N° d'ordre : XXXX

THÈSE

présentée en vue de l'obtention du grade de

DOCTEUR

DE L'UNIVERSITÉ PAUL SABATIER

Discipline : Mathématiques appliquées
Spécialité : Statistique

par **JEAN-MICHEL LOUBES**

Estimation non paramétrique par M-estimateurs Pénalisés

Directeurs de Recherche : **Sara** VAN DE GEER **et Michel** LEDOUX

Soutenue le 7/12/2001 devant le
jury composé de Mesdames et Messieurs les Professeurs :

Alain BERLINET	Université de Montpellier II
Fabrice GAMBOA	Université Toulouse III
Michel LEDOUX	Université Toulouse III
Xavier MILHAUD	Université Toulouse III
Dominique PICARD	Université Paris VII
Sara VAN DE GEER	Mathematical Institute of Leiden

au vu des rapports de Madame et Monsieur les Professeurs :

Alain BERLINET	Université de Montpellier II
Dominique PICARD	Université de Paris VII.

Laboratoire de Statistique et Probabilités
UFR MIG - UMR CNRS 58830
Université Paul Sabatier - Toulouse III

à ma famille

Remerciements:

Qu'il me soit tout d'abord permis d'exprimer ma gratitude à S. van de Geer pour avoir accepté de diriger mes recherches avec une sympathie patiente et compréhensive. Je lui sais gré d'avoir guidé mon travail et d'avoir toujours cherché à développer mes initiatives personnelles.

Je tiens également à remercier M. Ledoux, qui a bien voulu m'accueillir au laboratoire de Statistique et Probabilités, pour son écoute, ses nombreux conseils et ses encouragements à communiquer mes résultats.

Mes remerciements s'adressent aussi à D. Picard et A. Berlinet qui m'ont fait l'honneur d'être les rapporteurs de ma thèse. Leurs commentaires pertinents et l'intérêt qu'ils ont accordé à ce travail m'ont permis d'en améliorer la rédaction. En particulier je dois beaucoup à D. Picard qui m'a initié à la statistique et m'a encouragé à entreprendre cette recherche.

Je suis reconnaissant à F. Gamboa et à X. Milhaud d'avoir accepté de faire partie de mon jury, et de l'aide constante qu'ils m'ont apportée dans les différentes étapes de mon travail.

Une partie de la présente thèse a été réalisée au Mathematical Institute de Leiden en Hollande. Je remercie son directeur F. Bakker de m'y avoir fort bien accueilli et E. Belitser qui m'a permis de clarifier certains aspects de ma recherche.

Enfin, je tiens à ajouter que les échanges entre doctorants ont été fructueux, tant au sein du LSP qu'au Mathematical Institute. Leur aide m'a été précieuse. Je leur associerai également F. Pasquotto, E. Le pennec, T. Poullaouec, V. Rivoirard et V. Torri pour leurs conseils et encouragements amicaux.

Merci encore à ma famille “por ser quien soy”.

Contents

Introduction Générale

Dans le présent travail de thèse, nous avons voulu engager une réflexion sur les problèmes d'estimation non paramétrique, i.e lorsque l'espace des paramètres est infini. Très concrètement, nous avons étudié le comportement asymptotique d'une classe particulière d'estimateurs bien connus en statistique, les M-estimateurs pénalisés.

Ces estimateurs sont définis comme réalisant le minimum d'une fonction de perte sous certaines contraintes; ils ont été abondamment étudiés par de nombreux auteurs. Nous renvoyons à des ouvrages généraux et à leur bibliographie pour un début d'exhaustivité, cf par exemple Genont-Catalot et Picard in [GCP93], Birgé et Massart in [BM97], Silverman in [GS94], van de Geer in [vdG00], [vdG01] ou Wegkamp in [vdGW96]. Ils trouvent de nombreuses applications en pratique dans les domaines les plus divers: la physique et l'astronomie in [SM98], l'économie [DFR01] ou encore le traitement de l'image [ROF92] ou [Mum97].

S. van de Geer a montré que le comportement asymptotique de ces estimateurs peut essentiellement être déduit des propriétés géométriques de l'espace où a lieu la minimisation, notamment de leur entropie en utilisant des techniques liées à la théorie des processus empiriques. Ainsi, il n'est pas nécessaire de connaître explicitement l'expression de l'estimateur pour en déduire ses propriétés asymptotiques; cela permet de généraliser cette méthode à un grand nombre de situations. L'entropie, notion développée par Le Cam [LC91], Ibragimov et Ha'sminskii [HI86] ou Birgé [Bir83] pour mesurer la complexité des espaces fonctionnels, apparaît donc aujourd'hui comme un outil fondamental de la statistique asymptotique et permet d'aborder, sous un jour nouveau, les problèmes d'estimation usuels.

C'est dans ce cadre que nous nous sommes proposés de décrire en termes d'entropie le comportement asymptotique de M-estimateurs pénalisés. Cet objectif nous a conduit, dans un premier temps, à une double approche. D'une part, nous avons considéré des estimateurs obtenus par projection sur des bases, notamment des bases d'ondelettes, et dont les coefficients ont été soit lissés (multipliés par des paramètres de régularisation), soit seuillés (mis à zéro lorsqu'ils sont inférieurs à un certain seuil). D'autre part, nous nous sommes intéressés aux estimateurs obtenus en statistique Bayésienne. Ainsi, une même approche générale montre la convergence de ces estimateurs particuliers.

Dans un second temps, nous nous sommes attachés à dégager des procédures d'estimation adaptative qui rendent possible l'estimation d'une fonction, sans faire d'hypothèses a priori sur sa régularité tout en obtenant un estimateur qui converge à la vitesse optimale au sens minimax.

Nous avons choisi d'organiser la synthèse de nos travaux sous la forme suivante: une première partie d'introduction générale, le Chapitre 1, où nous présentons de façon succincte nos résultats, ensuite, viennent quatre chapitres qui les reprennent, les détaillent en proposant de nouvelles perspectives. Chacun d'eux (à l'exception du Chapitre 3) est articulé autour d'un article principal. La bibliographie est regroupée en fin de document.

Plus précisément, au Chapitre 2, nous commençons par présenter un théorème général qui décrit le comportement asymptotique de M-estimateurs en fonction de l'entropie de la classe de fonctions associée à la pénalité. En conformité avec cette approche, dans un article écrit en collaboration avec S. VAN DE GEER, nous avons étudié l'estimateur de seuillage doux en tant qu'estimateur minimisant une fonction de perte quadratique et une pénalité l^1 et généralisé ses propriétés d'adaptivité asymptotique à l'estimateur robuste du principe de déviation absolu pénalisé. En annexes de cette partie, nous présentons une généralisation de ces derniers résultats à l'estimateur du maximum de vraisemblance pénalisé ainsi que des simulations des estimateurs considérés et les moyens qui ont été mis en oeuvre pour les obtenir.

Nous avons consacré le Chapitre 3 à l'étude des estimateurs lissés, obtenus en minimisant une perte quadratique et une norme d'un espace de Besov. Les bases d'ondelettes nous ont permis d'écrire ces normes comme des combinaisons à poids des coefficients d'ondelettes. Nous avons étudié le comportement des estimateurs obtenus et avons expliqué comment, dans la pratique, il est possible de choisir le paramètre de lissage par des méthodes de validation croisée.

Au Chapitre 4 nous avons montré comment les estimateurs Bayésiens pouvaient être étudiés comme des M-estimateurs pénalisés. La fonction de perte, choisie par le statisticien, joue ici un rôle similaire à la loi a priori. Grâce à ces premiers résultats, nous construisons un estimateur basé sur la maximisation de la loi a posteriori, de forme analogue à celle d'un estimateur lissé autorisant une estimation adaptative. Cette méthode s'apparente à une procédure de sélection de modèles qui privilégierait ceux qui, parmi eux, sont constitués des fonctions les plus régulières.

Enfin, le Chapitre 5 constitue une extension de nos travaux à des fonctions très irrégulières, dont la régularité varie rapidement et n'a de sens que de façon locale: les fonctions multifractales qui échappent aux théories d'estimation actuelles. Pourtant, la connaissance a priori des caractéristiques de ces fonctions, définies par l'histogramme de leurs coefficients d'ondelettes nous offre la possibilité de construire un estimateur, maximisant la log-vraisemblance a posteriori, et de donner son comportement asymptotique. Ce travail a été réalisé en collaboration avec F. GAMBOA.

Chapter 1

Introduction et Préliminaires

Dans ce chapitre, nous commençons par rappeler les notions à la base de notre travail ainsi que les définitions des différents objets qui nous intéressent. Chaque section renvoie à un chapitre ultérieur en présentant, de manière succincte une partie des résultats qui y sont développés.

1.1 M-estimation

Dans tout notre travail, nous considérons le modèle de régression non paramétrique suivant:

$$\begin{cases} Y_i = \theta_0(z_i) + W_i, \ i = 1, \ldots, n \\ \theta_0 \in \Theta \end{cases} \tag{1.1.1}$$

Les Y_i désignent n observations qui proviennent d'une part d'une fonction $\theta_0 : \mathbb{L} \to \mathbb{R}$, le paramètre d'intérêt, observée aux points $(z_i) \in \mathbb{L}^n$ où $\mathbb{L}$ est un compact de $\mathbb{R}$, et d'autre part d'erreurs d'observation, les W_i, $i = 1, \ldots, n$, des variables aléatoires centrées. Notre connaissance a priori de la fonction se traduit par la condition $\theta_0 \in \Theta$, Θ désignant un espace fonctionnel connu. Considérons γ une fonction de perte qui mesure la différence entre deux quantités et une pénalité $I : \Theta \to \mathbb{R}$ portant sur la régularité des fonctions de l'espace Θ. Plus la fonction est irrégulière, dans un sens qu'il nous faudra préciser, et plus la pénalité est grande. Pour toute suite de réels λ_n^2, nous définissons $\hat{\theta}_n$, l'estimateur pénalisé de θ_0, comme toute variable aléatoire qui minimise la somme de cette fonction de perte et de la pénalité, la contribution des deux termes étant équilibrée par la suite λ_n^2. Plus précisément:

Definition 1.1.1. *Définissons le M-estimateur pénalisé $\hat{\theta}_n$ comme*

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \left(\gamma(Y, \theta) + \lambda_n^2 I(\theta) \right) \tag{1.1.2}$$

λ_n^2 peut être considéré comme un paramètre de régularisation. En effet, il existe un équilibre entre les deux termes de (1.1.2). Plus λ_n^2 est grand, plus le second terme va être prédominant et plus l'estimateur obtenu sera régulier. D'un autre côté, plus λ_n^2 sera proche de zéro, et plus l'estimateur sera proche des données, proche au sens déterminé par le choix de la fonction de perte. Si le minimum n'est pas atteint, nous considérons l'estimateur minimisant le précédent critère à une constante ϵ_n près:

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \left(\gamma(Y, \theta) + \lambda_n^2 I(\theta) + \epsilon_n \right) \tag{1.1.3}$$

Pour $\epsilon_n = o\left(\frac{1}{n}\right)$, les propriétés asymptotiques de l'estimateur sont inchangées. C'est pourquoi nous supposons dans la plupart de notre étude que le minimum est effectivement atteint et nous prenons $\epsilon_n = 0$. Dans quelques cas particuliers, décrits dans la partie 3.1.1, nous prouverons cette hypothèse.

Afin d'éviter d'avoir à imposer de conditions sur le schéma de discrétisation du problème $(z_i)_{i=1,\ldots,n}$, nous considérons des métriques basées sur cette discrétisation et nous énonçons les propriétés asymptotiques en utilisant ces métriques. Pour cela, nous définissons la mesure empirique

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{z_i} \tag{1.1.4}$$

ainsi que la norme quadratique empirique associée:

$$\forall \theta \in \Theta, \ \|\theta\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} \theta(z_i)^2 \tag{1.1.5}$$

La théorie des M-estimateurs a été largement étudiée par de nombreux auteurs, parmi lesquels nous pouvons citer tout particulièrement S. van de Geer, M. Wegkamp in [vdG90], [vdGW96], [vdG00], A. Berlinet, F. Liese et I. Vajda in [BLV00] ou [Vaj99], M. Kohler in [Koh99] ou [Koh00]. L'introduction d'une fonction de pénalité est justifiée par le fait que, dans les cas où peu de choses sont connues sur la fonction à estimer, ce qui se traduit par une classe de fonctions Θ trop grande, un estimateur qui ne minimiserait qu'une fonction de perte serait trop irrégulier et ne convergerait peut-être pas vers la vraie fonction. Par exemple, pour une perte quadratique, l'estimateur obtenu interpole simplement les données Y_i aux points z_i. Ainsi ajouter une contrainte, sous la forme d'une fonction pénalisant la régularité de l'estimateur, apparaît naturel dans le contexte de l'estimation fonctionnelle. L'estimation au moyen de fonctions de pénalités a par ailleurs été abordée par quelques auteurs dont Silverman in [SS95], ou encore P. Green in [GS94]. Dans ces ouvrages de référence les auteurs décrivent le comportement asymptotique des estimateurs en résolvant explicitement la minimisation. Mais, dans la plupart des situations rencontrées, il n'est pas possible de résoudre le problème d'optimisation et on ne dispose que d'une expression approchée de l'estimateur comme par exemple, en imagerie, pour la célèbre fonctionnelle de Mumford in [Mum97]. Dans ces cas, il est nécessaire d'utiliser des techniques liées à la théorie des processus empiriques et développées par S. van de Geer in [vdG90], [vdG00] ou [vdG01] qui prouve que le comportement asymptotique de l'estimateur ne dépend que de la complexité de l'espace sur lequel a lieu la minimisation, complexité caractérisée par l'entropie des sous-espaces $\{\theta \in \Theta, \ \|\theta - \theta_0\|_n \leq 1, \ I(\theta) \leq M\}$ et de la concentration des erreurs d'observation.

Definition 1.1.2. *Entropie $H(\delta, T)$:*
Soit T un sous-ensemble d'un ensemble métrique et $N(\delta, T)$ le nombre minimal de boules pour recouvrir T par des boules de rayon δ. La δ-entropie de T est alors:

$$H(\delta, T) = \log N(\delta, T)$$

Le concept d'entropie a été utilisé par de nombreux auteurs dans la littérature parmi lesquels nous pouvons citer Dudley [BDH+85], Vapnik et Cervonenkis [VC81] ou [Vap00], D. Pollard [Pol84], S. van de Geer et A. van der Vaart [vdVW96].

Dans le second chapitre de ce travail, nous étudierons les liens entre l'entropie d'une classe et la vitesse de convergence des M-estimateurs pénalisés, pour des fonctions de perte générales. C'est dans ce cadre que s'inscrit le théorème suivant:

Theorem 1.1.3. *Pour des erreurs sous Gaussiennes, et sous la condition suivante:*
il existe des constantes $\eta > 0$, $A > 0$ et $s \geq \frac{1}{2}$, telles que

$$H(\delta, \{\theta \in \Theta : \ I(\theta) \leq M, \ \|\theta - \theta_0\|_n \leq \eta\}) \leq A \left(\frac{M}{\delta}\right)^{\frac{1}{s}},$$

$$\textit{pour tout } \delta > 0, \ n \geq 1,$$

pour tout $M \geq M_n$, où $M_n \geq I(\theta_0)$. Alors pour un choix

$$\lambda_n^{-1} = \begin{cases} O_{\mathbf{P}}(n^{\frac{s}{2s+1}} M_n^{\frac{p}{2} - \frac{1}{2s+1}}), & if\ s > \frac{1}{2}, \\ O_{\mathbf{P}}(n^{\frac{1}{4}}) M_n^{\frac{p}{2} - \frac{1}{2}} (\log n)^{-\frac{1}{2}}, & if\ s = \frac{1}{2}, \end{cases}$$

l'estimateur pénalisé des moindres carrés

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \left(\|Y - \theta\|_n^2 + \lambda_n^2 I(\theta)\right)$$

vérifie

$$\|\hat{\theta}_n - \theta_0\|_n = O_{\mathbf{P}}(\lambda_n) M_n^{\frac{p}{2} - \frac{1}{2s+1}}.$$

Nous étendrons ce théorème à des fonctions de perte non quadratiques et à des erreurs non identiquement distribuées.

Les méthodes entropiques, de part leur caractère général, peuvent s'appliquer pour étudier bon nombre d'estimateurs: dans un modéle de régression, dans un modèle logit, dans un modèle d'estimation de densités. Ainsi, la Section 2.3 est consacrée à l'étude de la convergence d'un estimateur de maximum de vraisemblance pénalisé pour estimer une densité dans un espace de Besov.

Nous pouvons remarquer que ces méthodes présentent deux inconvénients:
D'une part, tous les résultats obtenus de cette manière reposent sur des calculs d'entropie,

qui sont parfois peu aisés à mener et qui nécessitent l'emploi de moyens détournés. Par exemple, L. Birgé et P. Massart in [BM97] parviennent à calculer l'entropie d'une boule d'un espace de Besov $B^s_{p\infty}$, $s > \frac{1}{p}$ en utilisant un théorème d'approximation. En Section 2.4.4, nous caractérisons un essemble comme enveloppe convexe de ses points extrémaux. Nous utilisons alors une inégalité de Ball et Pajor in [BP90] pour calculer son entropie et discutons de l'optimalité de cette majoration.
D'autre part le principal défaut de notre approche repose dans le fait que les pénalités, ainsi que le choix optimal du paramètre de lissage, font intervenir la régularité de la fonction à estimer. Or cette quantité est, en pratique, inconnue.
C'est pourquoi, dans un second temps, nous nous sommes particulièrement intéressés à des choix de pénalités ne faisant pas intervenir explicitement des quantités inconnues a priori, mais toujours liées à la régularité de la fonction. Dans la Section 2.2, qui retranscrit un article coécrit avec S. van de Geer, nous avons étudié des M-estimateurs avec des pénalités l^1. Ainsi nous construisons un estimateur adaptatif, c'est-à-dire construit sans connaissance a priori mais qui converge à la vitesse minimax. Considérons une base de $L^2(\mathbf{P}_n)$, les $\psi_j, j = 1, \ldots, n$ et décomposons les fonctions sur cette base en utilisant les notations suivantes: $\theta = \sum_{j=1}^n \alpha_j \psi_j$, $\theta_0 = \sum_{j=1}^n \alpha_{j,0}\psi_j$. Soit $\hat{\theta}_n$ l'estimateur des moindres carrés avec une pénalité l^1, i.e

$$\hat{\theta}_n = \arg \min_{\theta=\alpha_1\psi_1+\ldots+\alpha_n\psi_n} \left\{ \frac{1}{n}\sum_{i=1}^n (Y_i - \theta(z_i))^2 + 2\lambda_n^2 \sum_{j=1}^n |\alpha_j| \right\} = \sum_{j=1}^n \hat{\alpha}_{j,n}\psi_j \tag{1.1.6}$$

Theorem 1.1.4. *Supposons que la variance des erreurs est connue σ_0^2 (pour des raisons de simplicité) et prenons $\lambda_n^2 \geq \sigma_0\sqrt{\frac{2\log n}{n}}$. Pour $\mathcal{J}_n$ un sous-ensemble de $\{1, \ldots, n\}$ nous définissons*

$$N_n = |\mathcal{J}_n|,\ M_n = \sum_{j \notin \mathcal{J}_n} |\alpha_{j,0}|.$$

Alors $\hat{\theta}_n$ vérifie

$$||\hat{\theta}_n - \theta_0||_{Q_n} = O_{\mathbf{P}}(\lambda_n^2 N_n^{\frac{1}{2}} + \lambda_n M_n^{\frac{1}{2}}).$$

Si

$$\sum_{j=1}^n |\alpha_{j,0}|^\rho \leq 1,$$

pour un paramètre d'irrégularité $0 \leq \rho < 1$. Prenons $\mathcal{J}_n = \{j : |\alpha_{j,0}| > \lambda_n^2\}$. Le théorème montre que

$$||\hat{\theta}_n - \theta_0||_{Q_n} = O_{\mathbf{P}}(\lambda_n^{2-\rho}).$$

En d'autres termes, le choix optimal du paramètre de lissage $\lambda_n^2 = \sigma_0\sqrt{\frac{2\log n}{n}}$ permet d'obtenir la vitesse optimale de convergence, qui dépend de la régularité ρ de la vraie fonction, alors que l'estimateur est construit sans utiliser cette information supplémentaire. Dans notre travail, nous étudions de façon similaire l'estimateur du principe de déviation absolu pénalisé

$$\tilde{\theta}_n = \arg \min_{\theta=\alpha_1\psi_1+\ldots+\alpha_n\psi_n} \left\{ \frac{1}{n}\sum_{i=1}^n |Y_i - \theta(z_i)| + 2\lambda_n^2 \sum_{j=1}^n |\alpha_j| \right\} \tag{1.1.7}$$

Nous montrons qu'il possède des propriétés de robustesse et est asymptotiquement pseudo-minimax. Nous utilisons par la suite ce théorème, associé à des décompositions des fonctions dans des bases d'ondelettes pour fournir une alternative aux méthodes de seuillages développées par Donoho, Johnstone, Kerkyacharyan et Picard in [HKPT98] par exemple, et aux méthodes de sélection de modèles. Pour ce dernier point, nous mentionnons les travaux de L. Birgé et P. Massart in [BM97], [BM98], [BBM99] ou de Y. Baraud in [Bar00].
L'estimateur du principe de déviation absolue pénalisé peut sembler difficile à mettre en oeuvre numériquement. Aux techniques de minimisation l^1, décrites par Birkes et Doge in [BD93] et introduites à l'origine en raison de leurs propriétés de robustesse, ont longtemps été préférées des méthodes quadratiques, plus faciles à mettre en oeuvre pratiquement. En effet, minimiser une perte l^1 nécessite la mise en oeuvre de méthodes informatiques complexes proches de l'algorithme du simplexe. Dans la partie 2.4.3 nous présenterons un algorithme dual de point intérieur, basé sur les travaux de Sardy [SS99] qui nous a servi à effectuer nos simulations pour l'estimateur du principe de déviation absolu (1.1.7).

Après avoir décrit le comportement asymptotique des M-estimateurs pénalisés à partir de la seule entropie de la classe de fonctions où a lieu la minimisation, nous avons essayé de décrire leur distribution asymptotique. Si des théorèmes de Limite Centrale pour les M-estimateurs sont bien connus (pour des fonctions monotones de régression, nous nous référons à S. Leurgans in [Leu82] tandis que P. Groeneboom in [Gro85] a montré la normalité asymptotique de l'estimateur de Grenander, qui est lié aux problèmes d'estimation des moindres carrés), il n'en va pas de même pour les M-estimateurs pénalisés. Dans la Section 2.4.2, nous utiliserons des M-estimateurs pénalisés pour tester des hypothèses sur le modèle observé.

1.2 Ondelettes et Analyse Multirésolution

Nous rappelons brièvement les définitions et les premières propriétés des bases d'ondelettes. Nous donnons aussi quelques résultats de théorie d'approximation qui seront utiles pour estimer des fonctions dans des espaces de Sobolev ou de Besov. Pour des résultats plus exhaustifs, nous nous référons aux articles de Meyer et de Donoho, Johnstone, Kerkyacharian et Picard.

1.2.1 Analyse Multiresolution

Nous décrivons le concept d'analyse multirésolution défini par Y. Meyer et S. Mallat in [Mey87], [JM89] et [DMA97]. On peut construire une fonction $\phi \in L^2(\mathbb{R})$, telle que $\int \phi^2 = 1$, et qui vérifie les propriétés suivantes:

1. La famille de fonctions $\{\phi(x-k)\,,\, k \in \mathbb{Z}\}$ forme un système orthonormal dans $L^2(\mathbb{R})$. Soit l'espace $V_0 = Vect\{\phi(x-k)\,,\, k \in \mathbb{Z}\}$.
2. Soit $\phi_{j,k} = 2^{j/2}\phi(2^j x - k)$. Définissons l'espace V_j comme Vect$\{\phi_{j,k}, k \in \mathbb{Z}\}$. Les espaces V_j sont dits emboîtés au sens où $\forall j \in \mathbb{Z} \, : \, V_j \subset V_{j+1}$. En outre, ils vérifient

$\bigcap_{j\in\mathbb{Z}} V_j = \{0\}$, et $L^2(\mathbb{R}) = \overline{\bigcup_{j\in\mathbb{Z}} V_j}$. La fonction ϕ est alors appelée fonction d'échelle de l'analyse multiresolution (M.R.A) $(V_j, j \in \mathbb{Z})$.
Il est possible, lors de cette construction d'imposer pour la fonction ϕ les conditions de régularité suivantes:

3. ϕ est de classe C^r et chacune de ses dérivées jusqu'à l'ordre r est à décroissance rapide. L'analyse multirésolution est alors r-régulière.

4. La fonction ϕ est à support compact.

Sous ces conditions, Meyer in [Mey87] a montré le théorème suivant.

