|R_J|}$ with $\epsilon^{(0)} = (0, \ldots, 0)$ such that

$$\rho(\epsilon^{(i)}, \epsilon^{(i')}) \geq \frac{|R_J|}{8} = \frac{2^{j_1+j_2}}{8}, \quad \forall i \neq i' \in \{0, \ldots, M\}$$

and

$$\log(M) \geq \frac{\log 2}{8} 2^{j_1+j_2},$$

where the distance ρ is defined in (6.24).

Now parameters B_0, b, j_1 and j_2 are choosen so that the family $(F_{\epsilon^{(0)}}, \ldots, F_{\epsilon^{(M)}})$ satisfies the assumptions of Theorem 6.4.1 with

$$r = B_0 n^{\overline{\beta}/(\overline{\beta}+1)}.$$

Let

$$b = \frac{\log 2}{72 \|f_{(X,T)}\|_\infty a'}, \qquad c_0 = \left[\frac{L}{4\sqrt{b}} \left(1 - \sqrt{\frac{7}{2}} \right) \right]^{1/(1+\beta_1+\beta_2)}$$

and j_1 and j_2 be in $\mathbb{N}^*$ such that

$$\frac{c_0}{2} n^{\beta_2/(\beta_1+\beta_2+2\beta_1\beta_2)} \leq 2^{j_1} \leq c_0 n^{\beta_2/(\beta_1+\beta_2+2\beta_1\beta_2)}$$
$$\frac{c_0}{2} n^{\beta_1/(\beta_1+\beta_2+2\beta_1\beta_2)} \leq 2^{j_2} \leq c_0 n^{\beta_1/(\beta_1+\beta_2+2\beta_1\beta_2)}.$$

Let $B_0 = 32/bc_0^2$. Then for every $i, i' \in \{0, \ldots, M\}$

$$\begin{aligned} \|F_{\epsilon^{(i)}} - F_{\epsilon^{(i')}}\|^2 &\geq \frac{b}{n} \frac{2^{j_1+j_2}}{8} \\ &\geq \frac{bc_0^2}{32n} n^{(\beta_1+\beta_2)/(\beta_1+\beta_2+2\beta_1\beta_2)} \\ &= B_0 n^{-2\beta_1\beta_2/(\beta_1+\beta_2+2\beta_1\beta_2)} \\ &= B_0 n^{-\overline{\beta}/(\overline{\beta}+1)} \end{aligned}$$

which proves (2) in Theorem 6.4.1 with $r = B_0 n^{-\overline{\beta}/(\overline{\beta}+1)}$. Moreover

$$\frac{1}{M} \sum_{l=0}^{M} K(P_{\epsilon^{(l)}}^{(n)}, P_0^{(n)}) \leq a' \|f_{(X,T)}\|_\infty b 2^{j_1+j_2} = \frac{\log 2}{72} 2^{j_1+j_2} \leq \frac{\log M}{9}$$

which proves (3) in Theorem 6.4.1 with $\alpha = 1/9$.

Finally (1) in Theorem 6.4.1 is satisfied as soon as (6.22) and (6.23) are satisfied. As $\beta_1 > 0$ and $\beta_2 > 1$, there exists $n_0 \in \mathbb{N}$ which depends on ψ and L such that for every $n \geq n_0$, (6.22) is satisfied if (6.23) is satisfied, By definition of 2^{j_1} and 2^{j_2}, condition (6.23) is satisfied if

$$\sqrt{b} c_0 n^{-\beta_1\beta_2/(\beta_1+\beta_2+2\beta_1\beta_2)} \left((c_0^{\beta_1} + c_0^{\beta_2}) n^{\beta_1\beta_2/(\beta_1+\beta_2+2\beta_1\beta_2)} + 1 \right) \leq L \left(1 - \sqrt{\frac{7}{12}} \right)$$

which is guaranteed as soon as

$$\sqrt{b} c_0 (c_0^{\beta_1} + c_0^{\beta_2}) \leq \frac{L}{2} \left(1 - \sqrt{\frac{7}{12}} \right) \quad \text{and} \tag{6.25}$$

$$\sqrt{b} c_0 n^{-\beta_1\beta_2/(\beta_1+\beta_2+2\beta_1\beta_2)} \leq \frac{L}{2} \left(1 - \sqrt{\frac{7}{12}} \right). \tag{6.26}$$

(6.25) is satisfied as soon as

$$2c_0^{\beta_1+\beta_2+1} \leq \frac{L}{2\sqrt{b}} \left(1 - \sqrt{\frac{7}{12}} \right),$$

which is guaranteed by definition of c_0. Moreover, there exists an integer n_1 such that (6.26) is satisfied for every $n \geq n_1$.

Thus for every $n \geq \max(n_0, n_1)$, (1), (2) and (3) in Theorem 6.4.1 are satisfied with $r = B_0 n^{-\overline{\beta}/(\overline{\beta}+1)}$, which concludes the proof of Proposition 6.3.2. □

6.5 Appendix

6.5.1 Talagrand Inequality

We use the following form of Talagrand Inequality.

Theorem 6.5.1 *Let $(V_1, \ldots, V_n)$ be a sample of independent random variables, and $\mathcal{F}$ be a set of applications from $\mathbb{R}$ to $\mathbb{R}^n$ which has a countable dense subspace for the norm $\|.\|_\infty$. Let*

$$Z = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} (f^{(i)}(V_i) - \mathbb{E}[f^{(i)}(V_i)]) \right|.$$

Let b, v and $\mathbb{H}$ be such that

$$\sup_{f \in \mathcal{F}} \sup_{i=1,\ldots,n} \|f^{(i)}\|_\infty \leq b,$$

$$\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} Var(f^{(i)}(X_i)) \leq v,$$

$$\mathbb{E}[Z] \leq \mathbb{H}.$$

Then for every $\theta > 1$, there exist positive numerical constants $\overline{C}$, $\overline{C}'$, $\overline{K}$, $\overline{K}'$ such that for every n,

$$\mathbb{E}\left[(Z^2 - \theta \mathbb{H}^2)_+\right] \leq \overline{C} \frac{v}{n} \exp\left(-\overline{\kappa} \frac{n\mathbb{H}^2}{v}\right) + \overline{C}' \frac{b^2}{n^2} \exp\left(-\overline{\kappa}' \frac{n\mathbb{H}}{b}\right). \tag{6.27}$$

Proof of Theorem 6.5.1

Theorem 6.5.1 is obtained from the following result by Klein and Rio (2005).

Theorem 6.5.2 *Let $(V_1, .., V_n)$ be a sample of independent random variables. Let $\mathcal{S}$ be a countable set of applications from $\mathbb{R}$ to $[-1, 1]$. Let*

$$Z' = \sup_{s \in \mathcal{S}} \sum_{i=1}^{n} s^{(i)}(V_i).$$

Assume that $\mathbb{E}[s^{(i)}(V_i)] = 0$ for every $i \in \{1, \ldots, n\}$. Then

$$P[Z' \geq \mathbb{E}[Z'] + x] \leq \exp\left(-\frac{x^2}{2(V + 2\mathbb{E}[Z']) + 3x}\right)$$

where $V = \sup_{s \in \mathcal{S}} Var(\sum_{i=1}^{n} s^{(i)}(V_i))$.

• By density arguments, Theorem 6.5.2 can be generalised to a set of function $\mathcal{S}$ which has a countable dense subset for the norm $\|.\|_\infty$.

• We note that

$$Z = \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} (f^{(i)}(V_i) - \mathbb{E}[f^{(i)}(X_i)]) \right| = \sup_{f\in\mathcal{F}\cup(-\mathcal{F})} \left(\frac{1}{n}\sum_{i=1}^{n} (f^{(i)}(V_i) - \mathbb{E}[f^{(i)}(X_i)]) \right).$$

Moreover,

$$\sup_{f\in\mathcal{F}} \sup_{i=1,\dots,n} \|f^{(i)}\|_\infty = \sup_{f\in\mathcal{F}\cup(-\mathcal{F})} \sup_{i=1,\dots,n} \|f^{(i)}\|_\infty \quad \text{and}$$
$$\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} Var(f^{(i)}(X_i)) = \sup_{f\in\mathcal{F}\cup(-\mathcal{F})} \frac{1}{n}\sum_{i=1}^{n} Var(f^{(i)}(X_i)).$$

Thus it is enough to prove (6.27) for

$$Z = \sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^{n} (f^{(i)}(V_i) - \mathbb{E}[f^{(i)}(X_i)]) \right).$$

• Let $Z' = nZ/b$,

$$Z' = \sup_{s\in\mathcal{S}} \sum_{i=1}^{n} s^{(i)}(V_i)$$

with

$$\mathcal{S} = \left\{ s = (s^{(1)}, \dots, s^{(n)}), s^{(i)} : x \in \mathbb{R} \to \frac{f^{(i)}(x)}{b} - \mathbb{E}\left[\frac{f^{(i)}(V_i)}{b}\right] \in [-1,1], \forall i = 1,\dots,n \right\}.$$

The $\{V_i\}$'s are independent so

$$\begin{aligned} V = \sup_{s\in\mathcal{S}} Var\left(\sum_{i=1}^{n} s^{(i)}(V_i)\right) &= \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} Var\left(\frac{1}{b}(f^{(i)}(V_i) - \mathbb{E}[(f^{(i)}(V_i)]\right) \\ &= \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} Var\left(\frac{1}{b}(f^{(i)}(V_i)\right) \\ &\leq \frac{nv}{b^2}. \end{aligned}$$

Besides,

$$\mathbb{E}[Z'] = \frac{n}{b}\mathbb{E}[Z] \leq \frac{n\mathbb{H}}{b}.$$

Thus, for every $x > 0$, Theorem 6.5.2 implies

$$\begin{aligned} P[Z \geq \mathbb{H} + x] &\leq P[Z \geq \mathbb{E}[Z] + x] \\ &= P\left[Z' \geq \mathbb{E}[Z'] + \frac{nx}{b}\right] \\ &\leq \exp\left(-\frac{n^2x^2/b^2}{2(V + 2\mathbb{E}[Z']) + 3nx/b}\right) \\ &= \exp\left(-\frac{nx^2}{2v + 4b\mathbb{H} + 3bx}\right). \end{aligned}$$