Theorem 1.2.1. *Définissons l'espace W_j comme $V_{j+1} = V_j \bigoplus W_j$, il existe une fonction ψ, l'ondelette telle que:*

1. *$\{\psi(x-k), k \in \mathbb{Z}\}$ est une base orthonormale de W_0.*

2. *$\{\psi_{j,k}, k \in \mathbb{Z}\,, j \in \mathbb{Z}\}$ est une famille orthonormale de $L^2(\mathbb{R})$, où $\psi_{j,k}$ est définie comme précédemment.*

3. *ψ a la même régularité que la fonction ϕ et est aussi à support compact.*

Pour tout $j_0 \in \mathbb{Z}$, on a la décomposition suivante:

$$L^2(\mathbb{R}) = V_{j_0} \bigoplus \bigoplus_{j \geq j_0} W_j.$$

Pour $j_0 \in \mathbb{Z}$, toute fonction $\theta \in L^2$ peut se décomposer de la manière unique suivante:

$$\theta = \sum_{k\in\mathbb{Z}} \alpha_{j_0,k}\phi_{j_0,k} + \sum_{j\geq j_0}\sum_{k\in\mathbb{Z}} \beta_{j,k}\psi_{j,k}$$

où les coefficients d'ondelettes sont définis par:

$$\alpha_{j_0,k} = \int \theta(x)\phi_{j_0,k}(x)\,dx \text{ et } \beta_{j,k} = \int \theta(x)\psi_{j,k}\,dx.$$

Cette représentation est inhomogène dans la mesure où un niveau de résolution est privilégié. On pourra utiliser la représentation homogène suivante:

$$\theta = \sum_{j=-\infty}^{+\infty}\sum_{k} \beta_{jk}\psi_{jk}.$$

1.2.2 Espaces de Besov et approximation

Les espaces de Besov sont utilisés pour décrire les propriétés de régularité d'un grand nombre de fonctions. Pour une étude plus générale, nous nous référons aux ouvrages suivants: J. Peetre [Pee70], H. Triebel [ET92] ou R. De Vore et G. Lorentz [DL93]. Ils apparaissent naturellement comme des espaces de saturation liés aux vitesses minimax en théorie de l'approximation. Nous considérons des espaces de Besov B^s_{pq} définis par trois paramètres : $s > 0$ (un paramétre de régularité), $1 \leq p \leq \infty$ (paramètre L^p), et $1 \leq q \leq \infty$ (paramètre d'interpolation). Tout d'abord, nous rappelons la définition usuelle des espaces de Besov. Soient $1 > s > 0$, $1 \leq p, q \leq \infty$, et $\tau_h f(x) = f(x-h)$, opérateur de translation. Soit $\Delta_h f = \tau_h f - f$, $\omega_p(f,t) = \sup_{|h| \leq t} ||\Delta_h f||_p$ et $\omega^2_p(f,t) = \sup_{|h| \leq t} ||\Delta^2_h f||_p$.

Definition 1.2.2. *L'espace de Besov* $B^s_{pq}(\mathbb{R})$ *avec* $1 \leq p, q \leq \infty$, $s = n + \alpha$ *et* $0 < \alpha \leq 1$ *est défini par*

$$B^s_{pq} = \{f \in L^p \text{ , } \omega^2_p(f,t) = \epsilon(t)t^\alpha \text{ with } ||\epsilon||^*_{\mathrm{q}} < \infty\},$$

en choisissant

$$(||\epsilon||^*_q)^q = \int_0^\infty |\epsilon(t)|^q \frac{dt}{t}.$$

Nous pouvons remarquer que, $\forall \alpha \neq 1$,

$$\int \left(\frac{\omega^1_p(f,t)}{t^\alpha} \right)^q \frac{dt}{t} \leq \frac{1}{(1-\alpha)^q} \int \left(\frac{\omega^2_p(f,t)}{t^\alpha} \right)^q \frac{dt}{t}.$$

Ainsi, il est possible de remplacer ω^2 par ω^1 dans la définition précédente. Soit P_j l'opérateur de projection sur l'espace V_j et $D_j = P_{j+1} - P_j$ l'opérateur de projection sur l'espace W_j. Nous avons alors $P_j = \sum_k \psi_{jk}\bar{\psi}_{jk}$, ce qui permet d'écrire la décomposition suivante:

$$\begin{aligned} P_j f &= \sum_k \alpha_{jk}\phi_{jk} \\ &= \sum_k \alpha_{j0}\phi_{j0} + \sum_{m=0}^{j-1} \sum_k \beta_{mk}\psi_{mk}. \end{aligned}$$

La qualité de l'approximation d'une fonction d'un espace de Besov est mesurée par la propriété suivante énoncé dans [HKPT98] par exemple. Si ϕ est une fonction d'échelle N fois dérivable, soit ψ l'ondelette associée avec N moments nuls, l'équivalence suivante donne une caractérisation en termes d'approximation des espaces de Besov: pour toute fonction $f \in L^p(\mathbb{R})$ nous avons:

$$f \in B^s_{pq},\ 0 < s < N+1 \ => \ ||P_j f - f||_p = 2^{-js}\epsilon_j$$

$$\mathrm{o}\ (\epsilon_{\mathrm{j}}) \in \mathrm{l}^{\mathrm{q}}.$$

Nous pouvons ainsi donner une caractérisation des espaces de Besov au moyen des coefficients d'ondelettes

Theorem 1.2.3. *$f \in B^s_{pq}$ si et seulement si la semi-norme $J^{'}_{spq}$ est telle que*

$$J^{'}_{spq} = \| \alpha_{0.} \|_{l^p} + \left(\sum_{j \geq 0} (2^{j(s+1/2-1/p)}) \| \beta_{j.} \|_{l^p})^q \right)^{1/q} < \infty$$

Cette représentation des espaces de Besov en terme de coefficients d'ondelettes permet d'établir les relations entre les différents espaces de Besov. Ces différentes inclusions sont explicitées dans [HKPT98]

Les bases d'ondelettes fournissent une représentation des fonctions où l'information est concentrée en peu de coefficients. Ainsi peu de gros coefficients possèdent l'information du signal. Estimer ces quelques coefficients représentatifs fournit une bonne estimation de la fonction alors que les coefficients les plus petits, c'est-à-dire au dessous d'un certain niveau, sont mis arbitrairement à zéro. Cette mise à zéro des coefficients peut se faire de différentes façons: nous décrivons les deux estimateurs les plus classiques, l'estimateur par seuillage dur et l'estimateur par seuillage doux.

Definition 1.2.4. *Soit un seuil $\lambda > 0$ et considérons un estimateur empirique des coefficients d'ondelettes $\hat{\beta}_{jk}$, j, k.*
L'estimateur par seuillage dur est défini ainsi:

$$\tilde{f}_n = \sum_{j,k} 1_{|\hat{\beta}_{jk}| \geq \lambda} \hat{\beta}_{jk} \psi_{jk} \tag{1.2.1}$$

L'estimateur par seuillage doux est défini par:

$$\tilde{f}_n = \sum_{j,k} \mathrm{sgn}(\hat{\beta}_{jk})(|\hat{\beta}_{jk}| - \lambda)_+ \psi_{jk} \tag{1.2.2}$$

Ces estimateurs ont été étudiés par D. Donoho, I. Johnstone, G. Kerkyacharyan et D. Picard dans leurs travaux [KP93], [DJ94], [DJKP95], [DJKP96a], [DJKP96b], [DJKP97], [HKP98], [HKP99], P. Hall et P. Patil dans [HP95], [HP96] ou B. Delyon et A. Juditsky dans [DJ96a].
Les coefficients d'ondelettes sont calculés en pratique au moyen de la transformée en ondelettes discréte. Lorsque les points où est observée la fonction sont équidistants ($t_i = \frac{i}{n}$), les coefficients discrets w_{jk} sont reliés aux coefficients théoriques β_{jk} par la relation

$$w_{jk} \approx \sqrt{n} \beta_{jk}.$$

Le facteur $\sqrt{n}$ provient de la différence entre les conditions d'orthogonalité par rapport à la mesure empirique et à la mesure théorique.

Dans notre travail, nous relions ces estimateurs avec la théorie des M-estimateurs pénalisés. En effet, l'estimateur construit en minimisant une fonction de perte quadratique et une pénalité l^1 permet de retrouver l'estimateur par seuillage doux. Ses propriétés asymptotiques,

ainsi que le choix optimal du paramètre de seuillage découlent des propriétés entropiques des espaces de Besov comme nous le précisons dans la Section 2.2.
L'estimateur par seuillage dur peut, quant à lui, être obtenu en minimisant une perte quadratique et une pénalité portant sur le nombre de coefficients non nuls dans la décomposition en ondelettes comme nous le signalons dans la partie 3.2.

Tout au long du Chapitre 3, nous considérons des pénalités portant sur des combinaisons de coefficients d'ondelettes. Nous y montrons dans des cas précis que le problème de minimisation possède bien une solution dont nous étudions le comportement asymptotique. Plus précisément, nous construisons l'estimateur d'une fonction $\theta_0 \in B^s_{22}$:

$$\tilde{\theta}_n = \arg \min_{\beta_{jk}, j=j_0,\dots,j_1, k} \left(||\hat{\alpha}_{jk} - \beta_{jk}||_2^2 + \lambda_n^2 \sum_{j \geq j_0} \sum_k 2^{2js} \beta_{jk}^2 \right) \tag{1.2.3}$$

Cet estimateur minimise une perte quadratique ainsi qu'une pénalité de type B^s_{22}. Nous montrerons que l'estimateur lissé ainsi obtenu, si les niveaux de résolution j_0 et j_1 satisfont à quelques conditions, atteint asymptotiquement la vitesse optimale pour des pertes B^σ_{22} avec $\sigma < s$. Le théorème est le suivant:

Theorem 1.2.5. *L'estimateur des moindres carrés pénalisé par une norme de Besov B^s_{22} est tel que, sous les conditions suivantes: $\lambda_n^2 = n^{-\frac{2s}{1+2s}}$ et $j_0 = O(1)$, $j_1 = n^{\frac{1}{2s+1}}$, et pour une constante strictement positive C et $0 \leq \sigma < s_0$:*

$$E||\hat{\theta} - \theta||^2_{B^\sigma_{22}} \leq Cn^{-\frac{2(s-\sigma)}{2s+1}}.$$

Le choix optimal du paramètre est donné par $\lambda_n^2 = n^{-\frac{2s}{2s+1}}$. Or, en pratique, la constante s est inconnue. C'est pourquoi, par une méthode de validation croisée, nous déterminons un critère pour choisir au mieux le paramètre de lissage. Toutefois, cette méthode présente ne fournit pas de vitesse de convergence de l'estimateur pénalisé. C'est la raison pour laquelle nous nous sommes placés, dans la suite de notre travail, dans un cadre Bayésien.

1.3 Estimation Bayésienne et Pénalités

Dans la partie IV, nous nous sommes intéressé à une alternative aux méthodes d'estimation adaptative par seuillage, directes ou au moyen de pénalités l^1. Nous utilisons des pénalités, faisant certes intervenir la régularité de la fonction à estimer, mais nous considérons que ce paramètre est la réalisation d'une variable aléatoire. Cette approche Bayésienne, proche du point de vue de sélection de modèles permet de laisser les données choisir d'elles-mêmes (i.e a posteriori) le bon paramètre de régularité et de construire un estimateur atteignant asymptotiquement la vitesse minimax.
Plus précisément, le cadre de notre étude est le suivant: nous observons toujours un modl̀e de régression:

$$Y_i = \theta_0(z_i) + \epsilon_i, \quad i = 1, \dots, n$$

avec des erreurs $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$ indépendantes identiquement distribuées et un paramètre d'intérêt $\theta_0 \in L^2(\mu)$. Considérons une base orthonormée (ψ_j) et décomposons la fonction sur cette base $\theta = \sum_j \alpha_j \psi_j$. Par transformation orthonormale des données le problème s'écrit

$$d_j = \alpha_j + W_j, \quad j = 1, \ldots, J$$

où $W_j \sim \mathcal{N}(0, \frac{\sigma^2}{n})$. Le cadre Bayésien consiste à considérer α_j comme une réalisation d'une variable aléatoire inconnue. Pour cela, nous définissons une loi a priori π suivie par le paramètre α, et nous cherchons à évaluer la loi suivie par les données sachant le paramètre. Dès lors, le théorème de Bayes reliant les différentes lois du modèle

$$\mathcal{L}(\alpha|d) = \frac{\mathcal{L}(d|\alpha)\pi(\alpha)}{\mathcal{L}(d)}$$

permet de calculer la loi a posteriori suivie par le paramètre à estimer, α. La connaissance de la loi a posteriori, fonction des observations d_j, $j = 1, \ldots, J$ permet d'estimer les coefficients $(\hat{\alpha}_j)$ à partir soit du mode a posteriori, soit de la moyenne a posteriori, ou d'un quantile de la loi a posteriori. Un estimateur de la fonction initiale θ_0 est alors donné par $\hat{\theta} = \sum_{j=1}^{J} \hat{\alpha}_j \psi_j$.
Le choix de la loi a priori est donc fondamental pour obtenir de bonnes propriétés de l'estimateur. Des travaux récents modélisent les propriétés particulières des coefficients d'ondelettes. En effet, peu de coefficients représentent une grande quantité d'information et le nombre de coefficients non nuls tend vers zéro au fur et à mesure que le niveau de résolution augmente. Cette information peut être incorporée dans le choix de la loi a priori pour modéliser un tel comportement. Abramovich, Sapatinas et Silverman in [ASS98] considèrent l'a priori suivant:

$$\begin{aligned} w_{jk} &\sim (1-\pi_j)\delta_0 + \pi_j \mathcal{N}(0, \tau_j^2) \\ \tau_j^2 &= c_1 2^{-\alpha j} \\ \pi_j &= \min(1, c_2 2^{-\beta j}) \end{aligned}$$

Ce modèle apparait comme le modèle limite de celui proposé par Chipman in [CW99] ou Ruggeri aet Vidakovic in [RV99]

$$\begin{aligned} w_{jk}|\gamma_{jk} &\sim \gamma_{jk} N(0, c_j^2 \tau_j^2) + (1-\gamma_{jk}) N(0, \tau_j^2) \\ \gamma_{jk} &\sim \mathrm{Bern(p_j)} \\ c_j^2 &>> 1. \end{aligned}$$

Les coefficients suivent une loi a priori telle que les coefficients non nuls se raréfient lorsque j augmente.
Si ces modèles permettent de construire des estimateurs donnant de bons résultats en pratique, leur convergence n'a pas encore été démontrée. Très récemment, I. Johnstone et B. Silverman in [JS01] ont donné un début de réponse. C'est pourquoi dans notre travail nous considérons le modèle plus simple suivant

$$X_i = \theta_i + W_i, \; i = 1, \ldots, n \tag{1.3.1}$$

où les W_i sont des variables Gaussiennes indépendantes. Nous supposons en outre qu'il existe un paramètre de régularité s_0 tel que $\sum_{i=1}^n i^{2s_0}\theta_i^2 < \infty$ et qui appartient à un ensemble fini d'indices $\mathcal{S} = \{s_m,\ m \in \mathcal{M} \subset \mathbb{Z}\}$. A chaque s nous associons le modèle correspondant

$$\Theta(s) = \{\theta \in l^2,\ \sum_{i=1}^{\infty} i^{2s}\theta_i^2 < \infty\}$$

$\Theta = \cup_{s\in\mathcal{S}}\Theta(s)$ désigne l'ensemble de tous les modèles. A s connu, le paramètre suit la loi a priori suivante

$$\theta_i \sim \mathcal{N}\left(0, \lambda_n^2 i^{-2s}\right)$$

où λ_n^2 est choisi convenablement

$$\lambda_n^2 = \lambda_n^2(s) = n^{-\frac{1}{1+2s}}$$

La variable s devient un hyperparamètre du modèle et suit une loi a priori $q(s),\ s \in \mathcal{S}$. Ainsi le modèle est déterminé par les deux inconnues $(\theta, s_0) \in \Theta \times \mathcal{S}$. L'estimateur que nous considérons est l'estimateur le plus vraisemblable pour la loi a posteriori, c'est-à-dire défini par

$$(\hat{\theta}_n, \hat{s}_n) = \arg\max_{\Theta\times\mathcal{S}} \log p(\theta, s\,|X) \tag{1.3.2}$$

Après avoir montré les liens qui existent entre les estimateurs Bayésiens du mode et les M-estimateurs pénalisés, nous prouvons que, pour un choix convenable de l'a priori q, cet estimateur fournit un estimateur adaptatif de θ qui converge à la vitesse minimax. En étendant les résultats de Belitser et Ghosal in [BG00], nous montrons que cette méthode revient à étudier un estimateur pénalisé qui choisit des modèles au moins plus régulier que le vrai modèle $\Theta(s_0)$. En effet nous obtenons le lemme de sélection de modèles suivant qui décrit le comportement de $\hat{s}_n$:

Lemma 1.3.1. *Pour $s < s_0$, il existent des constantes strictement positives c_1, c_2 et un entier N tels que pour tout $n \geq N$*

$$\mathbf{P}(\hat{s}_n = s) \leq \exp(-c^2 n^{\frac{2s_0}{2s_0+1}\frac{1}{s_0+s}}) \tag{1.3.3}$$

$$\mathbf{P}(\hat{s}_n < s_0) \overset{n\to\infty}{\longrightarrow} 0. \tag{1.3.4}$$

La convergence de l'estimateur $\hat{\theta}_n$ en découle:

Theorem 1.3.2. *L'estimateur maximisant la log-vraisemblance a posteriori $\hat{\theta}_n$ vérifie la propriété suivante: il existe une constante C telle que:*

$$\mathbf{E}||\hat{\theta}(\hat{s}_n) - \theta_0||_n^2 \leq Cn^{-\frac{2s_0}{2s_0+1}} \tag{1.3.5}$$

Nous étendons ces résultats au cas particulier où les coefficients sont des coefficients d'ondelettes et étudions si ces méthodes d'estimation peuvent être transposées pour estimer des fonctions analytiques.

1.4 Formalisme Multifractal et Estimation

Dans cette dernière partie, nous nous proposons d'étendre les méthodes d'estimation Bayésiennes à des fonctions non régulières. De nombreux signaux présentent en effet un comportement très irrégulier, qui, dans les pires cas, suit des régimes différents. Un des exemples les plus frappants de tels signaux est donné par l'étude de la vitesse d'écoulement d'un fluide dans des zones de turbulence, cf l'étude réalisée par A. Arnéodo, E. Bacry et J. F. Muzy in [BAF$^+$91] ou [ABM99], ou bien U. Frisch in [Fri95]. L'analyse multifractale se propose de décrire de telles fonctions aux variations très rapides. Elle a connu ses premiers développements en probabilités (voir par exemple Brown et al. in [BMP92]). Elle a été introduite pour fournir des modèles statistiques en turbulence et a servi à modéliser des données financières (in [Man97]) ou des données sur un réseau informatique (in [RCRB99]).
La quantité qui convient pour décrire l'instabilité d'un signal, est son spectre de singularités $d_f(h)$ défini à partir de la notion de régularité ponctuelle.

Definition 1.4.1. *Pour toute fonction f définissons l'exposant de Hölder local en un point x_0 comme*

$$h_f(x_0) = \sup_{\alpha}\{\alpha,\ f \in \mathcal{C}^{\alpha}(x_0)\} \tag{1.4.1}$$

où $f \in \mathcal{C}^{\alpha}(x_0)$ s'il existe un polynôme P de degré inférieur ou égal à α tel que

$$|f(x) - P(x - x_0)| \leq c|x - x_0|^{\alpha}.$$

Le spectre de singularité d'une fonction f est alors défini comme une fonction de h, $d_f(h)$, repésentant la dimension de Haussdorf de l'ensemble $A_h = \{x,\ f \in C^h(x)\}$.

J.M Aubry et S. Jaffard in [Jaf00b] ou [Jaf00a] ont montré que cette quantité est reliée aux propriétés des coefficients d'ondelettes de séries aléatoires générées de la façon suivante: à chaque niveau j, les 2^j coefficients w_{jk} suivent une loi de probabilité ρ_j vérifiant quelques conditions techniques. Définissons

$$\rho(\alpha) = \lim_{\epsilon \to 0} \limsup\nolimits_{j \to \infty} \frac{\log_2(2^j \rho_j[\alpha - \epsilon, \alpha + \epsilon])}{j} \tag{1.4.2}$$

$$N_j(\alpha) = \#\{|w_{jk}| \geq 2^{-\alpha j}\} \tag{1.4.3}$$

$$\tilde{\rho}(\alpha) = \inf_{\epsilon > 0} \limsup\nolimits_{j \to \infty} \frac{\log_2(N_j(\alpha + \epsilon) - N_j(\alpha - \epsilon))}{j} \tag{1.4.4}$$

Les auteurs précédents ont montré que, sous certaines hypothèses, $\rho(\alpha) = \tilde{\rho}(\alpha)$ et si la fonction f se décompose en $f = \sum_{jk} w_{jk}\psi_{jk}$ alors

$$d_f(h) = h \sup_{\alpha \in [0,h]} \frac{\rho(\alpha)}{\alpha} \tag{1.4.5}$$

Dans le chapitre V, qui retranscrit un article coécrit avec F. Gamboa, nous estimons une fonction multifractale, exprimée dans une base d'ondelettes et observée dans un modèle de bruit blanc. Ces fonctions sont caractérisées par l'histogramme des coefficients du développement dans la base d'ondelettes. Il dépend de deux coefficients: η un coefficient de lacunarité et α un coefficient d'intensité. A chaque niveau j, un tel signal comporte $2^{\eta j}$ coefficients prenant la valeur $2^{-\alpha j}$ alors que les autres prennent la valeur zéro. Dans un premier temps, nous supposerons ces valeurs connues et nous montrerons que les problèmes d'estimation dans ce modèle rejoignent les problèmes de classification dans des modèles de mélange et nous donnerons des vitesses de convergence. La connaissance des caractéristiques de la structure du signal permet de reconstruire la fonction observée avec une vitesse de convergence exponentielle comme le montre le théorème suivant:

Theorem 1.4.2. *Supposons que nous observons une fonction multifractale*

$$f^* = \sum_j \sum_{k=0}^{2^j-1} w^*_{jk}\psi_{jk}$$

dont les coefficients sont tirés suivant la loi décrite précédemment, loi déterminée par les paramètres η et α vérifiant $1-2\alpha > 0$. Soit Π_1 la projection sur l'espace V_{j_1} où $2^{j_1} = n$ le nombre d'observations. Alors l'estimateur du mode $\hat{f}_n$ est tel qu'il existent deux constantes positives c et c_1 vérifiant:

$$\mathbf{E}\|\Pi_1 f^* - \hat{f}_n\|_2^2 \leq c_1 \exp(-c^2 n^{1-2\alpha}) n^{\eta+1-2\alpha} \tag{1.4.6}$$

Naturellement, en pratique les caractéristiques du signal ne sont pas observées. C'est pourquoi, dans un second temps, nous estimons ces paramètres par une approche pratique fondée sur le maximum de vraisemblance et l'utilisation d'un algorithme EM (cf [DLR77]), bien adapté à la résolution de tels problèmes. Toutefois, l'algorithme EM n'est pas appliqué sur l'ensemble des données mais de façon récursive sur des lignes de données pour j fixé. En outre il ne fournit pas de vitesse de convergence vers les vrais paramètres. C'est pourquoi, nous donnons des perspectives plus théoriques pour estimer ces paramètres en étudiant la distribution asymptotique des estimateurs empiriques de moments définis par:

$$\hat{\alpha}_n = \frac{1}{j_1 \log 2}\left(\log\left[\frac{\sum_{j\leq j_1}\sum_k d_{jk}}{\sum_{j\leq j_1}\sum_k d_{jk}^2 - \sigma^2}\right]\right)$$
$$\hat{\eta}_n = \alpha + \frac{1}{j_1 \log 2}\log\sum_{j=1}^{j_1}\sum_{k=0}^{2^j-1} d_{jk}$$

oles d_{jk} désignent les coefficients réellement observés. Ces résultats font actuellement l'objet d'une étude conjointe avec F. Gamboa de l'université de Toulouse Paul-Sabatier. Nous étendons ces résultats à des mesures non Gaussiennes in [GL01].

Chapter 2

Penalized M-estimation

2.1 General M-estimation

In this section, we give a general theorem that describes the asymptotic behavior of penalized M-estimators for a regression model in a large class of cases. Our model is the following:

$$\begin{cases} Y_i = \theta_0(z_i) + W_i, \ i = 1, \dots, n \\ \theta_0 \in \Theta \end{cases}$$

where $W_i, i = 1, \dots, n$ are observation errors satisfying some assumptions, we will make precise later on. Penalized M-estimators are defined as the solution of following the minimization problem

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \left(\gamma(Y, \theta) + \lambda_n^2 I(\theta) \right) \tag{2.1.1}$$

where $\gamma : \theta \in \Theta \to \gamma(Y, \theta) = \gamma_\theta(Y)$ is a convex loss function, Θ a functional set characterized by some regularity index s, $I : \Theta \to \mathbb{R}^+$ a penalty such that $\forall \theta \in \Theta$, $I(\theta) < \infty$ and λ_n^2 a smoothing parameter. Using a penalty is a way of adding a control over the regularity of the estimator, which prevents too rough estimators. The contribution between the two terms is balanced by the choice of the smoothing parameter that will play a key role. We want to be able to handle a large variety of situations depending on the choice of the loss-function, of the set and of the penalty. That is the reason why we do not want to use explicit expression of the estimators but, mainly, the geometric properties of the set where the minimization occurs. In many cases, the minimization program can not be solved directly and we do not have access to the true expression of the minimizer.