Then, we apply the above inequality with $x = y + \nu\mathbb{H}$ for some positive y and ν. As $x^2 \geq y^2 + 2\nu\mathbb{H}$, and for every positive numbers a, b, c,

$$\frac{1}{a+b+c} \geq \frac{1}{3}\min\left(\frac{1}{a}, \frac{1}{b}, \frac{1}{c}\right),$$

$$\begin{aligned} P[Z \geq (1+\nu)\mathbb{H} + y] &\leq \exp\left(-\frac{n(y^2 + 2\nu\mathbb{H}y)}{2v + 4b\mathbb{H} + 3by + 3b\nu\mathbb{H}}\right) \\ &\leq \exp\left(-\frac{n}{3}\min\left\{\frac{y^2}{2v}, \frac{y^2}{3by}, \frac{2\nu\mathbb{H}y}{(4b\mathbb{H} + 3b\nu\mathbb{H})}\right\}\right) \\ &= \exp\left(-\frac{n}{3}\min\left\{\frac{y^2}{2v}, \frac{y}{b}, \frac{2\nu y}{4b + 3b\nu}\right\}\right). \end{aligned}$$

Moreover, if $\nu \leq 1$,

$$\frac{2\nu y}{4b + 3b\nu} \geq \frac{2\nu y}{7b}$$

and if $\nu \geq 1$,

$$\frac{2\nu y}{4b + 3b\nu} \geq \frac{2\nu y}{4b\nu + 3b\nu} = \frac{2y}{7b}.$$

Hence, for every $\nu > 0$,

$$\frac{2\nu y}{4b + 3b\nu y} \geq 2\min(1, \nu)\frac{y}{7b}.$$

So,

$$\min\left\{\frac{2\nu y}{4b + 3b\nu}, \frac{y}{b}\right\} \geq \frac{y}{b}\min\left\{\frac{2\min(1,\nu)}{7}, 1\right\} = \frac{2\min(1,\nu)y}{7b}.$$

Thus,

$$P[Z \geq (1+\nu)\mathbb{H} + y] \leq \exp\left(-\frac{n}{3}\min\left\{\frac{y^2}{2v}, \frac{2\min(1,\nu)y}{7b}\right\}\right). \tag{6.28}$$

Besides, for every random variable X of density f_X,

$$\begin{aligned}\mathbb{E}[X_+] &= \int_0^\infty x f_X(x)dx \\ &= \int_0^\infty -x\frac{d(P[X \geq x])}{dx}dx.\end{aligned}$$

Noting that $\lim_{x\to+\infty} xP[X \geq x] \leq \mathbb{E}[X_+]$, we can integrate by part.

$$\mathbb{E}[X_+] = -\lim_{x\to+\infty} xP[X \geq x] + \int_0^\infty P[X \geq x]dx \leq \int_0^\infty P[X \geq x]dx.$$

Thus

$$\begin{aligned}\mathbb{E}\left[(Z^2 - \theta\mathbb{H}^2)_+\right] &\leq \int_0^{+\infty} P[(Z^2 - \theta\mathbb{H}^2)_+ \geq s]ds \\ &\leq \int_0^{+\infty} P[|Z| \geq \sqrt{\theta\mathbb{H}^2 + s}]ds.\end{aligned}$$

As $\theta > 1$, there exist θ_1, θ_2, $\theta_3 > 0$ such that $\theta = (1+\theta_1)(1+\theta_2)^2 + \theta_3$. Moreover, for every $x, y \geq 0$, $\sqrt{(1+\theta_1)x + (1+1/\theta_1)y} \geq \sqrt{x} + \sqrt{y}$, hence

$$\begin{aligned}\mathbb{E}\left[(Z^2 - \theta\mathbb{H}^2)_+\right] &\leq \int_0^{+\infty} P\left[|Z| \geq \sqrt{(1+\theta_1)(1+\theta_2)^2\mathbb{H}^2 + \theta_3\mathbb{H}^2 + s}\right] ds \\ &\leq \int_0^{+\infty} P\left[|Z| \geq (1+\theta_2)\mathbb{H} + \sqrt{\frac{\theta_3\mathbb{H}^2 + s}{1+1/\theta_1}}\right] ds.\end{aligned}$$

According to (6.28),

$$\begin{aligned}&\mathbb{E}\left[(Z^2 - \theta\mathbb{H}^2)_+\right] \\ &\leq \int_0^{+\infty} \exp\left(-\frac{n}{3}\min\left\{\frac{\theta_3\mathbb{H}^2 + s}{2v(1+1/\theta_1)}, \frac{2\min(1,\theta_2)\sqrt{\theta_3\mathbb{H}^2 + s}}{7b\sqrt{1+1/\theta_1}}\right\}\right) ds \\ &\leq \int_0^{+\infty} \exp\left(-\frac{n(\theta_3\mathbb{H}^2 + s)}{6v(1+1/\theta_1)}\right) ds + \int_0^{+\infty} \exp\left(-\frac{2\min(1,\theta_2)n\sqrt{\theta_3\mathbb{H}^2 + s}}{21b\sqrt{1+1/\theta_1}}\right) ds\end{aligned}$$

$\sqrt{\theta_3\mathbb{H}^2 + s} \geq \sqrt{\theta_3\mathbb{H}^2/2} + \sqrt{s/2}$, hence

$$\begin{aligned}&\mathbb{E}\left[(Z^2 - \theta\mathbb{H}^2)_+\right] \\ &\leq \int_0^{+\infty} \exp\left(-\frac{n(\theta_3\mathbb{H}^2 + s)}{6v(1+1/\theta_1)}\right) ds + \exp\left(-\frac{2n\min(1,\theta_2)(\sqrt{\theta_3}\mathbb{H} + \sqrt{s})}{21b\sqrt{2(1+1/\theta_1)}}\right) ds.\end{aligned}$$

We recall that

$$\int_0^{+\infty} \exp(-Cs)ds = \frac{1}{C} \quad \text{and} \quad \int_0^{+\infty} \exp(-C\sqrt{s})ds = \frac{2}{C^2}, \quad \forall C > 0.$$

Thus,

$$\begin{aligned}\mathbb{E}\left[(Z^2 - \theta\mathbb{H}^2)_+\right] \quad &\leq \frac{6(1+1/\theta_1)v}{n} \exp\left(-\frac{\theta_3}{6(1+1/\theta_1)}\frac{n\mathbb{H}^2}{v}\right) \\ &+ \left(\frac{21}{2\min(1,\theta_2)}\right)^2 (2(1+1/\theta_1))\frac{b^2}{n^2} \exp\left(-\frac{2\min(1,\theta_2)\sqrt{\theta_3}}{21\sqrt{2(1+1/\theta_1)}}\right)\end{aligned}$$

which ends the proof of Theorem 6.5.1. □

6.5.2 Linear algebra

Lemma 6.5.1 *Let V be a linear subspace of a vector space E with $Dim(V) = D < \infty$. Let $\langle s,t\rangle_0$ be a scalar product on E, and $\|t\|_0 = \sqrt{\langle t,t\rangle_0}$ the corresponding semi-norm. Then there exists a basis $(\varphi_1,\ldots,\varphi_D)$ of V which is orthogonal for the $\|.\|_0$-norm, and such that $\|\varphi_j\|_0 = 0$ or 1 for every $j = 1,\ldots,D$.*

Proof of Lemma 6.5.1

Let $(\psi_1,\ldots,\psi_D)$ be a basis of V. The proof follows the Gram Schmidt orthogonalisation procedure, but with a possibly linearly dependent family.

• Let $\tilde{\varphi}_1 = \psi_1$.

• Let $\tilde{\varphi}_2 = \psi_2 + a\tilde{\varphi}_1$ be such that

$$\langle \tilde{\varphi}_2, \tilde{\varphi}_1\rangle_0 = 0 \quad \Leftrightarrow \quad \langle \psi_2, \tilde{\varphi}_1\rangle_0 + a\|\tilde{\varphi}_1\|_0^2 = 0.$$

If $\|\tilde{\varphi}_1\|_0 = 0$, with Cauchy Schwartz Inequality, $\langle \psi_2, \tilde{\varphi}_1\rangle_0 = 0$ as well and we set

$$a = \begin{cases} 0 \text{ if } \|\tilde{\varphi}_1\|_0 = 0 \\ -\frac{\langle \psi_2, \tilde{\varphi}_1\rangle_0}{\|\tilde{\varphi}_1\|_0^2} \text{ otherwise.} \end{cases}$$

• For every $k \in \{1,\ldots,D-1\}$, we set $\tilde{\varphi}_{k+1} = \psi_{k+1} + \sum_{j=1}^k a_j\tilde{\varphi}_j$ where

$$a_j = \begin{cases} 0 \text{ if } \|\tilde{\varphi}_j\|_0 = 0 \\ -\frac{\langle \psi_{k+1}, \tilde{\varphi}_j\rangle_0}{\|\tilde{\varphi}_j\|_0^2} \text{ otherwise.} \end{cases}$$

Thus, for every $k \in \{1,\ldots,D\}$, $Vect(\psi_1,\ldots\psi_k\} = Vect(\tilde{\varphi}_1,\ldots,\tilde{\varphi}_k)$ and the $(\tilde{\varphi}_j)$'s are orthogonal for the $\|.\|_0$ semi-norm. Finally, let

$$\varphi_j = \begin{cases} \tilde{\varphi}_j \text{ if } \|\tilde{\varphi}_j\|_0 = 0 \\ \frac{\tilde{\varphi}_j}{\|\tilde{\varphi}_j\|_0} \text{ otherwise.} \end{cases} \quad \square$$

Concluding remark about regression-type estimators

In Chapters 2 and 6, we have estimated a regression function by minimisation of a least square contrast on models S_m. We have shown that the coefficients of the resulting estimator satisfy an equation which brings into play the Gram matrix of a basis of S_m for the empirical norm associated to the designs. But this form of coefficients also appears in other frameworks, like in Chapter 4 for hazard rate estimation.

More precisely, let g be a function to estimate from a sample $(U_i, W_i)_{i=1,\dots,n}$ and $\mathcal{M}_n = \{S_m, m \in J_n\}$ be a collection of models. Let $\{\phi_1, \dots, \phi_{D_m}\}$ be a basis of S_m, the non adaptive estimator $\widehat{g}_m$ on S_m is called a regression-type estimator if $\widehat{g}_m = \sum_{k=1}^{D_m} \widehat{a}_k \phi_k$ and the vector $\widehat{A}_m = [\widehat{a}_1, \dots, \widehat{a}_{D_m}]^t$ satisfies

$$\widehat{G}_m \widehat{A}_m = \widehat{V}_m \tag{6.29}$$

where $\widehat{G}_m$ is the Gram matrix of $\{\phi_1, \dots, \phi_{D_m}\}$ for the empirical norm

$$\|t\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} t^2(U_i)$$

and

$$\widehat{V}_m = \left(\frac{1}{n} \sum_{i=1}^{n} W_i \phi_1(U_i), \dots, \frac{1}{n} \sum_{i=1}^{n} W_i \phi_{D_m}(U_i) \right)^t.$$

In such a context, we generally assume that the norm $\|t\|_{f_U} = \sqrt{\int t^2(x) f_U(x) dx}$ is equivalent to the canonical L^2 norm $\|.\|$, which is guaranteed if f_U is lower and upper bounded by positive constants.