First, we will give a general theorem using entropy calculations and empirical process theory, then we will go further in studying the important case where the penalty is an l^1 norm and the estimator is a penalized least-absolute and a penalized least-square estimator. With this particular choice we will be able to find sharper rates of convergence for adaptive estimators. To conclude, we will adapt this result to other classical M-estimators and describe how such an estimating procedure can be implemented.

2.1.1 General theorem for penalized M-estimators

The theorem aims at handling various different cases. That is the reason why there is no optimality in the choice of constants, and particular methods can be used to improve the results as will be shown in the following parts of this work, see section 2.2. The first theorem is for quadratic loss which gives rise to penalized least squares estimators. The second theorem is about penalized M-estimators with general loss function that fulfills a specific assumption. Define the empirical measure:

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{z_i} \tag{2.1.2}$$

and the associated empirical norm

$$||\theta||_n^2 = ||\theta||_{P_n}^2 = \frac{1}{n} \sum_{i=1}^{n} \theta^2(z_i) \tag{2.1.3}$$

as well as the empirical scalar product $(W, \theta)_n = \frac{1}{n}\sum_{i=1}^n W_i\theta(z_i)$. We make the following additional assumptions over the model: there exist finite a constant K, such that

$$\begin{aligned}
&\exists R < \infty,\ \forall \theta \in \Theta,\ \sup_n \|\theta\|_n \leq R \\
&I(\theta_0) < \infty \qquad\qquad (2.1.4)\\
&\sup_n \sup_{i=1,\ldots,n} K^2(E\exp(\frac{W_i^2}{K^2}) - 1) \leq \sigma^2
\end{aligned}$$

So the errors measures W_i, $i = 1, \ldots, n$ are sub-Gaussian variables.

Theorem 2.1.1. *Theorem for quadratic loss. Under the assumption* (2.1.4)*, suppose that for some fixed constants $\eta > 0$, $A > 0$ and $s \geq \frac{1}{2}$, one has*

$$H(\delta, \{\theta \in \Theta:\ I(\theta) \leq M,\ \|\theta - \theta_0\|_n \leq \eta\}) \leq A\left(\frac{M}{\delta}\right)^{\frac{1}{s}},$$

$$for\ all\ \delta > 0,\ n \geq 1,$$

for all $M \geq M_n$, where $M_n \geq I(\theta_0)$. Then for

$$\lambda_n^{-1} = \begin{cases} O_{\mathbf{P}}(n^{\frac{s}{2s+1}} M_n^{\frac{p}{2} - \frac{1}{2s+1}}), & if\ s > \frac{1}{2}, \\ O_{\mathbf{P}}(n^{\frac{1}{4}}) M_n^{\frac{p}{2} - \frac{1}{2}} (\log n)^{-\frac{1}{2}}, & if\ s = \frac{1}{2}, \end{cases}$$

the Penalized Least-Square Estimator $\hat{\theta}_n$ satisfies

$$\|\hat{\theta}_n - \theta_0\|_n = O_{\mathbf{P}}(\lambda_n) M_n^{\frac{p}{2} - \frac{1}{2s+1}}.$$

Proof. The proof of this theorem can be found in [vdG00]. In the appendix, we recall briefly the main ideas. □

The following theorem gives asymptotic rates of convergence for penalized M-estimators with losses satisfying the Lipschitz condition:

$$\forall x_1,\ x_2 \in \mathbb{R},\ |\gamma_\theta(x_1) - \gamma_\theta(x_2)| \leq |x_1 - x_2| \qquad (2.1.5)$$

Theorem 2.1.2. *Theorem for general robust losses. Define $P^{(i)}$ the distribution of $X_i = (Y_i, z_i)$ and $\bar{P} = \frac{1}{n}\sum_{i=1}^n P^{(i)}$. Let for $0 \leq t \leq 1$, $\theta_t = t\theta_0 + (1-t)\theta$.*
Define $\gamma_\theta(X_i) = \gamma(Y_i - \theta(z_i))$. We make the assumption that there exist a positive constant ϵ and $0 \leq t_0 \leq 1$ such that, for $t \leq t_0$ we have:

$$\int (\gamma_{\theta_t} - \gamma_{\theta_0})\, d\bar{P} \geq \epsilon\|\theta_t - \theta_0\|_n^2 \qquad (2.1.6)$$

where ϵ is a strictly positive constant. Under the assumptions (2.1.4) *and* (2.1.5)*, suppose that for some fixed constants $\eta > 0$, $A > 0$ and $s \geq \frac{1}{2}$, one has*

$$H(\delta, \{\theta \in \Theta : \ I(\theta) \leq M, \ \|\theta - \theta_0\|_n \leq \eta\}) \leq A \left(\frac{M}{\delta}\right)^{\frac{1}{s}} \tag{2.1.7}$$

$$for \ all \ \delta > 0, \ n \geq 1,$$

for all $M \geq M_n$, where $M_n \geq I(\theta_0)$. Then for

$$\lambda_n^{-1} = \begin{cases} O_{\mathbf{P}}(n^{\frac{s}{2s+1}} M_n^{\frac{p}{2} - \frac{1}{2s+1}}), & if \ s > \frac{1}{2}, \\ O_{\mathbf{P}}(n^{\frac{1}{4}}) M_n^{\frac{p}{2} - \frac{1}{2}} (\log n)^{-\frac{1}{2}}, & if \ s = \frac{1}{2}, \end{cases}$$

we have

$$\|\hat{\theta}_n - \theta_0\|_n = O_{\mathbf{P}}(\lambda_n) M_n^{\frac{p}{2} - \frac{1}{2s+1}}.$$

For example, for a function $\theta : \ [0,1] \to \mathbb{R}$ of bounded variation, i.e such that for

$$I(\theta) = \mathrm{TV}(\theta) = \sum_{i=1}^{n} |\theta(z_i) - \theta(z_{i-1})| < \infty$$

S. van de Geer in [vdG00] proved that the assumptions of the theorem are fulfilled. Then the penalized least squares estimator $\hat{\theta}_n$ is consistent and for $I(\theta_0) > 0$:

$$\begin{cases} \|\hat{\theta}_n - \theta_0\|_n & = O_{\mathbf{P}}(\lambda_n I^{\frac{1}{2}}(\theta_0) \vee n^{-\frac{1}{2}}) \\ \lambda_n^{-1} & = O_{\mathbf{P}}(n^{\frac{1}{3}}) I^{\frac{1}{6}}(\theta_0) \end{cases} \tag{2.1.8}$$

Another example of the consequences the general theorem is given in section 2.4.5.

Proof. The proof of these theorems relies on the behavior of a particular empirical process in a neighborhood of the true parameter of the following form $\Lambda = \{\theta \in \Theta, \ \|\theta - \theta_0\| \leq 1, \ I(\theta) \leq M\}$. Once we have been able to write a concentration inequality for the empirical process, using the mere definition of the M-estimator and this inequality, the conclusion follows from explicit majorations. □

In a preliminary approach, we study the concentration inequality we use for penalized least squares estimation in the case of independent identically distributed errors and show that it can be extended to the Martingale case. Once this result has been proved, consistency of penalized least squares estimators for Martingale increments errors will follow.

Set a roughness parameter $\alpha = \frac{1}{s}$. First we establish a concentration inequality for the weighted empirical process

$$\nu_n : \ \theta \to \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i \theta(z_i)}{\|\theta\|_n^{1-\frac{\alpha}{2}} I(\theta)^{\frac{\alpha}{2}}}$$

where the random variables $W_1, \ldots, W_n$ are sub-Gaussian centered variables and the penalty I is minorated. We will start by an independent version and extend it to the case where the random variables are martingale increments.

Independent Version

Assume that the errors $W_1, \dots, W_n$ are independent random variables. The following proposition provides a concentration inequality over the supremum over $\Theta(M) \subset \Theta$ of an empirical that appears naturally in M-estimation.

Proproposition 2.1.3. *Let $\Theta(M) = \{\theta \in \Theta,\ I(\theta) \leq M\} \subset \Theta$ a function space such that there exists a constant A_0 such that for every $\delta > 0$*

$$\int_0^\delta H^{1/2}(u, \Theta(M), P_n)\, du \leq A_0 \delta^{1-\frac{\alpha}{2}} M^{\frac{\alpha}{2}}.$$

If $I : \Theta \longrightarrow [1, +\infty[$ then for sub-Gaussian independent random variables $W_i, i = 1, \dots, n$, there exists a constant K such that we have:

$$\mathbf{P}\left(\sup_{\theta\in\Theta} \frac{|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i(\theta(z_i) - \theta_0(z_i))|}{||\theta - \theta_0||_n^{1-\frac{\alpha}{2}} I^{\frac{\alpha}{2}}(\theta)} \geq T\right) \leq K \exp(-\frac{T^2}{K^2}).$$

The proof of this theorem relies on empirical process theory developed by Pollard in [Pol84]. Using a basic concentration inequality for empirical process, we extend it to a more complicated case using entropy considerations and a trick due to K. Alexander in [Ale85] and developed by S. van de Geer in [vdGW96], the peeling device. It enables us to split a concentration inequality of the supremum over a complicated set into several inequalities of the supremum over smaller sets.

Definition 2.1.4. *Peeling-device: For Λ a function space, let $\tau : \Lambda \longrightarrow [\rho, R]$ an increasing function where $\rho > 0$ and $R \leq \infty$. Let the sequence $(m_s)_{s=0}^S$ with $m_0 = \rho$ and $m_S = R$ with $R \leq \infty$, then we can peel the space Λ into smaller sets $\Lambda_s = \{\theta \in \Lambda, m_{s-1} \leq \tau(\theta) < m_s\}$ and we have*

$$\Lambda = \cup_{s\geq 1}\Lambda_s.$$

Now we can use majoration over the smaller sets and for this write for an empirical process $Z_n(\theta)$ and for a positive a:

$$\begin{aligned}\mathbf{P}\left(\sup_{\theta\in\Lambda} \frac{|Z_n(\theta)|}{\tau(\theta)} > a\right) &\leq \sum_{s=1}^S \mathbf{P}\left(\sup_{\theta\in\Lambda_s} \frac{|Z_n(\theta)|}{\tau(\theta)} > a\right) \\ &\leq \sum_{s=1}^S \mathbf{P}\left(\sup_{\theta\in\Lambda_s} |Z_n(\theta)| > a m_{s-1}\right).\end{aligned}$$

We can derive inequalities over the weighted empirical process $\frac{Z_n(\theta)}{\tau(\theta)}$, from inequalities over the unweighted empirical process $Z_n(\theta)$ over the sets Λ_s.

In order to majorate the following quantity:

$$\mathbf{P}\left(\sup_{\theta\in\Theta} \frac{|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i(\theta(z_i) - \theta_0(z_i))|}{||\theta - \theta_0||_n^{1-\frac{\alpha}{2}} I^{\frac{\alpha}{2}}(\theta)} \geq T\right)$$

where $I : \Theta \longrightarrow [1, +\infty[$ we use the peeling-device and consider the subsets $\Theta_k = \{\theta, 2^k \leq I(\theta) < 2^{k+1}\}$, $k \geq 0$ forming a partition of the global space: $\Theta = \cup_{k\geq 0}\Theta_k$. Therefore we have:

$$\mathbf{P}\left(\sup_{\theta\in\Theta} \frac{|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i(\theta(z_i) - \theta_0(z_i))|}{||\theta-\theta_0||_n^{1-\frac{\alpha}{2}} I^{\frac{\alpha}{2}}(\theta)} \geq T\right)$$

$$\leq \mathbf{P}\left(\sup_{k}\sup_{\theta\in\Theta_k} \frac{|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i(\theta(z_i) - \theta_0(z_i))|}{||\theta-\theta_0||_n^{1-\frac{\alpha}{2}} I^{\frac{\alpha}{2}}(\theta)} \geq T\right)$$

$$\leq \sum_{k\geq 0}\mathbf{P}\left(\sup_{\theta\in\Theta_k} \frac{|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i(\theta(z_i) - \theta_0(z_i))|}{||\theta-\theta_0||_n^{1-\frac{\alpha}{2}I(\theta)^{\frac{\alpha}{2}}}} \geq T\right)$$

$$\leq \sum_{k\geq 0}\mathbf{P}\left(\sup_{\theta\in\Theta_k} \frac{|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i(\theta(z_i) - \theta_0(z_i))|}{||\theta-\theta_0||_n^{1-\frac{\alpha}{2}}} \geq T2^{\frac{\alpha k}{2}}\right)$$

So we must study the empirical process

$$\frac{|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i(\theta(z_i) - \theta_0(z_i))|}{||\theta-\theta_0||_n^{1-\frac{\alpha}{2}}}, \ \forall\theta\in\Theta_k.$$

We recall that we have made the assumption that the functions are bounded in the set Θ in the empirical L^2 norm, i.e there exists a constant R such that:

$$\forall k, \ \sup_{\theta\in\Theta_k} ||\theta||_n \leq R.$$

Now we can apply the peeling-device once more but this time with the following decomposition:

$$\Theta_l = \cup_{l\geq 1}\{\theta\in\Theta_k, 2^{-l}R \leq ||\theta-\theta_0||_n \leq 2^{-l+1}R\}.$$

We now have:

$$\mathbf{P}\left(\sup_{\theta\in\Theta_k} \frac{|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i(\theta(z_i) - \theta_0(z_i))|}{||\theta-\theta_0||_n^{1-\frac{\alpha}{2}}} \geq T2^{\frac{\alpha k}{2}}\right) \leq$$

$$\mathbf{P}\left(\sup_{l\geq 1}\sup_{\theta\in\Theta_k, ||\theta-\theta_0||_n\leq 2^{-l+1}R} \frac{|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i(\theta(z_i) - \theta_0(z_i))|}{||\theta-\theta_0||_n^{1-\frac{\alpha}{2}}} \geq T2^{\frac{\alpha k}{2}}\right),$$

We set $T^{'} = T2^{\frac{\alpha k}{2}}$. If we choose properly some constants A_0, C, C_1 and under the following hypothesis

- The errors $W_i, i = 1, \ldots, n$ are sub-Gaussian.

$$\max_{i=1,\ldots,n} K^2(E\exp(\frac{|W_i^2|}{K^2}) - 1) \leq \sigma_0^2.$$

- The entropy of the class of functions is controlled for all $\delta > 0$ by:

$$\sqrt{n}\delta \geq 2C \left(\int_{\frac{\delta}{8\sigma}}^{R} H^{\frac{1}{2}}(u, \Theta_{kl}, P_n)\, du \right) 2^{\frac{\alpha k}{2}}$$

where
$$\Theta_{kl} = \{\theta \in \Theta, ||\theta - \theta_0||_n \leq R2^{-l+1}, I(\theta) \leq 2^{k+1}\}.$$

This conditions is verified for the stronger condition:

$$\int_0^\delta H^{1/2}(u, \Theta(M), P_n)\, du \leq A_0 \delta^{1-\frac{\alpha}{2}} M^{\frac{\alpha}{2}}$$

which is implied by the general assumptions over the model (2.1.4). The following lemma gives a majoration of this concentration inequality.

Lemma 2.1.5. *Under the assumptions* (2.1.4) *the following inequality holds*

$$\mathbf{P}\left(\sup_{\theta\in\Theta, ||\theta-\theta_0||_n \leq \delta} |\frac{1}{\sqrt{n}} \sum_{i=1}^n W_i(\theta(z_i) - \theta_0(z_i))| \geq 2CC_1 A_0 \delta^{1-\frac{\alpha}{2}} \right) \leq C \exp(-C_1^2 A_0^2 \delta^{-\alpha})$$

Proof. A demonstration of this result can be found in [vdG00]. □

Using this lemma together with previous inequality we have:

$$\begin{aligned}
&\mathbf{P}\left(\sup_{\theta\in\Theta} \frac{|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i(\theta(z_i) - \theta_0(z_i))|}{||\theta - \theta_0||_n^{1-\frac{\alpha}{2}}} \geq T' \right) \\
&\leq \sum_l \mathbf{P}\left(\sup_{\theta\in\Theta, ||\theta-\theta_0||_n \leq 2^{-l+1}R} |\frac{1}{\sqrt{n}} \sum_{i=1}^n W_i(\theta(z_i) - \theta_0(z_i))| \geq T'(2^{-l}R)^{1-\frac{\alpha}{2}} \right) \\
&\leq \sum_l C_1 \exp(-C_0^2 A_0^2 2^{(l+1)\alpha} R^{-\alpha}) \\
&\leq C^2 \exp(-\frac{(T')^2}{C^2}).
\end{aligned}$$

So if we apply this result to the first concentration inequality we find:

$$\begin{aligned}
&\mathbf{P}\left(\sup_{\theta\in\Theta} \frac{|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i(\theta(z_i) - \theta_0(z_i))|}{||\theta - \theta_0||_n^{1-\frac{\alpha}{2}} I(\theta)^{\frac{\alpha}{2}}} \geq T \right) \\
&\leq \sum_{k=1}^\infty C^2 \exp(\frac{-2^{\alpha k} T^2}{C^2}) \\
&\leq K \exp(\frac{-T^2}{K^2}).
\end{aligned}$$

which concludes the proof of the proposition (2.1.3).

We can see that the main condition imposed on the errors $W_i, i = 1, \dots, n$ is that they must be independent random variables satisfying a concentration inequality. As a matter of fact, the starting point of the demonstration of the lemma is a Hoeffding type inequality that requires such properties, but we could have written a similar proof starting with Bernstein inequality for martingales, inequalities that can be found in [LT91] for instance.

Martingale Version

In this section, we want to extend the previous theorem to the case where the errors are martingale increments. The preceding proof relies on the existence of a concentration inequality of the form

$$\mathbf{P}\left(\sup_{\theta\in\Theta, ||\theta-\theta_0||_n\leq\delta} |\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i(\theta(z_i)-\theta_0(z_i))| \geq 2CC_1A_0\delta^{1-\frac{\alpha}{2}}\right)$$

$$\leq C\exp(-C_1^2A_0^2\delta^{-\alpha}).$$

This result is a direct consequence of Hoeffding inequality. But if the the variables are dependent, this inequality does not hold any more and the following lemma replaces this inequality:

Lemma 2.1.6. *Let $S_n = \sum_{i=1}^{n} Z_i$, where the Z_i are martingale increments and define $\mathcal{F}_i = \sigma(Z_1, \dots, Z_n)$. Moreover suppose that for all $p \geq 1$ and $\forall i \in [1, n]$ there exist two constants C_0 and C_1 such that :*

$$E(|Z_i|^p|\mathcal{F}_{i-1}) \leq C_0C_1^p p^p \tag{2.1.9}$$

Then for all $t \geq 0$ and for $\hat{C}_0 = 4eC_0$ and $\hat{C}_1 = 2eC_1$:

$$\mathbf{P}(|S_n| \geq t) \leq 2\exp(-\frac{t^2}{2(\hat{C}_0C_1^2n + \hat{C}_1t)}).$$

Proof. The proof of this result can be found in many books, see for instance van der Vaart and Wellner in [vdVW96] but we give in the appendix an easy proof. □

Lemma 2.1.7. *If W_i, $i = 1, \dots, n$ are martingale increments satisfying condition (2.1.9), and if moreover for $a > 0$ the following bound holds:*

$$\sqrt{n}a \geq C\int_0^R H_B^{\frac{1}{2}}(u, \Theta, P_n)\,du$$

with C some (large) finite constant, there exist finite positive constants c, K_1 and K_2 such that we have a Bernstein-type majoration:

$$\mathbf{P}\left(\sup_{\theta\in\Theta}\frac{1}{n}|\sum_{i=1}^{n} W_i(\theta(z_i)-\theta_0(z_i))| \geq a\right) \leq c\exp\left(-\frac{na^2}{(K_1a + K_2R^2)}\right).$$

Proof. The proof of this result is similar to the proof of the inequality in the independent case, by replacing the Hoeffding inequality by the Bernstein-Martingale inequality, and using the chaining device and truncation device, explained in [vdG00]. We recall the main idea of the chaining device: for the set Θ, we consider families of covering-functions for different radius $2^{-s}R, s = 0, \dots, S$. So if we set $f_\theta^0 = 0$, then for a given S we can write

$$f_\theta = \sum_{s=1}^{S} (f_\theta^s - f_\theta^{s-1}) + (f_\theta - f_\theta^S).$$

The main idea is to take S large enough to obtain good approximation properties, and the remaining term $\sum_{s=1}^{S} (f_\theta^s - f_\theta^{s-1})$ can be handled easily since it only deals with a finite number of functions. □

So due to this lemma we can extend the theorem to the case where the errors $W_i, i = 1, \dots, n$ are martingale increments. Moreover we have proven the consistency of the least-square penalized estimator when the penalty is minorated.

Proproposition 2.1.8. *Let Θ a function space and $\Theta(M) = \{\theta \in \Theta,\ I(\theta) \leq M\}$ such that there exists a constant A_0 such that for every $\delta > 0$*

$$\int_0^\delta H_B^{1/2}(u, \Theta(M), P_n)\, du \leq A_0 \delta^{1-\frac{\alpha}{2}} M^{\frac{\alpha}{2}}.$$

If $I : \Theta \longrightarrow [1, +\infty[$ then for Martingale increments $W_i, i = 1, \dots, n$, there exists a constant K such that we have:

$$\mathbf{P}\left(\sup_{\theta \in \Theta} \frac{|\frac{1}{\sqrt{n}} \sum_{i=1}^n W_i(\theta(z_i) - \theta_0(z_i))|}{||\theta - \theta_0||_n^{1-\frac{\alpha}{2}} I^{\frac{\alpha}{2}}(\theta)} \geq T \right) \leq K \exp(-\frac{T^2}{K^2}).$$

Limiting case

The class of functions must satisfy the entropy condition:

$$\sqrt{n}\delta \geq 2C \left(\int_{\frac{\delta}{8\sigma}}^{R} H^{\frac{1}{2}}(u, \Theta_{kl}, P_n)\, du \right)$$

where

$$\Theta_{kl} = \{\theta \in \Lambda, ||\theta - \theta_0||_n \leq R2^{-l+1}, I(\theta) \leq 2^{k+1}\}.$$

This condition is fulfilled as soon as the integral of the square root of the entropy converges and has the following majoration:

$$\int_0^\delta H^{1/2}(u, \Theta_{kl}, P_n)\, du \leq A_0 \delta^{1-\frac{\alpha}{2}} 2^{\frac{(k+1)\alpha}{2}}.$$

But in the regression setting, we encounter the following condition over the entropy: there exists a constant C such that $H(\delta, \Theta, P_n) \leq C\delta^{-\frac{1}{s}}$, where $s \geq 2$ is a smoothness-coefficient.

But the case $s = 2$ corresponds to some interesting spaces and enables us to handle classical estimator such as the soft-thresholded estimator in a Besov space $B^s_{p,\infty}$, $s > \frac{1}{p}$. We can see that the entropy condition to apply the theorem can be weakened to $\log(\frac{8\sigma^2 R}{\delta}) \leq \sqrt{n}\delta$. Taking the log-term into account leads to theorem in the case where $s = \frac{1}{2}$.