In this manuscript, we have considered three regression-type estimators:
- the regression function estimator $\widehat{b}_m$ in Chapters 2 and 3,
- the hazard rate estimator $\widehat{h}_m$ in Chapters 4 and 5,
- the estimator of cumulative conditional distribution function $\widehat{F}_m$ in Chapter 6.

According to (6.29), $\widehat{g}_m$ is uniquely defined if and only if the matrix $\widehat{G}_m$ is invertible. Moreover, we need that the estimator $\widehat{g}_m$ does not become too large, so the invertibility of $\widehat{G}_m$ is not sufficient: indeed, the coefficients of $\widehat{g}_m$ can become very large if the eigenvalues of $\widehat{g}_m$ are close to 0. For the three estimators presented in this manuscript, we have considered three different ways to take in account this problem which are summarized and compared in this conclusion.

[1] The first approach, used in the regression fuction estimator from Baraud (2002), consists in forcing $\widehat{g}_m$ to remain smaller than a value k_n. More precisely, we replace $\widehat{g}_m$ by the following estimator

$$\tilde{g}_m = \begin{cases} \widehat{g}_m \quad \text{if} \quad \|\widehat{g}_m\| \leq k_n \\ \\ 0 \quad \text{otherwise} \end{cases} \tag{6.30}$$

and k_n is chosen so that the probability $P[\|\widehat{g}_m\| > k_n]$ is small. First of all, the norms $\|.\|_{f_U}$ and $\|.\|$ are equivalent, hence

$$P[\|\widehat{g}_m\| > k_n] \leq P[\|\widehat{g}_m\|_{f_U} > ck_n]$$

for some constant c. Then, we consider the set of large probability

$$A_m = \left\{ \left| \frac{\|t\|^2_{f_U}}{\|t\|^2_n} - 1 \right| \leq \frac{1}{4}, \forall t \in S_m \right\} \tag{6.31}$$

where the norms $\|.\|_{f_U}$ and its empirical counterpart $\|.\|_n$ are close, and

$$P[\{\|\widehat{g}_m\| > k_n\} \cap A_m] \leq P[\|\widehat{g}_m\|_n > c'k_n].$$

Besides, under some assumptions about the collection of models, the probability $P[A^c]$ is smaller than C/n^3 where

$$A = \cup_{m \in I_n} A_m. \tag{6.32}$$

Now, we have proved in the regression context that $\widehat{g}_m(U) = (\widehat{g}_m(U_1), \ldots, \widehat{g}_m(U_n))$ is the projection of $Y = (Y_1, \ldots, Y_n)$ on the subset

$$S_m(U) = Vect\{\phi_1(U), \ldots, \phi_{D_m}(U)\}$$

of $\mathbb{R}^n$, and this result holds for a general regression-type contrast. Indeed, let U be fixed and $(\varphi_1, \ldots, \varphi_{D_m})$ be a $\|.\|_n$-orthogonal basis of S_m and $\widehat{g}_m = \sum_{k=1}^{D_m} \widehat{b}_k \varphi_k$. Consider equality (6.29) in the basis $(\varphi_1, \ldots, \varphi_{D_m})$, then the matrix $\widehat{G}_m$ is equal to identity. Hence

$$\widehat{B}_m = \left[\widehat{b}_1, \ldots, \widehat{b}_{D_m}\right]^t = \left(\frac{1}{n} \sum_{i=1}^n W_i \varphi_1(U_i), \ldots, \ldots \frac{1}{n} \sum_{i=1}^n W_i \varphi_{D_m}(U_i) \right)^t$$

which is equivalent to

$$\langle \widehat{g}_m, \varphi_k \rangle_n = \frac{1}{n} \sum_{i=1}^{n} W_i \varphi_k(U_i), \quad \forall k = 1, \dots, D_m$$

$$\Leftrightarrow \langle \widehat{g}_m(U), \varphi_k(U) \rangle = \langle W, \varphi_k(U) \rangle, \quad \forall k = 1, \dots, D_m$$

where $\langle ., . \rangle$ denotes the canonical scalar product on $\mathbb{R}^n$. Thus, $\widehat{g}_m(U)$ is the projection of W on $S_m(U)$ so

$$\|\widehat{g}_m\|_n^2 = \frac{1}{n} \|\widehat{g}_m(U)\|^2 \leq \frac{1}{n} \|W\|^2 = \frac{1}{n} \sum_{i=1}^{n} W_i^2.$$

Therefore, Markov Inequality entails

$$P[\|\widehat{g}_m\|_n > c k_n] \leq \frac{C}{k_n^2} \mathbb{E}[W_1^2].$$

Finally, after the model selection procedure, the risk of $\tilde{g}_{\widehat{m}}$ decomposes in $\mathbb{E}\left[\|\tilde{g}_{\widehat{m}} - g\|_{f_U}^2 \mathbb{1}_{\{\|\widehat{g}_{\widehat{m}}\| \leq k_n\} \cap A}\right]$ which provides the main term in the oracle inequality and

$$\begin{aligned} \mathbb{E}\left[\|\tilde{g}_{\widehat{m}} - g\|_{f_U}^2 \mathbb{1}_{\{\|\widehat{g}_{\widehat{m}}\| > k_n\} \cup A^c}\right] &\leq \|g\|_{f_U}^2 P\left[\|\widehat{g}_{\widehat{m}}\| > k_n \cap A\right] + \left(\|g\|_{f_U} + k_n\right)^2 P\left[A^c\right] \\ &\leq \frac{C'}{k_n^2} + C'' \frac{k_n}{n^3} \end{aligned}$$

Then, if we set $k_n = n$, the term above is smaller than C'''/n.