Remark 2.1.9. *We can notice that this general theorem does not provide optimal constants due to the use of non optimal entropy majorations. Moreover, consistency results depend heavily on the choice of the smoothing parameter λ_n^2 which must tend towards zero, but not too quickly. This choice depends heavily on the concentration of the errors, but also on the smoothness of the class of functions Θ, in fact, it depends of s which is unknown in practice. As a result, to estimate a function without prior knowledge on it, we have two possibilities: either try to let the data pick the smoothing parameter by cross-validation or a Bayes approach, or change the penalty which will not use the smoothness index but will be still linked with the smoothness of the class. In the next section 2.2, we present adaptive estimation with l^1 penalty.*

2.1.2 Appendix

Proof of Theorem 2.1.1

Proof. By the definition of the estimator we have for all $\theta \in \Theta$:

$$||Y - \hat{\theta}_n||_n^2 + 2\lambda_n^2 I(\hat{\theta}_n) \leq ||Y - \theta||_n^2 + 2\lambda_n^2 I(\theta),$$

so for $\theta = \theta_0$ we can write

$$||\hat{\theta}_n - \theta_0||_n^2 + 2\lambda_n^2 I(\hat{\theta}_n) \leq 2(W, \hat{\theta}_n - \theta_0)_n + 2\lambda_n^2 I(\theta_0).$$

But using the concentration inequality over the empirical process we have

$$\sup_{\theta \in \Theta} \left(\frac{(W, \theta - \theta_0)_n}{||\theta - \theta_0||_n^{1-\frac{s}{2}} (I(\theta - \theta_0))^{\frac{s}{2}}} > T \right) = O_{\mathbf{P}}(n^{-\frac{1}{2}})$$

$$||\hat{\theta}_n - \theta_0||_n^2 + \lambda_n^2 I^p(\hat{\theta}_n) \leq O_{\mathbf{P}}(n^{-\frac{1}{2}}) ||\hat{\theta}_n - \theta_0||_n^{1-\frac{s}{2}} I^{\frac{s}{2}}(\hat{\theta}_n - \theta_0) + \lambda_n^2 I^p(\theta_0).$$

Solving this inequality in both cases $I(\hat{\theta}_n) > I(\theta_0)$ and $I(\hat{\theta}_n) \leq I(\theta_0)$ leads to the result of the theorem. □

Proof of Theorem 2.1.2

Proof. The proof is divided into five steps:

1. Write $\hat{\theta}_{t,n} = t\hat{\theta}_n + (1-t)\theta_0$ and take $t = \frac{t_0}{1+||\hat{\theta}_n - \theta_0||_n}$. Using convexity of the loss-function and the definition of the M-estimator we get:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \gamma(Y_i - \hat{\theta}_{n,t}(z_i)) + \lambda_n^2 I(\hat{\theta}_{n,t}) \\
&\leq t\left(\frac{1}{n}\sum_{i=1}^n \gamma(Y_i - \hat{\theta}_n(z_i)) + \lambda_n^2 I(\hat{\theta}_n)\right) + (1-t)\left(\frac{1}{n}\sum_{i=1}^n \gamma(Y_i - \theta_0(z_i)) + \lambda_n^2 I(\theta_0)\right) \\
&\leq \frac{1}{n}\sum_{i=1}^n \gamma(Y_i - \theta_0(z_i)) + \lambda_n^2 I(\theta_0)
\end{aligned}$$

So, the following inequality holds:

$$\int(\gamma_{\hat{\theta}_{n,t}} - \gamma_{\theta_0})d\bar{P} + \lambda_n^2 I(\hat{\theta}_{n,t}) \leq -\int(\gamma_{\hat{\theta}_{n,t}} - \gamma_{\theta_0})d(P_n - \bar{P}) + \lambda_n^2 I(\theta_0) \qquad (2.1.10)$$

2. Since $||\hat{\theta}_{n,t} - \theta_0||_n \leq t_0$, condition (2.1.6) is fullfilled so we have

$$\int(\gamma_{\hat{\theta}_{n,t}} - \gamma_{\theta_0})d\bar{P} \geq \epsilon||\hat{\theta}_{n,t} - \gamma_{\theta_0}||_n^2 \qquad (2.1.11)$$

(2.1.10) together with (2.1.11) gives the following majoration

$$\epsilon||\hat{\theta}_{n,t} - \gamma_\theta||_n^2 + \lambda_n^2 I(\hat{\theta}_{n,t}) \leq -\int(\gamma_{\hat{\theta}_{n,t}} - \gamma_{\theta_0})d(P_n - \bar{P}) + \lambda_n^2 I(\theta_0) \qquad (2.1.12)$$

3. Due to conditions (2.1.5) and the sub-Gaussian behavior of errors, the empirical process has sub-Gaussian tail behavior and Lemma 8.5 in [vdG00] apply: under the entropy condition (2.1.7) we have

$$\sup_{\theta,\, ||\theta-\theta_0||_n \leq 1} \frac{\left|\int(\gamma_\theta - \gamma_{\theta_0})d(P_n - \bar{P})\right|}{||\theta - \theta_0||_n^{1-\frac{1}{2s}}(1 + I(\theta_0) \vee I(\theta))^{\frac{1}{2s}}} = O_{\mathbf{P}}(n^{-\frac{1}{2}}) \qquad (2.1.13)$$

4. Now we find ourselves in the same situation than 2.1.1:

$$\epsilon||\hat{\theta}_{n,t} - \gamma_\theta||_n^2 + \lambda_n^2 I(\hat{\theta}_{n,t}) \leq O_{\mathbf{P}}(n^{-\frac{1}{2}})||\hat{\theta}_{n,t} - \theta_0||_n^{1-\frac{1}{2s}}(1 + I(\theta_0) \vee I(\hat{\theta}_{n,t}))^{\frac{1}{2s}} + \lambda_n^2 I(\theta_0) \quad (2.1.14)$$

5. Reasoning among the lines of 2.1.1 and using that $||\hat{\theta}_n - \theta_0||_n = \frac{1}{t}||\hat{\theta}_{n,t} - \theta_0||_n$ ends the proof.

□

Proof of Lemma 2.1.6

Proof. Let $0 < \lambda < \frac{1}{C_1 e}$, then

$$\sum_{p\geq 2} \lambda^p E(|z_i|^p|\mathcal{F}_{i-1})\frac{1}{p!} \leq C_0 C_1^2 \lambda^2 \sum_{p\geq 2} (\lambda C_1)^{p-2}\frac{p^p}{p!}.$$

But with the standard majoration for all $p \geq 2$,

$$p! \geq p^p e^{-p} e,$$

so we have the following inequality:

$$\begin{aligned} A &\leq C_0 C_1^2 e \lambda^2 \sum_{p\geq 2} (\lambda C_1 e)^{p-2} \\ &\leq \frac{C_0 C_1 e \lambda^2}{1-\lambda C_1 e}. \end{aligned}$$

So

$$\begin{aligned} E(e^{\lambda z_i}|\mathcal{F}_{i-1}) &\leq 1 + \sum_{p=0}^{+\infty} \lambda^p \frac{E(|z_i|^p|\mathcal{F}_{i-1})}{p!} \\ &\leq \exp\left(\frac{C_0 C_1 e \lambda^2}{1-\lambda C_1 e}\right). \end{aligned}$$

So finally

$$\begin{aligned} \mathbf{P}(S_n \geq t) &\leq e^{-\lambda t} E(e^{\lambda S_n}) \\ &\leq e^{-\lambda t} E(E(e^{\lambda S_n}|\mathcal{F}_{n-1})) \end{aligned}$$

By induction we condition by every filtration and choose a particular $\lambda_0 = \frac{t}{2(\hat{C}_0 C_1^2 n + \hat{C}_1 t)}$, and we get the following majoration:

$$\mathbf{P}(S_n \geq t) \leq \exp\left(-\frac{t^2}{2(\hat{C}_0 C_1^2 n + \hat{C}_1 t)}\right).$$

If we do the same for $-S_n$ the result comes obviously. □

2.2 Adaptive estimation in regression

Adaptive estimation in regression, using soft thresholding type penalties

Abstract We show that various regression estimators, such as the least squares estimator and the least absolute deviations estimator, can be made adaptive (up to logarithmic factors), by adding a soft thresholding type penalty to the loss function.

AMS 1991 subject classifications. Primary 62G05, secondary 62G20.

Key words and phrases. Adaptive estimation, empirical process, penalty, rate of convergence, regression, soft thresholding.

2.2.1 Introduction

Consider independent real-valued observations Y_i, with distribution depending on a location parameter $\theta_{i,0}$, $i = 1, \ldots, n$. This location parameter is defined by means of a given convex loss function $\gamma : \mathbf{R} \to \mathbf{R}$, namely, we suppose that $\mathbf{E}\gamma(Y_i - b)$ has a unique minimum in $b = \theta_{i,0}$, $i = 1, \ldots, n$. Hence, when $\gamma(\xi) = \xi^2$, then $\theta_{i,0}$ is the mean of Y_i, and when $\gamma(\xi) = |\xi|$, then $\theta_{i,0}$ is the median of Y_i, etc.

Let us write $\theta_{i,0} = \theta_0(z_i)$, where $z_i \in \mathcal{Z}$ is a covariable, $i = 1, \ldots, n$. The covariables $z_1, \ldots, z_n$ are introduced for ease of notation, although in fact, they often also have a “real life" interpretation. We now have the regression model

$$Y_i = \theta_0(z_i) + W_i, \; i = 1, \ldots, n, \tag{1.1}$$

where $W_1, \ldots, W_n$ may be considered as measurement error.

Let $\theta_0 \in \Theta$, with Θ a parameter space. We take this parameter space as (a subset of) the set of all real-valued functions on $\mathcal{Z}$. We shall study the asymptotic behaviour ($n \to \infty$) of estimators of θ_0. Throughout, as n varies, θ_0 as well as Θ are allowed to vary as well. In other words, we are actually considering triangular arrays of observations. However, to avoid too many indices, we will not always express dependence on n in our notation.

Take Q_n as the empirical measure of the covariables:

$$Q_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{z_i}.$$

We denote the $L_2(Q_n)$-norm of a function $\theta : \mathcal{Z} \to \mathbf{R}$ as

$$||\theta||_{Q_n} = (\int \theta^2 dQ_n)^{1/2}.$$

Now, let $\psi_1, \ldots, \psi_n$ be an orthonormal basis in $L_2(Q_n)$. Each function θ can be written as

$$\theta = \sum_{j=1}^{n} \alpha_j \psi_j,$$

in $L_2(Q_n)$. Moreover,

$$\|\theta\|^2_{Q_n} = \sum_{j=1}^{n} \alpha_j^2 = \|\alpha\|^2_n,$$

where $\alpha = (\alpha_1 \ldots, \alpha_n)'$, and where we we denote the Euclidean norm of a vector in $\mathbf{R}^n$ by $\| \cdot \|_n$.

We shall study the regression estimator with *soft thresholding type penalty*

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \left[\frac{1}{n} \sum_{i=1}^{n} \gamma(Y_i - \theta(z_i)) + \lambda_n^2 I_n(\theta) \right], \tag{1.2}$$

where for $\theta = \sum_{j=1}^{n} \alpha_j \psi_j$, we take the *penalty*

$$I_n(\theta) = \sum_{j=1}^{n} |\alpha_j|, \tag{1.3}$$

and where λ_n^2 is a (to be chosen) smoothing parameter.

The penalty $I_n(\theta)$ defined above is called a *soft thresholding* type penalty, because in the least squares (LS) case ($\gamma(\xi) = \xi^2$), the above estimator is in fact the standard soft thresholding estimator (see Subsection 3.2). Penalized M-estimators are common in the litterature, see for instance Baraud (2000), Birgé and Massart (2000), (1999), Köhler (1999), (2000) or Leurgans (1982). In particular, the least squares estimator (LSE) with soft thresholding is well studied, and very convenient from the computational point of view. It is however of interest to investigate other loss functions as well, because most theory for the least squares case depends on the assumption of Gaussian errors, or at least on the assumption of the existence of second moments of the errors.

Note also that LS method requires a smoothing parameter λ_n^2 which depends on the scale on which the observations are measured. It depends on (an estimator of) the variance of the errors. Robust regression estimators can do with a choice for λ_n^2 which works for all data. Unfortunately, for the robust methods, as yet we do not know the optimal (i.e., smallest) value for λ_n^2. This is due to our bounds for empirical processes, where not only universal constants, which have as yet not been calculated, appear, but also some (possibly redundant) logarithmic factors.

The paper is organized as follows. In the next section, we review some entropy results. Entropy is a measure of the complexity of a metric space, in our case the parameter space of regression functions. To illustrate the regression theory, and yet minimize the amount of approximation theory, we introduce a space of functions governed by a *roughness* parameter ρ, and calculate (a bound for) its entropy. We also discuss the relation with Besov spaces.

The rate of convergence of the "classical" LSE, that is, without penalty, follows from entropy calculations (see van de Geer (1990)). This result is cited in Theorem 3.1 in order to later on reveal the relation with adaptive estimation.

We then establish a rate of convergence for the LSE with soft thresholding. Here, we use a method of proof that does not need an explicit expression for the estimator, so that it is well tailored for transfer to other estimation methods. We show that the estimator indeed adapts (in a sense described there) to the amount of roughness ρ (see Example 3.2).

In Section 4, we present the extension to robust regression. Here, we need an inequality derived from empirical process theory.

Section 5 presents a small simulation study, where LS is compared to least absolute deviations (LAD), when the errors are double Exponential instead of Gaussian. The section ends with a discussion on adaptivity and robustness.

2.2.2 Entropy

Definition of entropy

Let T be a (subset) of a metric space. The *δ-covering number* $N(\delta, T)$ is the minimal number of balls with radius $\delta > 0$ necessary to cover T. The *δ-entropy* is $H(\delta, T) = \log N(\delta, T)$.

In our situation, we need entropies of subsets of $L_2(Q_n)$. Note that $L_2(Q_n)$ is essentially the n-dimensional Euclidean space $\mathbf{R}^n$. The $L_2(Q_n)$-norm of a function θ is the normalized Euclidean norm of the vector $(\theta(z_1), \ldots, \theta(z_n))'$. Also, for $\psi_1, \ldots, \psi_n$ an orthonormal basis in $L_2(Q_n)$,

$$\theta = \sum_{j=1}^{n} \alpha_j \psi_j, \ \|\theta\|_{Q_n} = \|\alpha\|_n$$

with $\| \cdot \|_n$ the Euclidean norm.

A roughness parameter

Let us illustrate the entropy concept for a simple case. Consider the set of functions $\theta = \sum_{j=1}^{n} \alpha_j \psi_j$, for which

$$\sum_{j=1}^{n} |\alpha_j|^{\rho} \leq 1,$$

for some $0 \leq \rho \leq 2$. We may think of ρ as a *roughness* parameter: if $\rho = 0$, we assume the convention $x^0 = 1$ if x is non zero and $0^0 = 0$. Throughout all this paper, for a given set A, we will use the notation $\#A$ for the cardinality of the set A. As a consequence we get

$$\sum_{j=1}^{n} |\alpha_j|^0 = \#\{\alpha_j, \ \alpha_j \neq 0\}$$

So the function θ may have at most 1 non-zero coefficient, whereas, on the other extreme, $\rho = 2$ only requires that θ is within the n-dimensional unit ball. We can point out that the sets $\{\alpha, \sum_{j=1}^{n} |\alpha_j|^p \leq 1\}$ increase for the inclusion as ρ becomes large. Thus, the smaller ρ the “smoother" θ will be. This is also reflected by the entropy calculation: the smaller ρ, the smaller the entropy. This is shown in Lemma (2.2.1) below (where we omit the two extreme but trivial cases $\rho = 0$ and $\rho = 2$).

Lemma 2.2.1. *Consider the following subset of* $\mathbf{R}^n$: $\mathcal{A}_n = \{\alpha = (\alpha_1, \ldots, \alpha_n)' : \sum_{j=1}^n |\alpha_j|^\rho \leq 1\}$, *with* $0 < \rho < 2$. *We have for some constant* A, *depending only on* ρ,

$$H(\delta, \mathcal{A}_n) \leq A\delta^{-\frac{2\rho}{2-\rho}} \left(\log n + \log \frac{1}{\delta}\right), \ \delta > 0. \tag{2.1}$$

Proof. Let $\epsilon = \delta^{\frac{2}{2-\rho}}$. Define for $\alpha \in \mathcal{A}_n$,

$$N_\alpha(\epsilon) = \#\{\alpha_j : \ |\alpha_j| > \epsilon\}.$$

Moreover, let

$$N(\epsilon) = \lfloor \epsilon^{-\rho} \rfloor,$$

where $\lfloor x \rfloor$ denotes the integer part of $x > 0$.

We have

$$\max_{\alpha \in \mathcal{A}_n} N_\alpha(\epsilon) \leq N(\epsilon).$$

It suffices to have a δ-approximation of the coefficients larger than ϵ, neglecting the other coefficients. That is, let $\alpha \in \mathcal{A}_n$, and suppose that for some $\bar{\alpha}$,

$$\sum_{|\alpha_j| > \epsilon} |\alpha_j - \bar{\alpha}_j|^2 \leq \delta^2.$$

In addition, suppose that $\bar{\alpha}_j = 0$ for all $|\alpha_j| \leq \epsilon$. Then we have

$$\|\alpha - \bar{\alpha}\|_n^2 \leq \delta^2 + \sum_{|\alpha_j| \leq \epsilon} |\alpha_j|^2 \leq 2\delta^2,$$

provided

$$\sum_{|\alpha_j| \leq \epsilon} |\alpha_j|^2 \leq \epsilon^{2-\rho}.$$

But this follows from

$$\begin{aligned}\sum_{|\alpha_j| \leq \epsilon} |\alpha_j|^2 &= \sum_{|\alpha_j| \leq \epsilon} |\alpha_j|^{2-\rho+\rho} \\ &\leq \sum_{|\alpha_j| \leq \epsilon} |\alpha_j|^\rho \epsilon^{2-\rho} \\ &\leq \epsilon^{2-\rho}.\end{aligned}$$

The number of ways to choose $N(\epsilon) \leq n$ coefficients out of n is

$$\binom{n}{N(\epsilon)} \leq n^{N(\epsilon)}.$$

Moreover, the δ-entropy of a space of dimension $N(\epsilon)$ is at most $5N(\epsilon)\log\frac{1}{\delta}$ (see e.g. van de Geer (2000)). So, we arrive at

$$H(\sqrt{2}\delta, \mathcal{A}_n) \leq N(\epsilon)\left(5\log\frac{1}{\delta} + \log n\right),$$

where $\epsilon = \delta^{\frac{2}{2-\rho}}$, and $N(\epsilon) \leq \epsilon^{-\rho}$. □ □

Remark 2.1 We used in the proof that if

$$\sum_{j=1}^{n} |\alpha_j|^\rho \leq 1,$$

then

$$N_\alpha(\epsilon) = \#\{\alpha_j : \ |\alpha_j| > \epsilon\} \leq \epsilon^{-\rho}.$$

It is also easy to see that in that case, for $\rho \leq 1$,

$$M_\alpha(\epsilon) = \sum_{|\alpha_j| \leq \epsilon} |\alpha_j| \leq \epsilon^{1-\rho}.$$

These inequalities we will use in Subsection 3.2 to prove adaptivity in ρ.

Relation with Besov spaces

In the literature on adaptive estimation, one often considers so-called Besov spaces $B_{\sigma,p,q}([0,1]^d)$. Such spaces are intrinsically connected to the analysis of curves since the scale of Besov spaces yelds the opportunity to describle the regularity of functions, with more accuracy than the classical Hölder scale. General references about Besov spaces are Besov, Il'in and Nikol'skii (1978), Berg and Löfström (1976), Triebel (1992) and DeVore and Lorentz (1993). This subsection discusses the link with our roughness parameter ρ. The notation $B_{\sigma,p,q}([0,1]^d)$ refers to the case of functions on the d-dimensional unit cube, with "smoothness" σ, and where p and q refer to L_p- and L_q-norms with respect to Lebesgue measure. We will not go into the details, but mainly want to show that, apart from logarithmic factors, such Besov spaces correspond to a roughness parameter ρ equal to $\rho = 2/(2s+1)$, where s is the "effective" smoothness σ/d, and where we assume $\rho \leq p$ (see Lemma 2.3). (Similar observations can be found in Donoho and Johnstone (1996).) The application of Lemma 2.2 then yields a bound for the entropy. However, in Besov spaces, the coefficients at higher *levels* tend to be smaller, i.e., there is a little more structure than as can be described by our roughness parameter ρ. As a result, it turns out that Besov spaces have entropies without logarithmic factors (see Lemma 2.3). Consider a wavelet basis ψ_{jk} of $L^2(\mathbf{P}_n)$ with regularity r such that $r \geq s$. We recall that a wavelet regularity is expressed through its number of vanishing moments, see e.g Meyer (1987) or Mallat (1990). Then a Besov norm is equivalent to an appropriate norm in the sequence space, that is, the space of the wavelet coefficients, see DeVore and Lorentz (1993) or Donoho, Johnstone, Kerkyacharyan and Picard (199for instance. We take a sequence space as a starting point (and consider for simplicity the case corresponding to $d = 1$ ($\sigma = s$) in the Besov interpretation). The coefficients α are now indexed by two integers: $\alpha = \{\alpha_{j,k}\}$, where k runs from 1 to 2^j, and where $j \in \{1, 2, \dots, J\}$ can be seen as a zoom-level.

Let $\mathcal{B}_{s,p,q}$ be the set of coefficients $\{\alpha_{j,k}\}$ that satisfy

$$\left(\sum_{j=1}^{J} 2^{j((2s+1)\frac{p}{2}-1)\frac{q}{p}} \left\{ \sum_{k=1}^{2^j} |\alpha_{j,k}|^p \right\}^{\frac{q}{p}} \right)^{\frac{1}{q}} \le 1. \tag{2.3}$$

This quantity is equivalent to the Besov semi-norm. Throughout, we assume $s \ge 0$, $p \ge 1$, and $q \ge 1$. In the Besov space interpretation, $\mathcal{B}_{s,p,q}$ (with $J = \infty$) corresponds (in the sense of norm equivalence) to a Besov ball in the space $B_{s,p,q}([0,1])$.

Lemma 2.2.2. *Suppose that $\alpha = \{\alpha_{j,k}\}$ satisfies (2.3), with $\rho = 2/(2s+1) \le \min(p, q)$, and $J < \infty$. Then*

$$\sum_{j=1}^{J} \sum_{k=1}^{2^j} |\alpha_{j,k}|^\rho \le J^{\frac{q-\rho}{q}}. \tag{2.4}$$

Proof. By Hölder's inequality, for a sequence $a_1, \dots, a_L$, and for $t \ge 1$,

$$\sum_{l=1}^{L} |a_l| \le L^{\frac{t-1}{t}} \left(\sum_{l=1}^{L} |a_l|^t \right)^{\frac{1}{t}}. \tag{2.5}$$

Apply this first with $L = J$, $|a_j| = \sum_{k=1}^{2^j} |\alpha_{j,k}|^\rho$, and $t = q/\rho$. Then we find

$$\sum_{j=1}^{J} \left\{ \sum_{k=1}^{2^j} |\alpha_{j,k}|^\rho \right\} \le J^{\frac{q-\rho}{q}} \left(\sum_{j=1}^{J} \left\{ \sum_{k=1}^{2^j} |\alpha_{j,k}|^\rho \right\}^{\frac{q}{\rho}} \right)^{\frac{\rho}{q}}. \tag{2.6}$$

Next, apply (2.5) with $L = 2^j$, $|a_{j,k}| = |\alpha_{j,k}|^\rho$, and $t = p/\rho$. This yields

$$\left\{ \sum_{k=1}^{2^j} |\alpha_{j,k}|^\rho \right\} \le \left\{ 2^{\frac{j(p-\rho)}{p}} (\sum_{k=1}^{2^j} |\alpha_{j,k}|^p)^{\frac{\rho}{p}} \right\}.$$

Do this for each $j = 1, \dots J$, and insert the result in (2.6):

$$J^{\frac{q-\rho}{q}} \left(\sum_{j=1}^{J} \left\{ \sum_{k=1}^{2^j} |\alpha_{j,k}|^\rho \right\}^{\frac{q}{\rho}} \right)^{\frac{\rho}{q}}$$

$$\leq J^{\frac{q-\rho}{q}} \left(\sum_{j=1}^{J} \left\{ 2^{\frac{j(p-\rho)}{p}} (\sum_{k=1}^{2^j} |\alpha_{j,k}|^p)^{\frac{\rho}{p}} \right\}^{\frac{q}{\rho}} \right)^{\frac{\rho}{q}}$$

$$= J^{\frac{q-\rho}{q}} \left(\sum_{j=1}^{J} 2^{j(\frac{p-\rho}{p})\frac{q}{\rho}} \left\{ \sum_{k=1}^{2^j} |\alpha_{j,k}|^p \right\}^{\frac{q}{p}} \right)^{\frac{\rho}{q}} \leq J^{\frac{q-\rho}{q}},$$

since

$$(\frac{p-\rho}{p})\frac{q}{\rho} = ((2s+1)\frac{p}{2} - 1)\frac{q}{p}.$$

□ □

Remark 2.2 One can also define spaces $\mathcal{B}_{s,p,q}$ with $p = \infty$ and/or $q = \infty$. Condition (2.3) is then to be understood with the usual adjustments. Note that $\mathcal{B}_{s,p,q} \subset \mathcal{B}_{s,p,\infty}$.

In our applications, the number of levels J is logarithmic in n. The entropy of the spaces $\mathcal{B}_{s,p,q}$ can now be bounded by combining Lemma 2.1 with Lemma 2.2. However, it turns out that this will result in a bound with an unnecessary $(\log n)$-term. The entropy bound without logarithmic factors can be found in Birman and Solomiak (1967) or in Birgé and Massart (1996).

We consider $\mathcal{B}_{s,p,q}$ as a subset of the Euclidean space $\mathbf{R}^{(2^J-2)}$ (possibly $J = \infty$), with Euclidean norm $\| \cdot \|$.

Lemma 2.2.3. *Let $\rho = 2/(2s+1) < p$. For a constant A depending on p and s,*

$$H(\delta, \mathcal{B}_{s,p,\infty}) \leq A\delta^{-\frac{1}{s}}, \quad \delta > 0. \tag{2.7}$$

Proof. This is shown in Birgé and Massart (1996). In fact, they show that the δ-entropy for the $L_{p'}$(Lebesgue measure)-norm, of a Besov ball with radius 1 in $B_{\sigma,p,\infty}([0,1]^d)$, is bounded by $A\delta^{-\frac{1}{s}}$, provided $s = \frac{\sigma}{d} > \frac{1}{p} - \frac{1}{p'}$. □ □

Remark 2.3 Inspection of the proof of the entropy result by Birgé and Massart (1996) reveals that if condition (2.3) holds, with $\rho = 2/(2s+1) < p$, then there exists an approximation $\bar{\alpha}$ with $N \asymp \delta^{-\frac{1}{s}}$ non-zero coefficients, such that $\|\alpha - \bar{\alpha}\| \leq \delta$. This we will use later on to re obtain adaptivity in the Besov case, of the LSE with soft thresholding, up to the correct logarithmic factors.

2.2.3 Least squares

In this section, we study the model $Y_i = \theta_0(z_i) + W_i$, $i = 1, \ldots, n$, with $W_1, \ldots, W_n$ i.i.d. $\mathcal{N}(0, \sigma_0^2)$-distributed errors.

Least squares without penalty

Suppose we know a priori that $\theta_0 \in \Theta$, where Θ is a given parameter space of regression functions. If Θ is "not too large", one need not penalize the LSE, i.e., one can use the "classical" LSE

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \theta(z_i))^2. \tag{3.1}$$

Theorem 2.2.4. *There exits a constant c depending only on the variance σ_0^2 of the errors, such that for*

$$J_n(\delta) \geq \int_{\delta^2/c}^{\delta} H^{1/2}(u, \{\theta \in \Theta : \ \|\theta - \theta_0\|_{Q_n} \leq \delta\}) du \vee \delta, \tag{3.2}$$

with $J_n(\delta)/\delta^2$ a non-increasing function of δ, and

$$\sqrt{n}\delta_n^2 \geq cJ_n(\delta_n), \tag{3.3}$$

one has for all $\delta \geq \delta_n$,

$$\mathbf{P}(\|\hat{\theta}_n - \theta_0\|_{Q_n} \geq \delta) \leq c \exp[-\frac{n\delta^2}{c^2}]. \tag{3.4}$$

Proof. *See e.g. van de Geer (1990, 2000).* □

Example 3.1 Let

$$\Theta = \{\theta = \sum_{j=1}^{n} \alpha_j \psi_j : \ \sum_{j=1}^{n} |\alpha_j|^\rho \leq 1\},$$

with $\psi_1, \ldots, \psi_n$ an orthonormal basis in $L_2(Q_n)$. Then it follows from Lemma 2.1 combined with the above theorem, that for $0 < \rho < 1$, the LSE $\hat{\theta}_n$ converges with rate

$$(\frac{\log n}{n})^{\frac{2-\rho}{4}}.$$

This corresponds to the minimax rate (see Donoho and Johnstone (1994)). The case $\rho = 0$ yields the rate

$$(\frac{\log n}{n})^{\frac{1}{2}}.$$

If $\rho \geq 1$, the entropy integral does not converge. For example, when $\rho = 1$, we find the rate

$$(\frac{\log^3 n}{n})^{\frac{1}{4}},$$

i.e., extra powers of $\log n$. The case $\rho > 1$ gives even rise to extra powers of n.

Observe that in the above example, the parameter space Θ, and hence the "classical" LSE $\hat{\theta}_n$, depends on ρ. In the next subsection, we will construct an estimator that does not depend on ρ, but that is shown to converge at the same rates (for $0 \leq \rho < 1$) as the one given above.

Least squares estimation using soft thresholding

In van de Geer (2000), least squares estimation using various penalties is considered. Here, we restrict ourselves to the special case of soft thresholding. (See also Remark 3.4.)

Let $\psi_1, \ldots, \psi_n$ be an orthonormal basis for $L_2(Q_n)$. Each $\theta : \mathcal{Z} \to \mathbf{R}$ can be written as $\theta = \sum_{j=1}^n \alpha_j \psi_j$ (in the sense of $L_2(Q_n)$-equivalence). In particular

$$\theta_0 = \sum_{j=1}^n \alpha_{j,0} \psi_j.$$

Let $\hat{\theta}_n$ now denote the LSE with soft thresholding, i.e.,

$$\hat{\theta}_n = \arg \min_{\theta=\alpha_1\psi_1+\ldots+\alpha_n\psi_n} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - \theta(z_i))^2 + 2\lambda_n^2 \sum_{j=1}^n |\alpha_j| \right\} = \sum_{j=1}^n \hat{\alpha}_{j,n} \psi_j. \tag{3.5}$$

An explicit expression, which explains why the method is called soft thresholding, is given in Remark 3.3 following Theorem 3.2 below. (We have added a factor 2 to the penalty term to simplify the expressions.) However, a rate of convergence can be established without using this explicit expression. This is of importance, because it will allow the extension to other estimation procedures, where no explicit expressions are available. The next theorem studies the general situation. It is followed by the example with roughness parameter ρ.

Theorem 2.2.5. *Suppose that the variance σ_0^2 of the errors is known (for simplicity) and take $\lambda_n^2 \geq \sigma_0 \sqrt{\frac{2 \log n}{n}}$. Let $\mathcal{J}_n$ be any subset of $\{1, \ldots, n\}$ and define*

$$N_n = |\mathcal{J}_n|, \; M_n = \sum_{j \notin \mathcal{J}_n} |\alpha_{j,0}|. \tag{3.6}$$

The LSE with soft thresholding $\hat{\theta}_n$ satisfies

$$||\hat{\theta}_n - \theta_0||_{Q_n} = O_{\mathbf{P}}(\lambda_n^2 N_n^{\frac{1}{2}} + \lambda_n M_n^{\frac{1}{2}}). \tag{3.7}$$

Proof. Write

$$V_j = \frac{1}{n} \sum_{i=1}^n W_i \psi_j(z_i), \; j = 1, \ldots, n.$$

Moreover, write

$$I_N(\alpha) = \sum_{j \in \mathcal{J}_n} |\alpha_j|, \; I_M(\alpha) = \sum_{j \notin \mathcal{J}_n} |\alpha_j|,$$

and (identifying a function θ with its coefficients α),

$$I_n(\alpha) = I_N(\alpha) + I_M(\alpha) = \sum_{j=1}^n |\alpha_j|.$$

Clearly, by the definition of the estimator $\hat{\theta}_n$,

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\theta}_n(z_i))^2 + 2\lambda_n^2 I_n(\hat{\theta}_n) \leq \frac{1}{n}\sum_{i=1}^{n}(Y_i - \theta_0(z_i))^2 + 2\lambda_n^2 I_n(\theta_0).$$

Rewrite this to

$$\|\hat{\alpha}_n - \alpha_0\|_n^2 + 2\lambda_n^2 I_n(\hat{\alpha}_n) \leq 2\sum_{j=1}^{n} V_j(\hat{\alpha}_{j,n} - \alpha_{j,0}) + 2\lambda_n^2 I_n(\alpha_0).$$

But this gives

$$\|\hat{\alpha}_n - \alpha_0\|_n^2 + 2\lambda_n^2 I_n(\hat{\alpha}_n) \leq 2 \max_{j=1,\ldots,n} |V_j| I_n(\hat{\alpha}_n - \alpha_0) + 2\lambda_n^2 I_n(\alpha_0).$$

If we take B_n as the set

$$B_n = \{\max_{j=1,\ldots,n} |V_j| \leq \lambda_n^2\},$$

we have

$$\mathbf{P}(B_n) \to 1.$$

On B_n, we find

$$\|\hat{\alpha}_n - \alpha_0\|_n^2 + 2\lambda_n^2 I_n(\hat{\alpha}_n) \leq 2\lambda_n^2 I_n(\hat{\alpha}_n - \alpha_0) + 2\lambda_n^2 I_n(\alpha_0), \tag{3.8}$$

or

$$\begin{aligned}\|\hat{\alpha}_n - \alpha_0\|_n^2 + 2\lambda_n^2 I_M(\hat{\alpha}_n) &\leq 2\lambda_n^2 I_N(\hat{\alpha}_n - \alpha_0) + 2\lambda_n^2 I_M(\hat{\alpha}_n - \alpha_0) \\ &\quad + 2\lambda_n^2 (I_N(\alpha_0) - I_N(\hat{\alpha}_n)) + 2\lambda_n^2 I_M(\alpha_0) \\ &\leq 4\|\hat{\alpha}_n - \alpha_0\|_n \lambda_n^2 N_n^{\frac{1}{2}} + 2\lambda_n^2 I_M(\hat{\alpha}_n) + 4\lambda_n^2 I_M(\alpha_0),\end{aligned}$$

or

$$\|\hat{\alpha}_n - \alpha_0\|_n^2 \leq 4\|\hat{\alpha}_n - \alpha_0\|_n \lambda_n^2 N_n^{\frac{1}{2}} + 4\lambda_n^2 I_M(\alpha_0).$$

Note now that $I_M(\alpha_0) = M_n$, by definition. Thus, we arrive at (3.7). □ □

Example 3.2 Suppose that

$$\sum_{j=1}^{n} |\alpha_{j,0}|^\rho \leq 1,$$

for some $0 \leq \rho < 1$. Take $\mathcal{J}_n = \{j : \ |\alpha_{j,0}| > \lambda_n^2\}$. Using Remark 2.1 with $\epsilon = \lambda_n^2$, we obtain that

$$\begin{aligned}N_n &\leq \lambda_n^{-2\rho} \\ M_n &\leq \lambda_n^{2(1-\rho)}\end{aligned}$$

and as a result we get

$$\lambda_n^2 N_n^{\frac{1}{2}} + \lambda_n M_n^{\frac{1}{2}} \leq 2\lambda_n^{2-\rho}.$$

Previous inequality together with Theorem 3.2 show that

$$\|\hat{\theta}_n - \theta_0\|_{Q_n} = O_{\mathbf{P}}(\lambda_n^{2-\rho}).$$

In other words, the optimal choice $\lambda_n^2 = \sigma_0\sqrt{\frac{2\log n}{n}}$ yields exactly the same rate as in Example 3.1, i.e., the LSE with soft thresholding (which does not use knowledge of ρ) adapts to ρ.

Remark 3.1 Application of Theorem 3.2 when α_0 lies in the space $\mathcal{B}_{s,p,q}$ defined in Subsection 2.2 does not directly give the correct rate. However, one can easily extend the theorem to show that (under the same condition on λ_n) one has

$$\|\hat{\theta}_n - \theta_0\|_{Q_n} = O_{\mathbf{P}}(\lambda_n^2 N_n^{\frac{1}{2}} + \lambda_n \bar{M}_n^{\frac{1}{2}} + \|\theta_0 - \bar{\theta}_n\|_{Q_n}),$$

where $\bar{\theta}_n = \sum_{j=1}^n \bar{\alpha}_{j,n}\psi_j$ is some (any) approximation of θ_0, and where

$$\bar{M}_n = \sum_{j \notin \mathcal{J}_n} |\bar{\alpha}_{j,n}|.$$

Remark 2.3 now gives

$$\|\hat{\theta}_n - \theta_0\|_{Q_n} = O_{\mathbf{P}}(\lambda_n^{2-\rho}),$$

when $\theta_0 \in \mathcal{B}_{s,p,q}$, with $\rho = 2/(2s+1) < p$. This coincides with the result in e.g. Donoho, Johnstone, Kerkyacharian and Picard (1996).

Remark 3.2 The condition of normally distributed errors can be dropped, provided we adjust the value of the smoothing parameter in the appropriate way. Namely, λ_n^2 should be chosen in such a way that

$$\mathbf{P}(\max_{j=1,\ldots,n} |V_j| \leq \lambda_n^2) \to 1,$$

where $V_j = \sum_{i=1}^n W_i\psi_j(z_i)/n$, $j = 1,\ldots,n$. Thus, the (optimal) choice of the smoothing parameter depends on the distribution of the errors. As a consequence, if the errors have heavy tails, the rate of convergence of the LSE with soft thresholding may be very slow.

Remark 3.3 Each coefficient is estimated separately. For completeness, we present here the expression of the estimated coefficients. Let us write

$$\tilde{\alpha}_{j,n} = \frac{1}{n}\sum_{i=1}^n Y_i\psi_j(z_i)$$

for the empirical coefficients. Then the penalized least squares estimators $\hat{\theta}_n = \sum_{j=1}^n \hat{\alpha}_{j,n}\psi_j$ is solution of the optimization program:

$$(\hat{\alpha}_{j,n}) = \arg\min_{\alpha_j}\left(\sum_{j=1}^n |\tilde{\alpha}_{j,n} - \alpha_j|^2 + 2\lambda_n^2\sum_{j=1}^n |\alpha_j|\right).$$

This minimization problem has the following explicit solution:

$$\hat{\alpha}_{j,n} = \begin{cases} \tilde{\alpha}_{j,n} - \lambda_n^2, & \text{if } \tilde{\alpha}_{\text{j,n}} > \lambda_{\text{n}}^2, \\ 0, & \text{if } |\tilde{\alpha}_{\text{j,n}}| \leq \lambda_{\text{n}}^2, \\ \tilde{\alpha}_{j,n} + \lambda_n^2, & \text{if } \tilde{\alpha}_{\text{j,n}} < -\lambda_{\text{n}}^2. \end{cases}$$

Remark 3.4 Theorem 10.2 in van de Geer (2000) is about penalized LS estimation in general. It cannot directly be applied to LS with soft thresholding, because the entropy of the set $\{\theta : I_n(\theta) \leq 1\}$ is not integrable. Adjusting the previous theorem in van de Geer (2000) to cover our case is however straightforward, and will yield the same result of Theorem 3.2, but with additional logarithmic factors, and restricted to the case $\mathcal{J}_n = \emptyset$, so that $N_n = 0$ and $M_n = I_n(\theta_0)$. In other words, the result using entropy methods of van de Geer (2000) is that $\|\hat{\theta}_n - \theta_0\|_{Q_n} = O_{\mathbf{P}}(\lambda_n I_n^{1/2}(\theta_0))$ times some log-factors. So when $I_n(\theta_0) \leq 1$ (say), one essentially arrives at a $n^{-\frac{1}{4}}$-rate of convergence, i.e., the entropy approach leads to rates for the roughest case $\rho = 1$. It may however be true that a more refined entropy approach, using local entropies, is capable of reproving adaptation (modulo log-factors) to the roughness of θ_0.

2.2.4 Robust adaptive estimators

Recall the regression model

$$Y_i = \theta_0(z_i) + W_i, \ i = 1, \ldots, n,$$

where $\theta_0 : \mathcal{Z} \to \mathbf{R}$ is an unknown regression function in the parameter space Θ.

The estimator with soft thresholding type penalty, based on the convex loss function γ, is defined as

$$\hat{\theta}_n = \min_{\theta = \alpha_1\psi_1 + \ldots + \alpha_n\psi_n \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^n \gamma(Y_i - \theta(z_i)) + \lambda_n^2 \sum_{j=1}^n |\alpha_j| \right\} = \sum_{j=1}^n \hat{\alpha}_{j,n}\psi_j. \tag{4.1}$$

The case $\gamma(\xi) = \xi^2$ was studied in the previous section. In this section, we look at the robust case, with γ satisfying

$$|\gamma(\xi) - \gamma(\tilde{\xi})| \leq |\xi - \tilde{\xi}|, \ \xi, \tilde{\xi} \in \mathbf{R}. \tag{4.2}$$

The following notation is convenient. Let $X_i = (Y_i, z_i)$, let P_n be the empirical distribution based on $X_1, \ldots, X_n$, $P^{(i)}$ the distribution of X_i, and $\bar{P} = \sum_{i=1}^n P^{(i)}/n$. In the introduction, we assumed that $\mathbf{E}\gamma(Y_i - b)$ has a unique minimum in $b = \theta_0(z_i)$. We now need the following stronger condition. Let $\gamma_\theta(X_i) = \gamma(Y_i - \theta(z_i))$. For $0 \leq t \leq 1$, write $\theta_t = t\theta + (1-t)\theta_0$. (We do not require $\theta_t \in \Theta$.) We will assume that for t_0 sufficiently small (not depending on n) and $\theta \in \Theta$, we have for all $0 \leq t \leq t_0$,

$$\int (\gamma_{\theta_t} - \gamma_{\theta_0}) d\bar{P} \geq \epsilon \|\theta_t - \theta_0\|_{Q_n}^2, \tag{4.3}$$

where $\epsilon > 0$ is a constant (not depending on n). If one is willing to assume that functions in Θ are bounded in sup-norm (by a constant K not depending on n), this is a reasonable condition, because $\int \gamma_\theta d\bar{P}$ is minimized at θ_0. However, boundedness in sup-norm is of course an awkward condition. (See also Remark 4.1 below.)

The choice for the smoothing parameter in Theorem 4.1 is taken on the safe side. It is based on the empirical process inequality of Lemma 4.2.

Theorem 2.2.6. *Suppose that (4.3) holds. Then for any subset $\mathcal{J}_n$ of $\{1,\ldots,n\}$ and for*

$$N_n = |\mathcal{J}_n|,\ M_n = \sum_{j \notin \mathcal{J}_n} |\alpha_{j,0}|$$

and $\lambda_n^2 \geq c\sqrt{\log n/n}$, where c is a universal constant, we have

$$||\hat{\theta}_n - \theta_0||_{Q_n} = O_{\mathbf{P}}((\lambda_n^2 N_n^{\frac{1}{2}} + \lambda_n M_n^{\frac{1}{2}}) \vee n^{-\frac{1}{2}}),$$

provided $\lambda_n^2 N_n^{\frac{1}{2}} + \lambda_n M_n^{\frac{1}{2}} \to 0$.

Proof. The proof follows the line of reasoning of Theorem 3.2 on the LSE.

Define $t = t_0/(1 + ||\hat{\theta}_n - \theta_0||_{Q_n})$. Consider the convex combination

$$\hat{\theta}_{t,n} = t\hat{\theta}_n + (1-t)\theta_0.$$

Using the convexity of the loss function γ, and of the soft thresholding type penalty, we obtain

$$\frac{1}{n}\sum_{i=1}^n \gamma(Y_i - \hat{\theta}_{t,n}(z_i)) + \lambda_n^2 I_n(\hat{\theta}_{t,n})$$

$$\leq t\left\{\frac{1}{n}\sum_{i=1}^n \gamma(Y_i - \hat{\theta}_n(z_i)) + \lambda_n^2 I_n(\hat{\theta}_n)\right\} + (1-t)\left\{\frac{1}{n}\sum_{i=1}^n \gamma(Y_i - \theta_0(z_i)) + \lambda_n^2 I_n(\theta_0)\right\}$$

$$\leq \frac{1}{n}\sum_{i=1}^n \gamma(Y_i - \theta_0(z_i)) + \lambda_n^2 I_n(\theta_0),$$

where in the second inequality, we used that $\hat{\theta}_n$ minimizes the penalized loss function. We rewrite this in a convenient form, namely

$$\int (\gamma_{\hat{\theta}_{t,n}} - \gamma_{\theta_0}) d\bar{P} + \lambda_n^2 I_n(\hat{\theta}_{t,n}) \leq -\int (\gamma_{\hat{\theta}_{t,n}} - \gamma_{\theta_0}) d(P_n - \bar{P}) + \lambda_n^2 I_n(\theta_0).$$

By assumption (4.3),

$$\int (\gamma_{\hat{\theta}_{t,n}} - \gamma_{\theta_0}) d\bar{P} \geq \epsilon ||\hat{\theta}_{t,n} - \theta_0||_{Q_n}^2.$$

Now, if $I_n(\hat{\theta}_{t,n}-\theta_0) \le n^{-\frac{1}{2}}$, it follows immediately that also $\|\hat{\theta}_{t,n}-\theta_0\|_{Q_n} \le n^{-\frac{1}{2}}$. Let us therefore assume that $I_n(\hat{\theta}_{t,n}-\theta_0) > n^{-\frac{1}{2}}$, and (for simplicity) that this is true for all n. Observe also that $\|\hat{\theta}_{t,n}-\theta_0\|_{Q_n} \le 1$. Let B_n be the set

$$|\int(\gamma_{\hat{\theta}_{t,n}}-\gamma_{\theta_0})d(P_n-\bar{P})| \le c\sqrt{\frac{\log n}{n}}I_n(\hat{\theta}_{t,n}-\theta_0).$$

We show in Lemma 4.2 that for a c a large enough universal constant,

$$\mathbf{P}(B_n) \to 1.$$

On B_n, we find

$$\epsilon\|\hat{\theta}_{t,n}-\theta_0\|^2_{Q_n}+\lambda_n^2 I_n(\hat{\theta}_{t,n}) \le \lambda_n^2 I_n(\hat{\theta}_{t,n}-\theta_0)+\lambda_n^2 I_n(\theta_0). \tag{4.4}$$

This is, apart from some constants, the same inequality as (3.8) in Theorem 3.2, and we may proceed as there. We then arrive at

$$\|\hat{\theta}_{t,n}-\theta_0\|_{Q_n} = O_{\mathbf{P}}(\lambda_n^2 N_n^{\frac{1}{2}}+\lambda_n M_n^{\frac{1}{2}}).$$

But then also

$$\|\hat{\theta}_n-\theta_0\|_{Q_n} = \|\hat{\theta}_{t,n}-\theta_0\|_{Q_n}/t = O_{\mathbf{P}}(\lambda_n^2 N_n^{\frac{1}{2}}+\lambda_n M_n^{\frac{1}{2}}).$$

□ □

Remark 4.1 It can be seen from the proof that it is allowed to let the value of t_0 in condition (4.3) to depend on θ as well as θ_0, as long as the last step in the proof remains true. Indeed, in the proof we took $t \le t_0$ depending on the parameter, a trick which allows one to restrict attention to a unit ball around θ_0 when the loss function and penalty are convex (see also van de Geer 1999).

Remark 4.2 Theorem 4.1 shows that, except for some logarithmic factors, for the robust penalized regression estimator one has exactly the same result as for the penalized LSE. Therefore, one can deduce e.g. adaptivity up to logarithmic factors, to the roughness parameter ρ (as in Example 3.2).

The robustness condition (4.2) implies that a choice of the smoothing parameter that works well for all error distributions, is possible. The optimal choice is as yet not clear, but the following empirical process inequality gives an upper bound (still depending on an unspecified universal constant c).