[2] Another procedure consists in considering directly a set where the matrix $\widehat{G}_m$ is invertible "enough", that is where the eigenvalues of $\widehat{G}_m$ are larger than a threshold which is chosen by the following argument. Consider the matrix $G_m = \mathbb{E}[\widehat{G}_m]$, then G_m is the Gram matrix of the basis $\{\phi_1, \dots, \phi_{D_m}\}$ for the norm $\|.\|_{f_U}$. By assumption, f_U is lower bounded by $h_0 > 0$. Hence let λ be an eigenvalue of G_m, and Z an eigenvector related to λ, then

$$\begin{aligned} G_m U = \lambda Z \quad &\Rightarrow \quad Z^t G_m Z = \lambda Z^t Z \\ &\Leftrightarrow \quad \left\| \sum_{k=1}^{D_m} z_k \phi_k \right\|_{f_U}^2 = \lambda \left\| \sum_{k=1}^{D_m} z_k \phi_k \right\|^2 \\ &\Rightarrow \quad h_0 \left\| \sum_{k=1}^{D_m} z_k \phi_k \right\|^2 \leq \lambda \left\| \sum_{k=1}^{D_m} z_k \phi_k \right\|^2 \\ &\Leftrightarrow \quad h_0 \leq \lambda. \end{aligned}$$

Thus, we build an estimator $\widehat{h}_0$ of h_0 and define

$$\tilde{g}_m = \begin{cases} \widehat{g}_m & \text{if} \quad \min(Sp(\widehat{G}_m)) \geq \frac{1}{2}\widehat{h}_0 \\ \\ 0 & \text{otherwise} \end{cases}$$

Then we upper bound $P\left[\min(Sp(\widehat{G}_m)) < (1/2)\widehat{h}_0\right]$.

$$\begin{aligned} P\left[\min(Sp(\widehat{G}_m)) < \frac{1}{2}\widehat{h}_0\right] &\leq P\left[\min(Sp(\widehat{G}_m)) < \frac{3}{4}h_0\right] + P\left[\frac{3}{4}h_0 < \frac{1}{2}\widehat{h}_0\right] \\ &= P\left[\min(Sp(\widehat{G}_m)) < \frac{3}{4}\min(Sp(G_m))\right] + P\left[h_0 < \frac{2}{3}\widehat{h}_0\right]. \end{aligned}$$

The study of the term $P[h_0 < (2/3)\widehat{h}_0]$ depends on the context, but $\widehat{h}_0$ is generally the minimum of a non adaptive estimator of f_X, and $P[h_0 < (2/3)\widehat{h}_0]$ is upper bounded with a deviation inequality (Bernstein,...). On the other hand, the deviation between $\min(Sp(\widehat{G}_m))$ and $\min(Sp(G_m)$ brings into play the difference between the norms $\|.\|_n$ and $\|.\|_{f_U}$ since $\widehat{G}_m$ ad G_m are the Gram matrix of the basis $\{\phi_1, \ldots, \phi_{D_m}\}$ respectively for the norms $\|.\|_n$ and $\|.\|_{f_U}$. This heuristic is stated more precisely:

$$\min(Sp(\widehat{G}_m)) = \min_{\sum_{k=1}^{D_m} u_k^2 = 1} U^t \widehat{G}_m U = \min_{\sum_{k=1}^{D_m} u_k^2 = 1} \left\| \sum_{k=1}^{D_m} u_k \phi_j \right\|_n^2 = \min_{t \in S_m, \|t\|=1} \|t\|_n^2.$$

Hence

$$\begin{aligned} P\left[\min(Sp(\widehat{G}_m)) < \frac{3}{4}h_0\right] &= P\left[\min_{t \in S_m} \frac{\|t\|_n^2}{\|t\|^2 h_0} < \frac{3}{4}\right] \\ &\leq P\left[\min_{t \in S_m} \frac{\|t\|_n^2}{\|t\|_{f_U}^2} < \frac{3}{4}\right] \\ &\leq P\left[A_m\right] \end{aligned}$$

where A_m is defined in (6.31).

[3] The third approach, developed in Chapter 6, consists in considering, in a first time, the risk associated to the empirical norm $\|.\|_n$. Indeed, as proved in [1], $\widehat{g}_m(U)$ is the projection of W on $S_m(U)$ so $\widehat{g}_m(U)$ is defined and unique, that is to say that $\widehat{g}_m$ is uniquely defined on the set $(U_1, \ldots, U_n)$. Thus, the risk $\mathbb{E}\left[\|\widehat{g}_m - g\|_n^2\right]$ arises naturally.

In Chapter 6, we even prove an oracle inequality with the risk $\mathbb{E}\left[\|\widehat{g}_{\widehat{m}} - g\|_n^2 | U\right]$ for U a.s. (which entails an oracle inequality for the risk $\mathbb{E}\left[\|\widehat{g}_{\widehat{m}} - g\|_n^2\right]$):

$$\mathbb{E}\left[\|\widehat{g}_{\widehat{m}} - g\|_n^2 | U\right] \leq C_1 \inf_{m \in J_n} \left\{ \inf_{t \in S_m} \|g - t\|_n^2 + pen(m) \right\} + \frac{C_2}{n}. \tag{6.33}$$

This result has the advantage to be directly transposable to the fixed design context: assume that $(U_1 = v_1, \ldots, U_n = v_n)$ are non random, and consider the risk associated to the non random norm $\|t\|_n^2 = (1/n)\sum_{i=1}^n t^2(v_i)$ (this risk is classical in fixed design regression, see for example Baraud (2000)), then the same oracle inequality holds. Moreoever, this result is obtained under very few assumptions on the collection of models and generates constants C_1 and C_2 smaller than in the L^2-risk oracle inequality.

Nevertheless, the L^2-risk $\mathbb{E}\left[\|\widehat{g}_{\widehat{m}} - g\|_{f_U}^2\right]$ is more classical. First of all, it enables to conduct a minimax study, which would be much more difficult with an empirical risk. Besides, in regression context, the behaviour of the estimator at other points than the design is one of the main purpose of regression function estimation (for example to know the quality of a prediction).

The oracle inequality for the L^2-risk is inferred from (6.33). Indeed, on the set A (defined in (6.32)) , $\mathbb{E}\left[\|\widehat{g}_{\widehat{m}} - g\|_{f_U}^2\right] \leq 4\mathbb{E}\left[\|\widehat{g}_{\widehat{m}} - g\|_n^2\right]$ is upper bounded by (6.33). To control the risk on A^c, similarly to [1] and [2], we need to restrict the definition of the estimator on a set on which $\widehat{g}_m$ is not too large, and put $\widehat{g}_m = 0$ otherwise. The distribution function F studied in Chapter 6 is especially simple since we know a priori that $F(x,t) \in [0,1]$ for every (x,t), but in a general case, we consider rather a restriction like (6.30). Moreover, we prove that

$$P[A^c] \leq \frac{C_3}{n}.$$

This result requires stronger assumptions on the collection of models, in particular a restriction of the dimension of the models, and provides a very large theoretical constant C_3.

Conclusion

Finally, we note that these three approaches, even if they are differently presented, are actually very similar. The risk is estimated on two sets: a set of "good estimation" where $\widehat{g}_m$ is well defined, which generates the main term $\inf_{m\in J_n}\left\{\inf_{t\in S_m}\|t-g\|_{f_U}^2 + pen(m)\right\}$ in the oracle inequality, and a set of "bad estimation" where $\|g - \widehat{g}_{\widehat{m}}\|_{f_U}$ is forced to remain bounded and which has a small probability depending on $P[A^c]$.

- In [1], the set of good estimation is $\{\|\widehat{g}_m\| \leq k_n\} \cap A$, and the set of bad estimation is $\{\|\widehat{g}_m\| > k_n\} \cup A^c$

- In [2], the set of good estimation is $\{\min(Sp(\widehat{G}_m)) \geq \widehat{h}_0\}$ and the set of bad estimation is $\{\widehat{h}_0 < (2/3)h_0\} \cap \{\min(Sp(\widehat{G}_m)) < (3/4)h_0\}$ and we have proved that $\{\min(Sp(\widehat{G}_m)) < (3/4)h_0\} \subset A$. Contrary to [1] where the threshold k_n is arbitrary fixed, the threshold $\widehat{h}_0$ has an interpretation and provides an information about the quality of the estimation.

- The procedure presented in [3] is different: contrary to the two cases above, we first consider a set of values where $\widehat{g}_m$ is uniquely defined: $(U_1, \ldots, U_n)$. Then, the upper bound of

the risk on this set of values can be immediately transferred to the set of good estimation $A \cap \{\widehat{g}_m\| \leq k_n\}$. The two steps described in this method are actually present in the proof of procedures [1] and [2]. Thus, the third approach allows us to distinguish the assumptions which come from the framework and and those which come from the equivalence between the empirical and L^2 norms. But if one is only interested in the L^2-risk, procedures [1] and [3] are equivalent.

Besides, we have summarized these three methods for a general regression-type estimator to underline the fact that each procedure can be applied to any of the three frameworks considered in this manuscript.

Bibliographie

Bibliographie

N. Akakpo and C. Durot. Histogram selection for possibly censored data. *Math. Methods Statist.*, to appear, 2010.

M. G. Akritas and I. Van Keilegom. Non parametric estimation of the residuals distribution. *Scan. J. Statis*, 28(3):549–567, 2001.

P.K. Andersen, O Borgan, R. Gill, and N. Kieding. *Statistical models based on counting processes*. Springer series in Statistics. Springer-Verlag, New York, 1993.

A. Antoniadis, G. Grégoire, and G. Nason. Density hazard rate estimation for right-censored data by using wavelet methods. *J. R. Statist. Soc. B*, 61(1):63–84, 1999.

D. Bagkavos and P. Patil. Variable bandwidths for nonparametric hazard rate estimation. *Comm. Statist. Theory Methods*, 38(6-7):1055–1078, 2009.

Y. Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Field*, 117 (6):467–493, 2000.

Y. Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6: 127–146, 2002.

L. Birgé. Interval censoring: a nonasymptotic point of view. *Math. Methods Statist.*, 8(3): 285–298, 1999.

L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

E. Brunel and F. Comte. Adaptive estimation of hazard rate with censored data. *Comm. Statist. Theory Methods*, 37(8-10):1284–1305, 2008.

E. Brunel and F. Comte. Cumulative distribution function estimation under interval censoring case 1. *Electron. J. Stat.*, 3:1–24, 2009.