Lemma 2.2.7. *There exists a constant c such that*

$$\mathbf{P}\left(\sup_{\|\theta-\theta_0\|_{Q_n}\le 1,\ I_n(\theta-\theta_0)>n^{-\frac{1}{2}}} \frac{|\int(\gamma_\theta-\gamma_{\theta_0})d(P_n-\bar{P})|}{I_n(\theta-\theta_0)} \ge c\sqrt{\frac{\log n}{n}}\right) \le c\exp[-\frac{\log n}{c^2}].$$

Proof. Without loss of generality, we may assume $\theta_0 \equiv 0$. Now, note that

$$||\theta||_{Q_n} \leq I_n(\theta) \leq \sqrt{n}||\theta||_{Q_n},$$

and that for some constant A, and all $M \leq \sqrt{n}$,

$$H(u, \{\theta : \ I_n(\theta) \leq M\}) \leq A\frac{M^2}{u^2}\log n, \ u > n^{-\frac{1}{2}}/8,$$

where we invoked Lemma 2.1. We show in the following lemma that for some constant C_0, we have the concentration inequality

$$\mathbf{P}\left(\sup_{||\theta||_{Q_n}\leq 1, \ I_n(\theta)\leq M} |\int(\gamma_\theta - \gamma_0)d(P_n - \bar{P})| \geq C_0\sqrt{\frac{\log n}{n}}M\right) \leq C_0\exp[-\frac{(\log n)(1 \vee M^2)}{C_0^2}].$$

Take j_0 as the smallest integer such that $j_0 + 1 > \log_2\sqrt{n}$. Then we find

$$\mathbf{P}\left(\sup_{n^{-\frac{1}{2}}<I_n(\theta)\leq 1} \frac{|\int(\gamma_\theta - \gamma_0)d(P_n - \bar{P})|}{I_n(\theta)} \geq 2C_0\sqrt{\frac{\log n}{n}}\right)$$

$$\leq \sum_{j=0}^{j_0}\mathbf{P}\left(\sup_{I_n(\theta)\leq 2^{-j}} |\int(\gamma_\theta - \gamma_0)d(P_n - \bar{P})| \geq C_0\sqrt{\frac{\log n}{n}}2^{-j}\right)$$

$$\leq (\log_2\sqrt{n} + 1)C_0\exp[-\frac{\log n}{C_0^2}]$$

$$\leq C_0\exp[-\frac{\log n}{c^2}].$$

Moreover,

$$\mathbf{P}\left(\sup_{||\theta||_{Q_n}\leq 1, \ I_n(\theta)>1} \frac{|\int(\gamma_\theta - \gamma_0)d(P_n - \bar{P})|}{I_n(\theta)} \geq 2C_0\sqrt{\frac{\log n}{n}}\right)$$

$$\leq \sum_{j=0}^{\infty}\mathbf{P}\left(\sup_{||\theta||_{Q_n}\leq 1, \ I_n(\theta)\leq 2^{j+1}} |\int(\gamma_\theta - \gamma_0)d(P_n - \bar{P})| \geq C_0\sqrt{\frac{\log n}{n}}2^{j+1}\right)$$

$$\leq \sum_{j=0}^{\infty}\exp[-\frac{(\log n)2^{2(j+1)}}{C_0^2}]$$

$$\leq C_1\exp[-\frac{\log n}{c^2}].$$

□ □

Lemma 2.2.8. *There exists a constant C_0 such that the following upper bound holds*

$$\mathbf{P}\left(\sup_{||\theta||_{Q_n}\leq 1,\ I_n(\theta)\leq M} |\int(\gamma_\theta-\gamma_0)d(P_n-\bar{P})| \geq C_0\sqrt{\frac{\log n}{n}}M\right) \leq C_0\exp[-\frac{(\log n)(1\vee M^2)}{C_0^2}].$$

Proof. Define the following empirical process

$$Z = \sup_{||\theta||_{Q_n}\leq 1,\, I_n(\theta)\leq M\}} \left|\int(\gamma_\theta-\gamma_0)\right| d(\mathbf{P}_n-\hat{\mathbf{P}}).$$

Set

$$U_i(\theta)=\gamma(Y_i-\theta(z_i))-\gamma(Y_i),\ i=1,\ldots,n.$$

An Höffding type inequality, proved by Massart (2000), says that for all $u\geq 0$ we get

$$\mathbf{P}\left(Z\geq \mathbf{E}(Z)+u\right)\leq \exp\left[-\frac{n^2u^2}{8b_n^2}\right]$$

where b_n satisfies

$$\sup_{||\theta||_{Q_n}\leq 1,\, I_n(\theta)\leq M}\sum_{i=1}^n ||U_i(\theta)-\mathbf{E}(U_i(\theta))||_\infty^2\leq b_n^2.$$

Since γ is $1-$Lipschitz, we have that

$$\begin{aligned}\sup_{||\theta||_{Q_n}\leq 1,\, I_n(\theta)\leq M}\sum_{i=1}^n ||U_i(\theta)-\mathbf{E}(U_i(\theta))||_\infty^2 &\leq 4\sup_{||\theta||_{Q_n}\leq 1,\, I_n(\theta)\leq M}\sum_{i=1}^n\theta^2(z_i)\\ &\leq 4n\end{aligned}$$

As a result, we have an upper bound for b_n and we take $b_n^2=4n$. A symmetrization procedure implies the following bound:

$$\begin{aligned}\mathbf{E}(Z)&=\mathbf{E}\left[\sup_{||\theta||_{Q_n}\leq 1,\, I_n(\theta)\leq M}\frac{1}{n}\left|\sum_{i=1}^n[U_i(\theta)-\mathbf{E}(U_i(\theta))]\right|\right]\\ &\leq 2\mathbf{E}\left[\sup_{||\theta||_{Q_n}\leq 1,\, I_n(\theta)\leq M}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_iU_i(\theta)\right|\right]\end{aligned}$$

where the ϵ_i's are Rademacher random variables. Theorem 4.12 in Ledoux and Talagrand's book (1991) based on a contraction principle and using the fact that γ is Lipschitz gives the following bound for the last quantity:

$$\begin{aligned}\mathbf{E}\left[\sup_{||\theta||_{Q_n}\leq 1,\, I_n(\theta)\leq M}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_iU_i(\theta)\right|\right]&\leq\mathbf{E}\left[\sup_{||\theta||_{Q_n}\leq 1,\, I_n(\theta)\leq M}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i\theta(z_i)\right|\right]\\ &\leq\mathbf{E}\left[\sup_{||\theta||_{Q_n}\leq 1,\, I_n(\theta)\leq M}\left|\frac{1}{n}\sum_{j=1}^n\alpha_j\sum_{i=1}^n\epsilon_i\psi_j(z_i)\right|\right]\\ &\leq M\mathbf{E}\left[\max_{j=1,\ldots,n}\left|\frac{1}{n}\sum_{i=1}^n\epsilon_i\psi_j(z_i)\right|\right]\end{aligned}$$

The last quantity is of order $M\sqrt{\frac{\log n}{n}}$. As a consequence we get

$$\mathbf{P}[Z \leq 2M\sqrt{\frac{\log n}{n}} + u] \geq 1 - \exp\left(-\frac{nu^2}{32}\right)$$

Taking $u \approx \sqrt{\frac{\log n}{n}}$ completes the proof of the lemma. □

2.2.5 Least absolute deviations

Let us (for simplicity) assume that $W_1, \ldots, W_n$ are i.i.d. copies of a random variable W. Suppose W has median zero. We may then consider the least absolute deviations (LAD) loss function. The LAD estimator with soft thresholding type penalty is

$$\hat{\theta}_n = \min_{\theta=\alpha_1\psi_1+\ldots+\alpha_n\psi_n\in\Theta} \left\{ \frac{1}{n}\sum_{i=1}^n |Y_i - \theta(z_i)| + \lambda_n^2 \sum_{j=1}^n |\alpha_j| \right\} = \sum_{j=1}^n \hat{\alpha}_{j,n}\psi_j.$$

Note that this estimator has a certain scale invariance, in the sense that the optimal value of the smoothing parameter does not depend on the scale of the data. In particular, the smoothing parameter will not depend on (an estimate of) the variance of the data.

In order to apply Theorem 4.1, we need to verify condition (4.3). Suppose W has density f_W with respect to Lebesgue measure, and that for some $\epsilon > 0$,

$$f_W(w) \geq \epsilon, \text{ for all } |w| \leq \epsilon.$$

Let us write

$$|\theta|_\infty = \max_{z\in\mathcal{Z}} |\theta(z)|.$$

Assume that for some constant $K < \infty$,

$$\sup_{\theta\in\Theta} |\theta|_\infty \leq K.$$

Then indeed, one can easily see that (4.3) holds with $t_0 = \min(\frac{1}{2}, \frac{\epsilon}{2K})$.

Remark 5.1 It is not a good idea to apply LAD (or other robust methods) with soft thresholding in the sequence space. To see why, let us denote the empirical coefficients by

$$\tilde{\alpha}_{j,n} = \frac{1}{n}\sum_{i=1}^n Y_i\psi_j(z_i), \ j = 1, \ldots, n.$$

Renormalize to

$$\tilde{Y}_j = \sqrt{n}\tilde{\alpha}_{j,n}, \ j = 1, \ldots, n,$$

with expectation (assuming the errors are centered)

$$\mathbf{E}\tilde{Y}_j = \sqrt{n}\alpha_{j,0} := \vartheta_{j,0}, \ j = 1, \ldots, n.$$

The LAD estimator of ϑ_0 is now

$$\hat{\vartheta}_n = \arg\min_{\vartheta \in \mathbf{R}^n} \left\{ \sum_{j=1}^{n} |\tilde{Y}_j - \vartheta_j| + \sqrt{n}\lambda_n^2 \sum_{j=1}^{n} |\vartheta_j| \right\}.$$

Thus, as soon as $\sqrt{n}\lambda_n^2 > 1$, $\hat{\vartheta}_n \equiv 0$. Suppose now that ϑ_0 remains bounded, say that $|\vartheta_{j,0}| \leq 1$ for all j. The result of Theorem 4.1 is then trivial:

$$\frac{1}{n}\sum_{j=1}^{n} \vartheta_{j,0}^2 = \frac{1}{n}\sum_{j \in \mathcal{J}_n} \vartheta_{j,0}^2 + \frac{1}{n}\sum_{j \notin \mathcal{J}_n} \vartheta_{j,0}^2$$

$$\leq \frac{N_n}{n} + \frac{1}{n}\sum_{j \notin \mathcal{J}_n} |\vartheta_{j,0}| = \frac{N_n}{n} + \frac{1}{\sqrt{n}}\sum_{j \notin \mathcal{J}_n} |\alpha_{j,0}|.$$

2.2.6 Simulation study

In our simulation study, we allow ourselves the following deviations from the main theory. Firstly, we will not restrict the functions θ to be bounded in sup-norm by some constant. Secondly, we take values λ_n^2 proportional to $\sqrt{2\log n/n}$ as given in Remark 4.3. (Note that in the LAD case, the random variables $(\gamma_\theta(X_i) - \gamma_{\theta_0}(X_i))/(\theta(z_i) - \theta_0(z_i))$ are $+1$ or -1 as soon as $|W_i| \geq |\theta(z_i) - \theta_0(z_i)|$).

The signals have been generated using the software MATLAB. We consider the Heavisine function and the Doppler function with n=100 observations, and decompose these two functions onto a wavelet basis using a Daubechies wavelet with 8 vanishing moments. As error distribution, we considered the standard centered Gaussian distribution with variance 3, and also the Laplacian distribution, i.e., double exponential distribution, with mean zero and variance 3.

The LS estimator is computed using its explicit expression whereas the LAD estimator must be numerically computed. Since the LAD estimator is defined as the solution of an l^1 minimization, standard minimization algorithm do not give good results. That is the reason why, following ideas developped by Bruce and Sardy (1999) for the Huber loss function, we consider the minimization problem as an optimization issue with Lagrange multipliers and the associated dual, see for instance Rockafellar (1970). A primal-dual algorithm with a log-barrier penalty, described by Chen, Donoho and Saunders (1999) provides an efficient numerical method to compute the LAD estimator.

We have looked at 9 cases, corresponding to different values of the smoothing-parameter λ_n^2 including the theoretical optimal value for the Gaussian case 0.303 that corresponds to the 8^{th} line. The four tables summarize the performance of the LS and LAD estimators in term of mean square error (MSE). The numbers represent an average over 20 simulations. In order to make comparison of LS and LAD relevant, we have put on a same line the results with $\sigma_0\lambda_n^2$ for the LS and λ_n^2 for the LAD. We also added a line where comparisons are made for the optimal cases, i.e., smallest $||\hat{\theta}_n - \theta_0||_{Q_n}$ (corresponding to different smoothing parameters).

In these simulations, we can see that LAD works better in the Laplacian case, and LS works better in th Gaussian case (as is to be expected). We note furthermore that the value $\lambda_n^2 = \sqrt{2\log n/n}$ is not optimal in the LAD case: it is too large. In the LS case, the corresponding value $\lambda_n^2 = \sigma_0\sqrt{2\log n/n}$ is also too large when the errors are Laplacian, but it is optimal when the errors are Gaussian.

We show the results for some significant simulations. The LS estimator is represented in dotted lines and the LAD estimator is represented with solid lines. The figures 1-2 show the results obtained for the two different functions Doppler and Heavisine with the Gaussian noise. The last figure, figure 3, shows the results when taken Heavisine function corrupted by Laplacian noise. We can observe that LAD catches better the irregularity of the two functions Heavisine and Doppler, when LS is too smooth. However LAD with a wavelet basis has limitations because its ability to estimate spatially inhomogeneous signals conflicts with the goal of robustness to filter noise.

Heavisine function with Gaussian errors.

λ^2	MSE for LS	MSE for LAD
0.0303 (1)	0.7535	0.605
0.0607 (2)	0.5229	0.3994
0.1011 (3)	0.4782	0.3737
0.1517 (4)	0.4934	0.4507
0.2124 (5)	0.4749	0.4612
0.2427 (6)	0.3451	0.4828
0.2731 (7)	0.2821	0.5003
0.3034 (8)	0.2238	0.5601
0.6070 (9)	0.5852	0.6242
optimum	0.2238 at (8)	0.3737 at (3)

Heavisine function with Laplacian noise.

λ^2	MSE for LS	MSE for LAD
0.0303 (1)	1.7051	1.5157
0.0607 (2)	1.010	0.954
0.1011 (3)	.8201	0.6238
0.1517 (4)	0.7853	0.5896
0.2124 (5)	0.6021	0.4324
0.2427 (6)	0.5925	0.4654
0.2731 (7)	0.5896	0.5870
0.3034 (8)	0.6012	0.6925
0.607 (9)	0.6238	0.7021
optimum	0.5896 at (7)	0.4324 at (5)

Doppler signal with Gaussian errors.

λ^2	MSE for LS	MSE for LAD
0.0303 (1)	0.5862	0.8103
0.0607 (2)	0.5210	0.7801
0.1011 (3)	0.3521	0.6610
0.1517 (4)	0.2625	0.5218
0.2124 (5)	0.2212	0.3451
0.2427 (6)	0.1521	0.2821
0.2731 (7)	0.1330	0.3299
0.3034 (8)	0.090	0.4445
0.6070 (9)	0.3928	0.5510
optimum	0.090 at (8)	0. 2821 at (6)

Doppler signal with Laplacian errors.

λ^2	MSE for LS	MSE for LAD
0.0303 (1)	0.736	0.901
0.0607 (2)	0.6260	0.7700
0.1011 (3)	0.5218	0.6101
0.1517 (4)	0.5680	0.5018
0.2124 (5)	0.6321	0.3451
0.2427 (6)	0.7081	0.2821
0.2731 (7)	0.8588	0.8229
0.3034 (8)	0.9097	0.8429
0.607 (9)	0.9254	0.9545
optimum	0.5218 at (3)	0.2521 at (6)

2.3 Penalized Density Estimation

In this part, we focus on the density estimation problem using different penalization techniques. In a first part, we will investigate a smoothing approach obtained by minimizing a quadratic loss-function penalized by Besov pseudo-norms. The lack of adaptivity of this method will lead us to use empirical process theory to study the behavior of another estimator: we penalize the log-likelihood by a functional which depends on the roughness of the logarithm of the density. More precisely, we will use a functional based on the wavelet coefficients of the log-density.

2.3.1 A smoothing technique

The model is the following: let X be a random variable that ranges over a subinterval of $\mathbb{R}$ and has unknown probability distribution $\mathbf{P}$ with density f_0 with respect to Lebesgue measure, lying in a functional space $\mathcal{F}$,

$$\frac{dP}{d\lambda} = f_0.$$

Suppose we observe n independent observations $X_1, \ldots, X_n$ of X and use this random sample for inference about the density f_0. Linear estimators of the density can be obtained using a wavelet decomposition of the density. As a matter of fact, we make the assumption that f_0 belongs to a Besov space $B^s_{p\infty}$, $s > \frac{1}{p}$. Consider a wavelet ψ with $r > s$ regularity and compactly supported that provides an unconditional basis of such space. So every function f can be decomposed onto this basis as follows:

$$f = \sum_j \sum_k \beta_{jk}\psi_{jk}, \ \beta_{jk} =< f, \psi_{jk} > .$$

D. Donoho, I. Johnstone, G. Kerkyacharyan and D. Picard in [DJ95] or [DJKP95] for instance have constructed the linear wavelet estimator:

$$\hat{f}(x) = \sum_{(j,k)} \hat{\alpha}_{jk}\psi_{jk}(x)$$

where $\hat{\alpha}_{jk} =< f_0, \psi_{jk} >$ is the estimated wavelet coefficient :

$$\hat{\alpha}_{jk} = \frac{1}{n}\sum_{i=1}^{n} \psi_{jk}(X_i).$$

Since

$$\begin{aligned} E\hat{\alpha}_{jk} &= E\psi_{jk}(X_1) \\ &= \int \psi_{jk}(x) f_0(x) dx \\ &= \alpha_{jk} \end{aligned}$$

$\hat{f}$ is an unbiased estimator of the density. But this estimator has to be truncated to use only a finite number of resolution levels. Moreover it may be non smooth, which justifies the use of smoothing techniques to improve its behavior. Indeed, we will try to solve the approximation issue which consists in finding a function $\tilde{f}$ close enough to this estimator but with a regularity that can be easily controlled. For this consider the smoothing problem

$$\tilde{f} = \arg \min_{g=\sum_{j=j_0}^{j_1} \sum_k \beta_{jk}\psi_{jk} \in \mathcal{F}} \left(||D_{j_0,j_1} f_0 - g||_2^2 + I(g) \right) \tag{2.3.1}$$

$$= \arg \min_{g=\sum_{j_0}^{j_1} \sum_k \beta_{jk}\psi_{jk} \in \mathcal{F}} \left(\sum_{j=j_0}^{j_1} \sum_k |\hat{\alpha}_{jk} - \beta_{jk}|^2 + I((\beta_{jk})_{(j,k)}) \right) \tag{2.3.2}$$

The optimization problem is transformed due to orthonormality of wavelet transform. We already know that, for special choices of penalty, the minimization problem can be solved. The explicit solutions lead to well-known wavelet estimator: for a choice $I((\beta_{jk})_{(j,k)}) = 2\lambda_n^2 \sum_{(j,k)} |\beta|_{jk}$ and resp. $I((\beta_{jk})_{(j,k)}) = \lambda_n^2 \#\{|\beta_{jk}| > 0\}$ we obtain the soft-thresholded wavelet estimator

$$\tilde{f} = \sum_{(j,k)} \mathrm{sgn}(\hat{\alpha}_{jk})(|\hat{\alpha}_{jk}| - \lambda_n^2)_+ \psi_{jk}$$

and resp. the hard-thresholded wavelet estimator

$$\tilde{f} = \sum_{(j,k)} \hat{\alpha}_{jk} 1_{|\hat{\alpha}_{jk}| > \lambda_n^2} \psi_{jk}.$$

The asymptotic behavior of such estimators have been studied by D.Donoho, I.Johnstone, G. Kerkyacharian and D. Picard in [DJ95] or [DJKP96b]. They proved that such estimators achieve the minimax rate of convergence over compact sets of $B^s_{p\infty}$ for $B^\sigma_{p\infty}$-losses with $\sigma \geq s$, up to some logarithmic factors.

If we want to control more directly the smoothness of the estimator, we must choose a norm-type penalty with a B^s_{22} pseudo-norm. So we have

$$I(g) = \lambda_n^2 \sum_{j=j_0}^{j_1} 2^{2js} \sum_k \beta_{jk}^2 = \lambda_n^2 ||g||_{B^s_{22}}$$

and

$$\tilde{f}_{\lambda_n} = \arg \min_{\beta_{jk}} \left(\sum_{j_0}^{j_1} (\hat{\alpha}_{jk} - \beta_{jk})^2 + \lambda_n^2 \sum_{j_0}^{j_1} 2^{2js} \sum_k \beta_{jk}^2 \right).$$

This approach is similar to the approximation point of view with the use of the K-Peetre functional, described by Peetre in [Pee70] . In that case, we have proven that the estimator is a smoothed estimator

$$\tilde{f}_{\lambda_n} = \sum_k \hat{\alpha}_{j_0 k} \phi_{j_0 k} + \sum_{j=j_0}^{j_1} \sum_k \tilde{\beta}_{jk} \psi_{jk}$$

where for all $j = j_0 \dots j_1$

$$\tilde{\beta}_{jk} = \frac{\hat{\beta}_{jk}}{(1+\lambda_n^2 2^{2js})}.$$

This estimator is consistent provided the same hypothesis as in the case of the regression holds. It achieves the minimax rate of convergence for a convenient choice of the smoothing parameter λ_n as asserts the following theorem.

Theorem 2.3.1. *Assume that $f_0 \in L^\infty \cup B^s_{22}$. Moreover assume that the resolution levels satisfy $j_0 = 0(1)$ and $2^{j_1} = o(n^{\frac{1}{1+2s}})$. For a choice of the smoothing parameter such that $\lambda_n^2 = n^{-\frac{2s}{1+2s}}$, there exists a finite positive constant c such that for every $0 \le \sigma < s$:*

$$E||\tilde{f}_{\lambda_n} - f_0||^2_{B^\sigma_{22}} \le cn^{\frac{2(\sigma-s)}{2s+1}}$$

So the smoothed estimator still achieves the minimax rate of convergence for B^σ_{22} losses for $\sigma < s$.

This method provides a consistent estimator but the main drawback is that it can not be turned into an adaptive method without estimating the smoothness coefficient s of the space where the unknown function lies. As a matter of fact, the choice of the optimal smoothness coefficient depends on the unknown quantity s. Yet we can hope to choose the regularization parameter by a data-depending technique using cross-validation. For every parameter λ_n we associate the estimator $\tilde{f}_{\lambda_n}$ which is consistent as soon as $\lambda_n \to 0$. We already know that there exists an optimal choice of the parameter that minimizes over λ_n

$$E||\tilde{f}_{\lambda_n} - f_0||^2 = \int \tilde{f}^2_{\lambda_n} - 2\int f\tilde{f}_{\lambda_n} + \int f^2$$

So cross-validation minimizes an empirical version of that criterium:

$$F(\lambda_n) = \int \tilde{f}^2_{\lambda_n} - \frac{2}{n}\sum_{i=1}^{n} \tilde{f}^{(i)}_{\lambda_n}(X_i)$$

where $\tilde{f}^{(i)}_{\lambda_n}$ is constructed using coefficients $\hat{\beta}^{(i)}_{jk} = \frac{1}{n-1}\sum_{k\neq i}\psi_{jk}(X_k)$ using the cross-validation method described by G. Wahba in [Wah90] or in [GW91]. Here again, a proof, similiar to the proof of theorem (3.1.16) in the regression model, leads to the existence of a minimizer of the cross-validation criterium that can be chosen to construct the estimator.

But this method is a regularization technique with no theoretical links with the estimation problem. We should use an another methodology and consider as estimator the penalized log-likelihood estimator.

2.3.2 M-estimation of the density

The settings of the estimation problem is the same as in the first part: we observe n independent realizations of variables $X_1, \dots, X_n$ with unknown density with respect to Lebesgue measure λ, $f_0 = \frac{d\mathbf{P}}{d\lambda}$ lying in a set of functions $\mathcal{F}$ with value in $[0, 1]$. The log-likelihood estimator has been studied in the literature, see for instance Stone in [Sto90], Silverman in [Sil82] or Barron and Sheu in [BS91] who considered this issue using log-spline basis.

For a given penalty I, define $\tilde{f}_n$ as

$$\tilde{f}_n = \arg\max_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^{n} \log f(X_i) - \lambda_n^2 I(f) \right). \qquad (2.3.3)$$

In these papers, it is stated that it is more convenient to assume some regularity properties over the logarithm of the density in order to gain positivity of the estimated density. So we define:

$$\gamma_0 = \log(f_0) + b(\gamma_0),$$

with $b(\gamma_0) = -\int \log(f_0)d\mathbf{P}$. For every density $f \in \mathcal{F}$, we consider the variable $\gamma = \log f + b(\gamma)$ lying in the correspondent functional class Γ. Since the density integrates to one, we have the useful relation

$$b(\gamma) = \log \int e^{\gamma(x)} d\lambda(x)$$

Now we can define the maximum penalized likelihood estimator for a penalty function $I : \mathcal{F} \to \mathbb{R}^+$ or alternatively $J : \Gamma \to \mathbb{R}^+$. The penalized maximum log-likelihood estimator (2.3.3) can be written, using the change of notations and the substitution of the penalty as

$$\hat{\gamma}_n = \arg\max_{\gamma\in\Gamma} \left(\frac{1}{n}\sum_{i=1}^{n} \gamma(X_i) - b(\gamma) - \lambda_n^2 J(\gamma) \right). \qquad (2.3.4)$$

We will penalize the roughness of the log-density and use a pseudo-norm. Another direction would be the use of sieves and a penalty over the dimension of the approximating space. Such survey can be found in a paper from G. Castellan in [Cas00]. This point of view is part of model selection theory developed by L. Birgé and P. Massart in [BM97] for instance.

Definition 2.3.2. *Given a smoothing parameter $\lambda_n^2 \to 0$, consider the following penalized M-estimator*

$$\hat{\gamma}_n = \arg\max_{\gamma\in\Gamma} \left(\frac{1}{n}\sum_{i=1}^{n} \gamma(X_i) - b(\gamma) - \lambda_n^2 J(\gamma) \right)$$

where $\Gamma = \{\gamma \in B^s_{p,\infty},\ s > \frac{1}{p}\}$.

This estimator has been studied in a general way by S. van de Geer in [vdG00], but the result was not adaptive. We want to investigate for a precise choice of penalty. The starting point is an inequality that can be found in [vdG00] which links the L^2-loss of the estimator and the true function with the empirical process $\int(\hat{\gamma}_n - \gamma_0)dP_n$ and proves consistency of penalized maximum likelihood. We recall here that the entropy of a Besov space has been calculated by L. Birgé and P. Massart in [BM97]. So there exists a finite constant A such that

$$H_B(\Gamma, \delta, \mathbf{P}) \leq H_\infty(\Gamma, 2\delta/2) \leq A(\frac{2M}{\delta})^{\frac{1}{s}}.$$

Using the definition of the penalized M-estimator we have the following lemma

Lemma 2.3.3.

$$b(\hat{\gamma}_n) - b(\gamma_0) + \lambda_n^2 J(\hat{\gamma}_n) \leq \int (\hat{\gamma}_n - \gamma_0) d(P_n - \mathbf{P}) + \lambda_n^2 J(\gamma_0).$$

Consider a wavelet basis (ψ_{jk}) of Γ with respect to Lebesgue measure, with enough regularity and decompose the log-density onto this basis $\gamma = \sum_{(j,k)} \beta_{jk}\psi_{jk}$. Now we consider the following penalty for the penalization term:

$$J(\gamma) = \sum_{(j,k)} |\beta_{jk}|.$$

This penalty is related to the norm as can be easily shown by the following lemma.