C. Butucea. Exact adaptive pointwise estimation on Sobolev classes of densities. *ESAIM Probab. Statist*, 5:1–31, 2001.

F. Comte and E. Brunel. Penalised contrast estimation of density and hazard rate with censored data. *Sankhya*, 67(3):441–475, 2005.

F. Comte, J. Dedecker, and M.L. Taupin. Adaptive density deconvolution with dependent inputs. *Math. Methods Statist.*, 17(2):87–112, 2008.

R. A. DeVore and G. G. Lorentz. *Constructive approximation*. Springer-Verlag, Berlin, 1993.

S. Diehl and W. Stute. Kernel density and hazard function estimation in the presence of censoring. *J. Multivariate Anal.*, 25(2):299–310, 1988.

D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2):508–539, 1996.

S. Efromovich. Estimation of the density of regression errors. *Ann. Stat.*, 33(5):2194–2227, 2005.

S. Efromovich. Optimal nonparametric estimation of the density of regression errors with finite support. *Ann. Inst. Statist.Math.*, 59(4):617–654, 2007.

P. Groeneboom and J.A. Wellner. *Information bounds and nonparametric maximum likelihood estimation*, volume 19 of *DMV Seminar*. Birkhäuser Verlag, Basel, 1992.

R. Hochmuth. Wavelet characteriztions for anisotropic Besov spaces. *Appl. Comput. Harmon. Anal.*, 12(2):179–208, 2002.

C. Huber and B. MacGibbon. Lower bounds for estimating a hazard. *Handbook of Statistics*, 23(5):209–226, 2004.

M.G. Hudgens, M.H. Maathuis, and P.B. Gilbert. Nonparametric estimation of the joint distribution of a survival time subject to interval censoring and a continuous mark variable. *Biometrics*, 63(2):372–380, 2007.

E.L. Kaplan and P. Meier. Non parametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, 53:457–481, 1958.

S. Kiwitt, E-R. Nagel, and N. Neumeyer. Empirical likelihood for the error distribution in nonparametric regression models. *Math. Methods Statist.*, 34:511–534, 2008.

T. Klein and E. Rio. Concentration around the mean of empirical processes. *Ann. Probab.*, 33 (3):1060–1077, 2005.

C. Lacour. Adaptive estimation of the transition density of a Markov chain. *Ann. Inst. H. Poincaré Probab. Statist.*, 43(5):571–597, 2007.

C. Laurent, C. Ludena, and C. Prieur. Adaptive estimation of linear functionals by model selection. *Electron. J. Stat.*, 2:993–1020, 2008.

E. le Pennec and V. Rivoirard. Adaptive dantzig density estimation. *Annales de l'IHP*, To appear.

M.R. Leadbetter and G.S. Watson. Hazard analysis. *Bioemtrika*, 51(1 and 2):175–184, 1964.

O. V. Lepski. Asymptotically minimax adaptive estimation. i. upper bounds. optimally adaptive estimates. *Theory Probab. Appl.*, 36(4):682–697, 1991.

O.V. Lepski and V.G. Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Satist.*, 25(6):2512–2546, 1997.

O. V. Lepskii and A. Goldenshluger. Structural adaptation via lp-norm oracle inequalities. *Theory Probab. Related Fields*, (1-2):47–71, 2009.

S. Ma and M.R. Kosorok. Adaptive penalized M-estimation with current status data. *Ann. Inst. Statist. Math.*, 58(3):511–526, 2006.

P. Massart. *Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23*, chapter Concentration inequalities and model selection. Lecture Notes in Mathematics. Springer, Berlin, 2007.

Y. Meyer. *Ondelettes et operateurs*. Hermann, Paris, 1990.

H.G. Muller and J.L. Wang. Hazard rate estimation under random censoring with varying kernels and bandwidth. *Biometrics*, 50(1):61–76, 1994.

W. Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966, 1972.

J.P. Nielsen. Variable bandwidth kernel hazard estimators. *Nonparam. Stat.*, 15(3):355–376, 2003.

S. M. Nikol'skii. Approximation of functions of several variables and imbedding theorems. *Jr. Die Grundlehren der Mathematischen Wissenschaften*, 1975.

P.N. Patil. On the least squares cross validation bandwidth in hazard rate estimation. *Ann. Statist.*, 21(4):1792–1810, 1993.

S. Plancade. Estimation of the density of regression errors by pointwise model selection. *Math. Methods Statis.*, 18(4):341–374, 2009.

S. Plancade. Model selection for hazard rate estimation in presence of censoring. *Metrika*, to appear.

S. Plancade. Non parametric estimation of the density of the regression noise. *C. R. Acad. Sci. Paris*, 346(7-8):461–466, 2008.

P. Reynaud-Bouret. Penalized projection estimators of the Aalen multiple intensity. *Bernoulli*, 12(4):633–661, 2006.

M.A. Tanner and W.H. Wong. The estimation of the hazard function from randomly censored data by the kernel method. *Ann. Statist.*, 11(3):989–993, 1983.

A. B. Tsybakov. *Introduction à l'estimation non paramtrique*. Springer-Verlag, Berlin, 2004.

S. van de Geer. Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.*, 21(1):14–44, 1993.

M.J. van der Laan and A. van der Vaart. Estimating a survival distribution with current status data and high-dimensional covariates. *Int. J. Biostat.*, 2:Art. 9, 42, 2006.

B.S. Yandell. Nonparametric inference for rates with censored survival data. *Ann. Satist.*, 11 (4):1119–1135, 1983.