Lemma 2.3.4. $\forall \gamma \in \Gamma$:

$$||\gamma - \gamma_0|| \leq J(\gamma - \gamma_0).$$

With that choice of penalty, we can prove the following theorem describing the asymptotic behaviour of the estimator, using ideas of the proof of the same result in the regression model, proof which can be found in [LvdG00]:

Theorem 2.3.5. *Under the following condition:* $\exists 0 < C < \infty$, $\sup_{\gamma\in\Gamma} |\gamma| \leq C$*, the penalized log-likelihood estimator defined for* $\gamma_0 \in B^s_{p\infty}$ *as*

$$\hat{\gamma}_n = \arg \max_{\gamma = \sum_{(j,k)} \beta_{jk}\psi_{jk} \in \Gamma} \left(\frac{1}{n} \sum_{i=1}^{n} \gamma(X_i) - b(\gamma) - \lambda_n^2 \sum_{(j,k)} |\beta_{jk}| \right),$$

is pseudo-adaptive over the Besov class of functions $\{B^s_{p\infty},\ s > 1/p\}$ *since it is convergent at the minimax rate of convergence up to a logarithmic factor for a quadratic loss.*

Remark 2.3.6. *The condition* $\sup_{\gamma\in\Gamma} |\gamma| \leq C$*, is close to the usual condition in log-likelihood estimation* $\exists \eta_0 > 0,\ f = \frac{d\mathbf{P}}{d\lambda} \geq \eta_0^2$

To conclude, both methods have their drawbacks and advantages. The maximum penalized likelihood method extend to density estimation the theory of M-estimation using soft-thresholding penalties. If it provides an adaptive estimator that achieves the asymptotic minimax rate, it will be important to look for good computer algorithms to solve the optimization problem, whereas the first method can be easily implemented since the estimators are all well defined and depend linearly on the data. But the smoothing coefficient must be chosen using cross-validation and is an approximation of the optimal theoretical parameter.

2.3.3 Appendix

Proof of theorem 2.3.1

Proof. The proof of the consistency follows the same lines as in the case of the regression. Indeed we have for a B^{σ}_{22}-loss the bias variance decomposition:

$$\begin{aligned}
E||\tilde{f}_{\lambda_n} - f||^2_{B^{\sigma}_{22}} &= E||\tilde{f}_{\lambda_n} - E(\tilde{f}_{\lambda_n})||^2_{B^{\sigma}_{22}} + ||E(\tilde{f}_{\lambda_n}) - f|||^2_{B^{\sigma}_{22}} \\
&= \frac{1}{n}\sum_k \left(\int f\psi^2_{j_0k} - \alpha^2_{j_0k}\right) + \frac{1}{n}\sum_{j=j_0}^{j_1} \frac{2^{2j\sigma}}{(1+\lambda_n^2 2^{2js})^2} \sum_k \left(\int \psi^2_{jk} f - \beta^2_{jk}\right) \\
&+ \sum_{j_0}^{j_1} 2^{2j\sigma} \frac{\lambda_n^4 2^{4js}}{(1+\lambda_n^2 2^{2js})^2} \sum_k \beta^2_{jk} + \sum_{j>j_1} 2^{2j\sigma} \sum_k \beta^2_{jk} \\
&\leq \underbrace{\sum_{j>j_1} 2^{2j\sigma} \sum_k \beta^2_{jk}}_{(I)} + \underbrace{\sum_{j=j_0}^{j_1} 2^{2j\sigma} \frac{\lambda^4 2^{4js}}{(1+\lambda^2 2^{2js})^2} \sum_k \beta^2_{jk}}_{(II)} \\
&+ \underbrace{\frac{1}{n}\int (\sum_k \phi^2_{j_0k}(x)) f(x)dx}_{(III)} \\
&+ \underbrace{\frac{1}{n}\sum_{j_0}^{j_1} \frac{2^{2j\sigma}\lambda^4 2^{4js}}{(1+\lambda^2 2^{2js})^2} \int (\sum_k \psi^2_{jk}(x) f(x) dx)}_{(IV)}.
\end{aligned}$$

Reasoning by analogy and using the same arguments as in Chapter III, the following majorations hold:

$$\begin{aligned}
(I) &\leq 2^{2(\sigma-s)} 2^{2j_1(\sigma-s)} \sum_{j>j_1} 2^{2js} \sum_k \beta^2_{jk} \\
&\leq C 2^{2j_1(\sigma-s)} ||f^2||^2_{B^s_{22}} \\
(II) &\leq \lambda^{2\frac{s-\sigma}{s}} \sum_{j=j_0}^{j_1} 2^{2js} \sum_k \beta^2_{jk} \\
&\leq \lambda^{2\frac{s-\sigma}{s}} ||f^2||^2_{B^s_{22}} \\
(III) &\leq K\frac{2^{j_0}}{n} \\
(IV) &\leq \frac{K}{n}\sum_{j=j_0}^{j_1} \frac{2^{j(1+2\sigma)}}{(1+\lambda^2 2^{2js})^2},
\end{aligned}$$

where K is a constant depending only on $||f||_\infty$.
So we have the final majoration: there exists a finite constant C such that

$$E||\tilde{f}_\lambda - f||^2_{B^\sigma_{22}} \leq C(\frac{2^{j_0}}{n} + \frac{2^{j_1(1+2\sigma)}}{n} + 2^{-2j_1(s-\sigma)} + \lambda_n^{2\frac{s-\sigma}{s}}),$$

so provided we have the make the assumptions that $2^{j_0} = o(n)$ and that $2^{j_1} = o(n^{\frac{1}{1+2s}})$, then

$$E||\tilde{f}_\lambda - f||^2_{B^\sigma_{22}} = o(1).$$

Moreover the majoration proves the asymptotic minimax rate. □

Proof of lemma 2.3.3

Proof. The proof comes from the definition of $\hat{\gamma}_n$ and uses the fact that γ is, by construction, a centered variable.

$$\begin{aligned}
\int \hat{\gamma}_n dP_n - b(\hat{\gamma}_n) - \lambda_n^2 J(\hat{\gamma}_n) &\geq \int \gamma dP_n - b(\gamma) - \lambda_n^2 J(\gamma) \\
\int \hat{\gamma}_n dP_n - b(\hat{\gamma}_n) - \lambda_n^2 J(\hat{\gamma}_n) &\geq \int \gamma_0 dP_n - b(\gamma_0) - \lambda_n^2 J(\gamma_0) \\
\int (\hat{\gamma}_n - \gamma_0) dP_n + \lambda_n^2 J(\gamma_0) &\geq b(\hat{\gamma}_n) - b(\gamma_0) + \lambda_n^2 J(\hat{\gamma}_n) \\
\int (\hat{\gamma}_n - \gamma_0) d(P_n - \mathbf{P}) + \lambda_n^2 J(\gamma_0) &\geq b(\hat{\gamma}_n) - b(\gamma_0) + \lambda_n^2 J(\hat{\gamma}_n)
\end{aligned}$$

□

Proof of lemma 2.3.4

Proof. By orthonormality of the basis we can write for $m = (j, k)$ an index:

$$||\gamma - \gamma_0|| = (\sum_m |\beta_m - \beta_m^0|^2)^{1/2}.$$

But we have

$$\sum_m |\beta_m - \beta_m^0|^2 \leq (\sum_m |\beta_m - \beta_m^0|)^2$$

which concludes the proof. □

Proof of theorem 2.3.5

Proof. The following majoration stands for all $\gamma \in \Gamma$,

$$\begin{aligned} |\int (\gamma - \gamma_0) d(P_n - \mathbf{P})| &= |\int \sum_m (\beta_m - \beta_m^0) \psi_m d(P_n - \mathbf{P})| \\ &\leq \sup_m \left| \int \psi_m d(P_n - P) \right| J(\gamma - \gamma_0). \end{aligned}$$

We must derive a concentration inequality over $\left|\int \psi_m d(P_n - \mathbf{P})\right|$, from a Bernstein type inequality.

$$\mathbf{P}\left(|\int \psi_m d(P_n - P)| > T_n \right) \leq 2 \exp(-\frac{nT_n^2}{2(\sigma^2 + 3/2||Y||_\infty)})$$

where $Y_i = \psi_m(X_i) - E(\psi_m(X_i))$ are independent random variables with zero mean.

$$\sigma^2 \leq ||f_0||_\infty$$

$$||Y||_\infty \leq 2^{j/2} M.$$

So Bernstein inequality can be written in the following way: there exists a finite constant A,

$$\mathbf{P}(\sup_m |\int \psi_m d(P_n - \mathbf{P})| > T_n) \leq 2 \exp(-A((nT_n^2) \wedge (nT_n))).$$

Using this inequality we obtain:

$$\begin{aligned} \mathbf{P}(\sum_{m \in \Lambda} |\int \psi_m d(P_n - \mathbf{P})| > T_n) &\leq \sum_{m \in \Lambda} 2 \exp(-AnT_n^2) \\ &\leq |\Lambda| 2 \exp(-AnT_n^2). \end{aligned}$$

If we choose $T_n = c\sqrt{\frac{\log n}{n}}$ then if the set of indices Λ is polynomial in n, for c large enough we have

$$\mathbf{P}(\sup_{m \in \Lambda} |\int |\psi_m| d(P_n - \mathbf{P})| > T_n) \leq 2 \frac{|\Lambda|}{n^{Ac^2}} \rightarrow O.$$

Now recall our model: we consider a wavelet basis $m = (j, k)$ and we begin to approximate the log-density by its projection onto the space V_{j_1} for a choice of j_1 that will be precised later on. Moreover, we assume that the log-density belongs to a Besov space $B_{p\infty}^s$ and is bounded in the supremum norm. As a result,

$$\begin{aligned} \mathbf{P}(\sup_{0 \leq j \leq j_1, k} |\int \psi_{jk} d(P_n - \mathbf{P})| \geq T_n) &\leq \sum_{j=0}^{j_1} \sum_k \mathbf{P}(|\int \psi_{jk} d(P_n - \mathbf{P})| \geq T_n) \\ &\leq \sum_{j=0}^{j_1} \sum_k 2 \exp(-A_j((nT_n^2) \wedge (nT_n))) \\ &\leq 2 \sum_{j=0}^{j_1} 2^j \exp(-AnT_n^2) \end{aligned}$$

for a choice of T_n and j_1 such that

$$T_n = c\sqrt{\frac{\log n}{n}}$$

$$2^{j_1} \leq 1/c\sqrt{\frac{n}{\log n}},$$

we have

$$\mathbf{P}(\sup_{0 \leq j \leq j_1, k} | \int \psi_{jk} d(P_n - \mathbf{P})| \geq T_n) \leq 2 2^{j_1} \exp(-AnT_n^2)$$

$$\leq 2/c \frac{n^{1/2 - Ac^2}}{\sqrt{\log n}}.$$

As soon as we have chosen c large enough, the last quantity tends to zero as n increases. The condition over the choice of the constant c can be written as:

$$c^2 \geq \max(||f_0||_\infty, 2/3||\psi||_\infty).$$

Then on an event of probability one we can write that for every $\lambda_n^2 \geq c\sqrt{\frac{\log n}{n}}$ we have

$$\sup_{(j,k) \in \Lambda} |\psi_{jk} d(P_n - \mathbf{P})| \leq \lambda_n^2$$

If we set γ_1 the projection of γ on V_{j_1}, we have:

$$||\hat{\gamma}_n - \gamma_0|| \leq ||\hat{\gamma}_n - \gamma_1|| + ||\gamma_1 - \gamma_0||.$$

From the property of the wavelet basis and the choice of the level j_1 we have, since $\gamma_0 \in B_{p\infty}^s$:

$$||\gamma_1 - \gamma_0|| \leq 2^{-j_1 s} \leq \left(\frac{n}{\log n}\right)^{-s/2}.$$

Using consistency of the estimator and a Taylor's expansion of $b(\gamma)$ we get:

$$b(\hat{\gamma}_n) - b(\gamma_0) = E\|\hat{\gamma}_n(X_1) - \gamma_0(X_1)\|^2/(1 + O(1))$$

but since $\frac{d\mathbf{P}}{d\lambda} \geq \eta^2$, we have

$$\frac{||\hat{\gamma}_n - \gamma_0||^2}{1 + O_{\mathbf{P}}(1)} + \lambda_n^2 J(\hat{\gamma}_n) \leq \int (\hat{\gamma}_n - \gamma_0) dP_n + \lambda_n^2 J(\gamma_0). \tag{2.3.5}$$

As a result, for the stochastic term, we have the following inequality:

$$\frac{||\hat{\gamma}_n - \gamma_1||^2}{1 + O_{\mathbf{P}}(1)} + \lambda_n^2 J(\hat{\gamma}_n) \leq \lambda_n^2 J(\hat{\gamma}_n - \gamma_1) + \lambda_n^2 J(\gamma_1).$$

So we obtain, following the same guideline than in the proof of the rate of convergence of the penalized least squares estimator in [LvdG00]:

$$||\hat{\gamma}_n - \gamma_1|| \leq \lambda_n^2 N_n^{1/2} + \lambda_n J_M^{1/2} \leq \left(\frac{\log n}{n}\right)^{\frac{s}{2s+1}}$$

And by comparison of the two rates of convergence, we have:

$$||\hat{\gamma}_n - \gamma_0|| \leq \left(\frac{\log n}{n}\right)^{\frac{s}{2s+1}}$$

□

2.4 Appendix

2.4.1 Some extensions for adaptive M-estimation

Adaptive M-estimation with l^1 penalty

We refer to the article of Loubes and van de Geer [LvdG00] We make the comment that we have obtained an adaptive estimator up to logarithmic factors via a penalized least absolute deviation estimation method. The only condition on the errors is that they must be enough concentrated around zero. This assumption is not too restrictive since, if there exists a continuous density of the errors f, then as soon as $f(0) \neq 0$, the condition will be fulfilled, which is the case in most of common laws of probability errors and we have

$$\exists \epsilon > 0,\ \forall 0 \leq a \leq \epsilon,\ \mathbf{P}(0 \leq W \leq a) \geq \epsilon a.$$

This constant can be approximated easily: indeed for a small enough,

$$\mathbf{P}(0 \leq W \leq a) \approx a f(0)$$

So take $\epsilon \approx f(0)$.

Extension to other error measures

We want to prove consistency of our penalized estimator in the integrated norm $||.||$. For this, we will use the following theorem due to S. van de Geer in [vdG00], that enables us to make a comparison between the norm and the empirical norm. This result needs a more refined control over the class of functions than the standard entropy: the bracketing entropy.

Definition 2.4.1. *For a class of functions Θ, consider, for $\delta > 0$ a collection of couples of functions $(\theta_{L,j}, \theta_{L,j})$, $j = 1, \ldots, N$ such that, for every $\theta \in \Theta$, there exists a couple in the previous collection such that*

$$\theta_L \leq \theta \leq \theta_U$$
$$\|\theta_U - \theta_L\|_n \leq \delta$$

Define $N_B(\delta, \Theta)$ the smallest N for which such δ-covering happens. The bracketing entropy is defined as

$$H_B(\delta, \Theta) = \log N_B(\delta, \Theta)$$

This quantity is defined by S. van de Geer in [vdG00] for instance.

Theorem 2.4.2. *If there exists a sequence δ_n, such that $n\delta_n \to \infty$ and $H_B(\delta_n, \Theta, \mathbf{P}) \leq n\delta_n^2$, then for all $0 < \eta < 1$, the following inequalities stands:*

$$\limsup \mathbf{P}\left(\sup_{\theta \in \Theta, ||\theta|| > 2^5 \delta_n / \eta} |\frac{||\theta||_n}{||\theta||} - 1| \geq \eta \right) = 0$$

$$\limsup \mathbf{P}\left(\sup_{\theta \in \Theta, ||\theta|| < 2^5 \delta_n / \eta} |||\theta||_n - ||\theta|| \geq 2 \right) = 0.$$

Proof. The proof is given in the appendix. It relies on empirical processes techniques and the use of the peeling device. □

Recall that we consider a M-estimator defined by:

$$\hat{\theta}_n = \arg \min_{\theta=\sum_{(j,k)} \beta_{jk}\psi_{jk}} \left(||Y-\theta||_n^2 + \lambda_n^2 \sum_{(j,k)} |\beta_{jk}| \right).$$

We have already proven its consistency for the empirical norm and we want to study the quadratic error of this estimator. We make the additional assumption that

$$I(\hat{\theta}_n) < \infty \tag{2.4.1}$$

so that $\hat{\theta}_n \in \Theta$. So we have two possibilities for $||\hat{\theta}_n - \theta_0||_2$:

- If $||\hat{\theta}_n - \theta_0|| \leq 2^5 n^{\frac{-s}{2s+1}}/\eta$, then the penalized M-estimator converges with a good rate convergence towards the true function θ_0.
- If $||\hat{\theta}_n - \theta_0|| > 2^5 n^{\frac{-s}{2s+1}}/\eta$, then we can apply the following theorem with $\delta_n = n^{-\frac{s}{2s+1}}$. Since there exists a finite constant A so that

  $$H_B(\delta_n, B^s_{p\infty}) \leq A\delta_n^{-1/s},$$

  the hypothesis are fulfilled so we have the following inequality on an event with probability equal to one:

  $$|\frac{||\hat{\theta}_n - \theta_0||_n}{||\hat{\theta}_n - \theta_0||} - 1| \leq \eta,$$

  which gives

  $$||\hat{\theta}_n - \theta_0|| \leq \frac{1}{1-\eta}||\hat{\theta}_n - \theta_0||_n$$

  but since $||\hat{\theta}_n - \theta_0||_n = O_{\mathbf{P}}(n^{-\frac{s}{2s+1}})$, we can write

  $$||\hat{\theta}_n - \theta_0|| \leq Cn^{-\frac{s}{2s+1}}.$$

  So, also in that case, the penalized M-estimator converges at the rate of convergence $n^{-\frac{s}{2s+1}}$.

Remark 2.4.3. *The main assumption $I(\hat{\theta}_n) < \infty$ can not be removed from the proof. It may be hard to prove it. The only control we can obtain over this quantity is a very rough majoration which is proven in the appendix*

$$I(\hat{\theta}_n) = O_{\mathbf{P}}(\lambda_n^{-2} \log n) \tag{2.4.2}$$

Since consistency requires $\lambda_n^2 \geq \sqrt{\frac{n}{\log n}}$, this majoration does not help proving the hypothesis (2.4.1).

Generalization to classical M-estimators

We give here an application to various robust loss functions, used in M-estimation, which have been described by Berlinet in [BLV00] in the case where the errors $W_i, i = 1, \ldots, n$ are identically independently distributed. Huber and Hampel started with a systematic treatment of such robust estimators. There are three functions mainly used. In order to extend our general theorem to these specific cases, we have to check if the additional condition is fulfilled, that is if there exists $\epsilon > 0$ so that for $t \leq \epsilon$, the following inequality holds

$$\int \left(\gamma_{\theta_t} - \gamma_{\theta_0}\right) d(P_n - \mathbf{P}) \geq \epsilon ||\theta_t - \theta_0||_n^2$$

which is a direct consequence of the next condition: there exists $\epsilon > 0$ such that $\forall 0 \leq a \leq \epsilon$,

$$\min_i E\left(\gamma(W_i + a) - \gamma(W_i)\right) \geq \epsilon a^2.$$

1. The α-quantile loss function, defined by Hampel in [Ham83] is given by: $\forall \alpha \in (0, 1)$

$$\rho(x) = \begin{cases} \alpha x & \text{if } x \geq 0 \\ (\alpha - 1)x & \text{if } x < 0 \end{cases}$$

   By direct calculations, we can find that

$$E(\gamma(W + a) - \gamma(W)) \geq \alpha a \mathbf{P}(W \geq 0),$$

   So provided the errors verify $\mathbf{P}(W \geq 0) \geq a$ for $a \leq \epsilon$, the theorem can be applied.

2. The Vajda loss-function, cf Vajda in [Vaj99], is defined by, $\forall \alpha > 0$

$$\rho(x) = 1 - \exp(-\alpha x^2).$$

   Such a function puts the stress on small coefficients since if x is small $\rho(x) \approx \alpha x^2$ and if x is large $\rho(x) \approx 1$, so the estimator tends to select an estimator very close to the data with a smoothing effect.
   If the errors are Gaussian, centered with variance σ^2, we find that

$$E((W + a)^2 - W^2) \geq \alpha a^2$$

   as soon as $\alpha \sigma^2 \leq \frac{3}{2}$. For general errors, we obtain a similar condition about the concentration of the errors around zero:

$$\forall 0 \leq \alpha \leq \epsilon, \; E\left((W + a)\exp(-\alpha(W + a)^2)\right) \geq \epsilon.$$

3. The Huber loss-function described by Huber in [Hub81] is an intermediate between the quadratic loss and the l^1 loss. The loss-function behaves like a quadratic function

for small coefficients and like a l^1 function for large coefficients. Set $\tau > 0$ a cut-off parameter, and define

$$\rho(x) = \begin{cases} \frac{x^2}{2} & \text{if } |x| \le \tau \\ \tau|x| - \frac{\tau^2}{2} & \text{if } |x| > \tau. \end{cases}$$

The Huber M-estimator is consistent with the rate of convergence given by the general theorem provide the following hypothesis hold: there exists $\epsilon > 0$ and a constant $c < \infty$ so that $\forall 0 \le a < \epsilon$,

$$\begin{aligned} \mathbf{P}(|W + a| > \tau) &\ge 1 - \epsilon \\ \tau^2 \mathbf{P}(|W| > \tau) &\ge E(W^2 1_{|W|>\tau}) - \epsilon \\ E((W + a)^2 1_{|W+a|\le\tau}) &\ge 4\epsilon + ca^2. \end{aligned}$$

appendix

Proof of theorem 2.4.2

Proof. We recall briefly the main ideas of the proof. We consider a δ_n-bracktened covering of Θ by functions (θ_L, θ_U). Then for $\theta \in \Theta$, we have $\sup_{\theta\in\Theta} ||\theta||_\infty \le M$, we use the peeling device to write that for some integer s, we have $M(s-1)\delta_n \le ||\theta|| \le Ms\delta_n$, so since $||\theta||_n^2 \le ||\theta_U||_n^2 + ||\theta_U - \theta_L||_n^2$, we can write:

$$\mathbf{P}\left(\frac{||\theta||_n}{||\theta||} \ge 1 + \eta\right) \le \mathbf{P}\left(||\theta_U||_n^2 - ||\theta_U||^2 + ||\theta_U - \theta_L||_n^2 \ge \eta s \delta_n^2\right).$$

Now we can use a Berntein-type inequality:

$$\forall a > 0, \mathbf{P}\left(|||\theta||_n^2 - ||\theta||^2| > a\right) \le 2\exp(-\frac{na^2}{8(a + ||\theta||^2)}).$$

$$\begin{aligned} &\mathbf{P}(\sup_{\theta\in\Theta, ||\theta||>2^5\delta_n/\eta} \frac{||\theta||_n}{||\theta||} \ge 1 + \eta) \\ &\le \sum_{s \ge 2^5/\eta} 4\exp(H_B(\delta_n, \Theta, \mathbf{P}) - \frac{n^2\delta_n^2 s^2 \eta^2}{2^8}) \\ &\le 4 \sum_{s\ge 2^5/\eta} \exp(-\frac{n\delta_n^2 s^2\eta^2}{2^7}) \overset{n\to\infty}{\longrightarrow} 0\ . \end{aligned}$$

So we can conclude that

$$\lim_{n\longrightarrow\infty} \mathbf{P}\left(\sup_{\theta\in\Theta, ||\theta||>2^5\delta_n/\eta} \frac{||\theta||_n}{||\theta||} \ge 1 + \eta\right) = 0$$

We can prove in a similar way that

$$\lim_{n\longrightarrow\infty} \mathbf{P}\left(\sup_{\theta\in\Theta,||\theta||>2^5\delta_n/\eta} \frac{||\theta||_n}{||\theta||} \leq 1-\eta\right) = 0$$

which concludes the proof. □

Proof of assertion (2.4.2):

Proof. The penalized least squares estimator is given by:

$$\hat{\theta}_n = \arg\inf_{\theta\in\Theta} \left(||Y-\theta||_n^2 + \lambda_n^2 I(\theta)\right)$$

where the penalty I is positive and equal to zero for the null function.
By the definition of the M-estimator we know that for every $\theta \in \Theta$ we have

$$||Y-\hat{\theta}_n||_n^2 + \lambda_n^2 I(\hat{\theta}_n) \leq ||Y-\theta||_n^2 + \lambda_n^2 I(\theta).$$

So for the function $\theta = 0$ and by positivity of the loss-function the inequality becomes:

$$0 \leq I(\hat{\theta}_n) \leq \lambda_n^{-2}||Y||_n^2.$$

But $||Y||_n \leq ||\theta||_0 + \max_{i=1,\dots,n} ||W_i||_\infty$. With the following lemma, we can control the behavior of $||W_i||_\infty$, $i = 1, \dots, n$..

Lemma 2.4.4. *Let W_i independent random variables centered with variance σ^2 and, for all positive real s: $E\exp(s|W_i|) \leq \exp(\frac{s^2\sigma^2}{2})$. Then we have for all sequence T_n:*

$$\mathbf{P}(\max_{i=1,\dots,n} |W_i| \geq T_n) \leq \sigma\frac{\sqrt{2\log n}}{T_n}$$

So the choice of $T_n = \log n$ leads to

$$\max_{i=1,\dots,n} |W_i| \leq \log n \text{ a.s}$$

With this majoration we can conclude that

$$I(\hat{\theta}_n) = O_{\mathbf{P}}(\lambda_n^{-2}\log n).$$

□

Proof of Lemma 2.4.4

Proof.

$$\begin{aligned}\exp(sE(\max_{i=1,\ldots,n}|W_i|)) &\leq E\exp(s\max_{i=1,\ldots,n}|W_i|)\\ &\leq E\max_{i=1,\ldots,n}\exp(s|W_i|)\\ &\leq \sum_{i=1}^{n}E\exp(s|W_i|)\\ &\leq n\exp(\frac{s^2\sigma^2}{2}).\end{aligned}$$

So we obtain

$$sE\max_{i=1,\ldots,n}|W_i| \leq \log(n)+\frac{s^2\sigma^2}{2},$$

and for a value of $s^2=\frac{2\log n}{\sigma^2}$ we have for $T_n=\log n$:

$$\mathbf{P}(\max_{i=1,\ldots,n}|W_i|\geq \log n)\leq 2\sqrt{\frac{2}{\log n}}$$

and so infinitely often we have

$$\max_{i=1,\ldots,n}|W_i|\leq \log n.$$

□

2.4.2 Asymptotic Distribution of Penalized M-Estimators

Testing Problems with M-estimators

After having studied the asymptotic behavior of M-estimators, we would like to apply the results of preceding sections to find the asymptotic law of the estimator. This problem is difficult in the general case, since we do not have a direct expression of the estimator which is only defined as the solution of an optimization problem. That is why we focus on various testing situations for the model using the M-estimator.

A first testing problem is to test whether the observations in our model come from an unknown real function or if we only observe a pure white noise model. That is the reason why we want to test the null hypothesis: $H_0 : \theta_0 = 0$. For this we will consider a test statistic based on $\hat{\theta}_n$, which is defined by:

$$\hat{\theta}_n = \arg \inf_{\theta \in \Theta} \left(||Y - \theta||_\gamma \right), \tag{2.4.3}$$

where

$$||Y - \theta||_\gamma = \frac{1}{n} \sum_{i=1}^{n} \gamma(\theta_i - Y_i)$$

for γ a loss-function, convex in θ. Under some conditions on the regularity of the class Θ, i.e under good entropy bounds, and for a good choice of the smoothing parameter λ_n^2, the solution of the optimization problem $\hat{\theta}_n$ converges to the true parameter $\theta_0 = \arg\min_{\theta \in \Theta} E||Y - \theta||_\gamma$. So if H_0 is true, then the above quantity is likely to be minimized at zero, therefore we want to find a constant, c_α, such that the null hypothesis is rejected at a level α if and only if $||\hat{\theta}_n||_\gamma > c_\alpha$. If the estimator is significant, the true function is likely to be a non zero function. So this means that c_α satisfies

$$\begin{cases} \mathbf{P}(||\hat{\theta}_n||_\gamma \leq c_\alpha) & = 1 - \alpha, \quad \text{or} \\ \mathbf{P}(||\hat{\theta}_n||_\gamma > c_\alpha) & = \alpha \end{cases}$$

We want to show that this constant can be calculated using a concentration inequality in both cases, either when the estimator is a penalized least squares estimator, or a penalized least-absolute deviation estimator.

1. Penalized Least Squares Estimator:
   Here the loss function is quadratic $\gamma(x) = x^2$, so we have $||\theta||_\gamma = ||\theta||_n^2 = \frac{1}{n} \sum_{i=1}^{n} \theta(z_i)^2$. The M-estimator is defined as

   $$\hat{\theta}_n = \arg \inf_{\theta \in \Theta} \left(\frac{1}{n} \sum_i |Y_i - \theta(z_i)|^2 \right).$$

   So, using the mere definition of the estimator, we can write that:

   $$||Y - \hat{\theta}_n||_n^2 \leq ||Y||_n^2.$$

Now, by the definition of the empirical norm $||.||_n$ and the discrete scalar product

$$< f, g >_n = \frac{1}{n}\sum_{i=1}^{n} f(z_i)g(z_i)$$

we have:

$$||\hat{\theta}_n||_n^2 = ||Y - \hat{\theta}_n||_n^2 - ||Y||_n^2 + 2 < Y, \hat{\theta}_n >_n .$$

So the following majoration stands:

$$\begin{aligned} \mathbf{P}(||\hat{\theta}_n||_n^2 > c_\alpha^2) &\leq \mathbf{P}(2 < \hat{\theta}_n, Y >_n > c_\alpha^2) \\ &\leq \mathbf{P}(\sup_{\theta\in\Theta} (2 < Y, \theta >_n) > c_\alpha^2) \\ &\leq \mathbf{P}(\sup_{\theta\in\Theta} < \theta, Y >_n > \frac{c_\alpha^2}{2}). \end{aligned}$$

We recognize a deviation inequality. Provided there exists a control over the entropy of the set Θ, a majoration of this quantity can be given. It depends on the behavior of the underlying empirical process. For instance, if the class of functions Θ is such that:

- there exists R, a positive constant such that $\sup_{\theta\in\Theta} ||\theta||_n \leq R$
- there exists $s > 1/2$ and a constant C such that, for all positive δ

$$H(\delta, \Theta, P_n) \leq C\delta^{-\frac{1}{s}}$$

then the following property describes the concentration of the empirical process::

Proprosition 2.4.5.

$$\forall\delta > 0,\ \sqrt{n}\delta > C\left(\int_0^R H^{1/2}(u, \Theta, P_n)du \vee R\right)$$

$$\mathbf{P}(\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\theta(z_i)W_i\right| \geq \delta) \leq C\exp(-\frac{n\delta^2}{R^2C^2}).$$

Proof. The proof of this result comes from a starting Hoeffding-type inequality and a typical use of the chaining device to transform the probability of a supremum over an infinite set into the supremum over a finite set. It can be found in [vdG00]. □

As a result, under such assumptions, the following majoration holds:

$$\mathbf{P}(\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\theta(z_i)W_i\right| \geq \frac{c_\alpha^2}{2}) \leq C\exp(-\frac{nc_\alpha^4}{4R^2C^2}).$$

So if we choose properly the constant c_α

$$c_\alpha^2 \approx \frac{2CR}{\sqrt{n}}\sqrt{\log(\frac{C}{\alpha})}$$

then provided

$$c_\alpha^2 \geq \frac{2C}{\sqrt{n}}\frac{2s}{2s-1}R^{1-\frac{1}{2s}} \vee \frac{2CR}{\sqrt{n}}$$

condition which is automatically fulfilled if α is small enough, i.e $\alpha \leq C\exp(\frac{s^2}{C^2(2s-1)^2}R^{-1/s})$, then

$$\mathbf{P}(\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \theta(z_i)W_i\right| \geq \frac{c_\alpha^2}{2}) \leq \alpha.$$

So if we can calculate the constants in the concentration inequality, we can calculate c_α and so reject the assumption of a pure white noise model at a level at least α as soon as $||\hat{\theta}_n|| > c_\alpha$. For this, we look at the starting inequality of the proof. The constants can be majorated by $2\sqrt{2}$, which gives

$$c_\alpha^2 = \frac{1}{\sqrt{n}}4R\sqrt{\log(\frac{2\sqrt{2}}{\alpha})}.$$

Hence we have the following property:

Proproposition 2.4.6. *In a regression model over a class $\Theta = \{\theta, \|\theta\|_n \leq R\}$ of functions whose entropy satisfies $H(\delta, \Theta, P_n) \leq C\delta^{-\frac{1}{s}}$ with $s > 1/2$, then for $c_\alpha^2 = 2CR\frac{1}{\sqrt{n}}\sqrt{\log(\frac{C}{\alpha})}$ the test $||\hat{\theta}_n||_n > c_\alpha$ is a test at level at least α for the pure white noise model.*

Remark 2.4.7. *The concentration inequality provides a majoration of the true constant which may be non optimal. We expect the real constants to be smaller than the critical value we obtain. As a result, the power of the test is expected to be in fact smaller than the power we consider. Moreover, the test is too rough since, using a result from Wegkamp in [Weg], we get $\sqrt{n}\|\theta\|_n^2 \to 0$. So we should find in an optimal test $c_\alpha^2 = o\left(\frac{1}{\sqrt{n}}\right)$ which is not the case.*

Remark 2.4.8. *Using penalization over balls of Sobolev spaces H^s, $s > \frac{1}{2}$ fulfills the assumptions of the previous proposition. This is the case when $\Theta = \{\theta \in H^s, \|\theta\|_n \leq R\}$. Given ψ_j an orthonormal family in $L^2(P_n)$, the assumptions of the proposition are also satisfied if there exists a roughness parameter ρ for which $\Theta = \{\theta = \sum_j \beta_j\psi_j,\ I(\theta) = \sum_j |\beta_j|^\rho < \infty\}$. In this case we can take $\rho = \frac{2}{1+2s}$ as it is done in [LvdG00].*
However, the penalized least squares estimator with soft-thresholding penalty does not fulfill the conditions an can not be used in that way.

2. Penalized Least Absolute Deviation Estimator.
   In that case, consider $\gamma(x) = |x|$ and $||\theta||_\gamma = \frac{1}{n}\sum_{i=1}^n |\theta(z_i)|$, the empirical norm is not quadratic anymore and the M-estimator is defined as

$$\hat{\theta}_n = \arg\inf_{\theta\in\Theta}\left(\frac{1}{n}\sum_i |Y_i - \theta(z_i)|\right) \tag{2.4.4}$$

   We have:

$$\epsilon||\hat{\theta}_n||_n^2 \leq \int(\gamma_{\hat{\theta}_n} - \gamma_0)d\bar{P}, \tag{2.4.5}$$

   where we have set $\gamma_\theta = |Y - \theta(z)|$ and $\bar{P} = \frac{1}{n}\sum_{i=1}^n P^{(i)}$. So we can write:

$$\mathbf{P}(||\hat{\theta}_n||_n^2 > c_\alpha^2) \leq \mathbf{P}\left(\frac{1}{\epsilon}\int(\gamma_{\hat{\theta}_n} - \gamma_0)d\bar{P} > c_\alpha^2\right).$$

   But the M-estimator is defined as a solution of a L^1-minimization problem (2.4.4), so the following majoration holds using the mere definition of the minimizer:

$$\int(\gamma_{\hat{\theta}_n} - \gamma_0)d\bar{P} \leq -\int(\gamma_{\hat{\theta}_n} - \gamma_0)d(P_n - \bar{P}) \tag{2.4.6}$$

   So using the inequality (2.4.5) together with (2.4.6) we can write:

$$\begin{aligned}\mathbf{P}\left(||\hat{\theta}_n||_n^2 > c_\alpha^2\right) &\leq \mathbf{P}\left(|\int(\gamma_{\hat{\theta}_n} - \gamma_0)d(P_n - \mathbf{P})| > \epsilon c_\alpha^2\right)\\ &\leq \mathbf{P}\left(\sup_{\theta\in\Theta}|\int(\gamma_\theta - \gamma_0)d(P_n - \mathbf{P})| > \epsilon c_\alpha^2\right)\end{aligned}$$

   Now if we have a control over the preceding empirical process, we can derive a deviation inequality and so define a bound for the constant c_α in the test problem. Now we recall the following proposition:

Proprosition 2.4.9. *If the class of functions $\Theta = \{\theta = \sum_{jk}\beta_{jk}\psi_{jk}, \sum_{jk}|\beta_{jk}| < \infty\}$, is such that for a constant A, and $\delta > n^{-\frac{1}{2}}/8$*

$$H(\Theta, \delta) \leq A\frac{1}{\delta^2}\log n$$

there exists a universal constant C such that

$$\mathbf{P}\left(\sup_{\theta\in\Theta,\,||\theta-\theta_0||_n\leq 1}|\int(\gamma_\theta - \gamma_{\theta_0})d(P_n - \bar{P})| \geq C\sqrt{\frac{\log^3 n}{n}}\right) \leq C\exp\left(-\frac{\log^3 n}{C^2}\right)$$

Proof. The proof of this result can be found in [vdG90] or [vdG00] and rely as usual on a deviation inequality whose complexity is controlled by an entropy argument. □

So since the hypothesis of the theorem are satisfied, if the constant c_α is chosen such that

$$\mathbf{P}\left(\sup_{||\theta-\theta_0||_n\leq 1} |\int(\gamma_\theta - \gamma_{\theta_0})d(P_n - \bar{P})| \geq c_\alpha^2\right) \leq \alpha$$

we will conclude that $\mathbf{P}(||\hat{\theta}_n|| > c_\alpha) \leq \alpha$.
Take $M = 1$, for the choice

$$c_\alpha^2 \approx \frac{1}{\epsilon}C\sqrt{\frac{\log^3 n}{n}}\sqrt{\log(\frac{C}{\alpha})}$$

the deviation inequality holds and we have built a test at least at level α for testing the white noise model with a statistic based on the Least Absolute Deviation Estimator. However the universal constant C remains to be found. Moreover, here again the real power of the test should be a lot less than α.

Remark 2.4.10. *We could try to test another hypothesis: the no effect alternative. In a non-parametric regression setting, we look at the inference between the response variable and the predictor variable. Such problem is similar to the one investigated by Cox and Koh in [CKWY88] who used a Bayesian approach or by Cox and Wahba in [CK89]. In a more precise way, the authors were interested in testing the null hypothesis H_0: θ is a polynomial function of degree m versus the alternative hypothesis H_1: θ is smooth in the sense that there exists a $m-1$-fold integrated Wiener process Z such that*

$$\theta = P_m + \sqrt{b}Z,$$

where we have set b a finite positive constant and P_m a polynomal with degree m. They showed that there was no uniformly most powerful test for testing H_1 versus H_0. Now if we go back to our setting, the problem can be written: the model is the following

$$Y_i = \theta(z_i) + \beta_0 + \epsilon_i,\ i = 1, \ldots, n.$$

Testing for no effect is equivalent to testing if $\theta = 0$. But if we use a wavelet decomposition, let ψ_{jk} a wavelet with r vanishing moments where s is the regularity of the function and $r \geq s$, we have

$$\theta_0 = \sum_{(j,k)} \beta_{jk}\psi_{jk}.$$

The penalized Least Squares estimator can be written in term of a minimization over the wavelet coefficients, and since $<\beta, \psi_{jk}>= 0$ due to the regularity of the wavelet, the hypothesis is equivalent to testing a pure white noise model for its wavelet representation. As a consequence, the test based on the Penalized Least Squares estimator is also a test at level α for testing a no effect hypothesis. We can point out that Antoniadis, Gijbels and Gregoire in [AGG97] constructed test using a model selection argument, and showed that the two preceding test problems can be extended to the case where we look at a martingale structure in time series using the same type of tests.

Asymptotic distribution of M-estimator

In this part, we will prove asymptotic normality of the distribution separately for each wavelet coefficient of the Penalized Least Square estimator. By the definition of the estimator (4.1.2) minimizing a penalized loss function over the class Θ we have the following equivalent definition:

$$\begin{cases} \forall h \in \Theta,\ \forall t \in [0,1],\ \hat{\theta}_{n,t} = \hat{\theta}_n + th, \\ \frac{d}{dt}|_{t=0} \left(||Y - \hat{\theta}_{n,t}||_n^2 + \lambda_n^2 I(\hat{\theta}_{n,t}) \right) = 0 \end{cases} \tag{2.4.7}$$

If we have set

$$\theta^0 = \sum_{(j,k)} \alpha_{jk}^0 \psi_{jk}$$

$$\forall \theta \in \Theta,\ \theta = \sum_{(j,k)} \alpha_{jk} \psi_{jk}$$

$$(W, h) = \sum_{(j,k)} \alpha_{jk} V_{jk} \quad \text{and}\ V_{jk} = \frac{1}{n} \sum_{i=1}^{n} W_i \psi_{jk}(z_i),$$

using a Taylor expansion of the quantity (2.4.7) we find that,

$$\sqrt{n}(\hat{\alpha}_{jk} - \alpha_{jk}^0) = \frac{\frac{1}{n}\sum_{i=1}^{n} W_i h(z_i) + \sqrt{n}\lambda_n^2 F^{'}((\hat{\alpha}_{jk})_{jk})(< h, \psi_{jk} >) + R_n}{\frac{1}{n}\sum_{i=1}^{n} h(z_i)\psi_{jk}(z_i)},$$

where the residual term R_n is defined as

$$R_n = \sqrt{n} \sum_{((l,m)\neq j,k)} \frac{1}{n} \sum_{i=1}^{n} h(z_i)\psi_{lm}(z_i)(\hat{\alpha}_{lm} - \alpha_{lm}^0),$$

and where we have set

$$F = I(\hat{\theta}_{n,t}).$$

We have to choose a non optimal smoothing coefficient since we make the assumption that:

$$\sqrt{n}\lambda_n^2 \longrightarrow 0 \tag{2.4.8}$$

If we choose in the preceding calculations $h = \psi_{jk}$, we have $\forall (l,m) \neq (j,k)$, $< h, \psi_{l,m} >= 0$, so the residual term is equal to zero and we obtain using the central limit theorem and since $F^{'}$ is bounded that:

$$\sqrt{n}(\hat{\alpha}_{jk} - \alpha_{jk}^0) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\sigma_W^2}{\int \psi_{jk}^2}\right) \tag{2.4.9}$$

Remark 2.4.11. *For soft-thresholding type penalties, we have proved in [LvdG00] that the consistency of the penalized least square estimator requires the condition $\lambda_n^2 \geq \sqrt{\frac{\log n}{n}}$ which contradicts the condition (2.4.8) $\sqrt{n}\lambda_n^2 \to 0$. So such a method does not prove a central limit theorem for soft-thresholded estimator. Recently, some authors such as D. Picard and K.*

Tribouley in [PT00] have investigated this problem and have tried to give asymptotic confidence intervals for such estimators. However it can be used for M-estimator with $l^2 -$ *penalty like an* H^s *or a* B^s_{22} *semi-norm. Indeed consistency only requires* $\lambda_n^2 \geq n^{-\frac{2s}{2s+1}}$ *and the condition* (2.4.8) *gives raise to* $s > \frac{1}{2}$*, which is a common assumption in this framework see for instance [MvdG97].*

Remark 2.4.12. *The proof of the asymptotic convergence in law can not be directly extended to the case of the least absolute deviation estimator since its proof relies on the differentiability of the loss function. However, to prove a central limit theorem, such condition can be weakened and we only have to impose differentiability in quadratic norm. For example, D. Pollard in theorem VII.1.5 of [Pol84], uses stochastic equicontinuity to prove a central limit theorem for M-estimators with general loss functions. In particular, for an* l^1 *loss, the estimator* $\hat{\theta}_n$*, minimizing over the set* Θ *the quantity* $M_n = \frac{1}{n}\sum_{i=1}^n |Y_i - \theta_i|$*, is such that*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \frac{4}{\pi}I)$$

But proving a similar result for the penalized least absolute deviation estimator would require the same condition (2.4.8)*. Again, such an assumption is contradictory to the condition* $\lambda_n^2 \geq \sqrt{\frac{\log^3 n}{n}}$ *necessary for the consistency of the estimator.*

2.4.3 Numerical Approximation of Penalized Adaptive M-Estimator

In this part, we describe the algorithms we used to compute the estimators we have studied. Such estimators are defined as solution of optimization problems. The model we observe is the standard regression model

$$\begin{cases} Y_i = \theta_0(z_i) + W_i,\ i = 1, \dots, n, \\ \theta_0 \in \Theta, \end{cases}$$

where Θ is a functional set. We consider the M-estimator $\hat{\theta}_n$ defined by:

$$\hat{\theta}_n = \arg \inf_{\theta \in \Theta} \left(||Y - \theta||_p^p + \lambda_n^2 I(\theta) \right)$$

where the norm $||.||_p$, the index p and the penalty $I : \Theta \to \mathbb{R}^+$ have to be chosen properly. In many cases this minimization can be solved directly and an explicit version of the estimator is used in our simulations. In the other cases, especially when the loss-function is not quadratic, the minimization problem can not be solved directly and standard algorithms fail to achieve the minimization. That is why we will have to use more complex algorithms such as the interior point algorithm.

Penalized least-squares estimators

In this case $||\theta||_n^2 = \frac{1}{n} \sum_{i=1}^n \theta(z_i)^2$. For our simulation study, we rewrite the minimization problem as follows. Given an orthonormal basis of the space Θ, ψ_j, we decompose each function onto this basis:

$$\begin{aligned} \theta &= \sum_{j=1}^n \beta_j \psi_j \\ &= \Phi \beta. \end{aligned}$$

where Φ is an operator over $\mathbb{R}^n$. So the estimator $\hat{\theta}_n = \sum_{j=1}^n \hat{\beta}_j \psi_j$ is defined by its coefficients and is a solution of

$$\begin{aligned} \hat{\theta}_n &= \arg \inf_{\beta = (\beta_j)_{j=1,\dots,n}} \left(||Y - \Phi\beta||_2^2 + \lambda_n^2 I(\beta) \right) \\ &= \arg \inf_{\beta = (\beta_j)_{j=1,\dots,n}} \left(\sum_{j=1}^n |Y_i - (\Phi\beta)_i|^2 + \lambda_n^2 I(\beta) \right) \end{aligned}$$

The numerical problem of minimizing a quadratic norm and a penalization term can be solved in all cases using a relaxation algorithm described as follows.

A relaxation algorithm is a recursive algorithm. We describe here an iteration:

- choose a starting point $u^0 = (u_1^0, \dots, u_n^0)$.

- select u_i^{n+1} such that $\forall x \in \mathbb{R}$ the following inequality holds:

  $$J(u_1^{n+1},\dots,u_{i-1}^{n+1},u_i^{n+1},u_{i+1}^{n},\dots,u_n^{n}) \le J(u_1^{n+1},\dots,u_{i-1}^{n+1},x,u_{i+1}^{n},\dots,u_n^{n}).$$

  Repeat this selection for $i = 1,\dots,n$.

The following theorem describes the efficiency of this program.

Theorem 2.4.13. *Define a function $J : \mathbb{R}^n \to \mathbb{R}$*

$$J(v) = J_0(v) + \sum_{i=1}^{n} \alpha_i |v_i|^p$$

where $\alpha_i, i = 1,\dots,n$ are positive constants, p is positive real, and J_0 a strictly convex continuously differentiable function such that $J(v) \to \infty$, when $||v|| \to \infty$. The relaxation algorithm, described above, converges to the solution of the minimization problem of $J(v)$ over $v \in \mathbb{R}^n$.

Proof. The proof of this result can be found in [GLT76] for instance. □

This algorithm is a direct algorithm which means that it does not use the expression of the derivate of J_0, so it is easier to calculate the expression of the minimizer when p ranges from 1 to 3. We recall here the direct expression of the M-estimators in the three cases where we can find an explicit solution:

- $p = 1$. We find the minimum by deriving the constraint function and analyzing whether the minimum is in the open set or in the boundaries of the set: it is given by

  $$\hat{\beta}_j = \mathrm{sgn}(\beta_\mathrm{j})(|\beta_\mathrm{j}| - \lambda_\mathrm{n}^2|\alpha_\mathrm{j}|/2)_+.$$

  We can recognize the soft-thresholded estimator.

- $p = 2$. The minimization problem is:

  $$\hat{\beta}_j = \arg\min_x \left((\beta_j - x)^2 + \lambda_n^2 x^2\right)$$

  It is the easiest case. Using differentiation properties of both loss function and penalty, we know that the minimum of the penalized contrast is reached for:

  $$\hat{\theta} = \sum_{j=1}^{n} \hat{\beta}_j \psi$$
  $$\hat{\beta}_j = \frac{1}{1+\lambda_n^2|\alpha_j|}\beta_j$$

  This estimator is a smoothed linear estimator.

- $p = 3$. The minimization problem is

$$\hat{\beta}_j = \arg\min_x \left((\beta_j - x)^2 + \lambda_n^2 |x|^3\right)$$

The solution is a zero of the gradient so if we differentiate the functional we get:

$$\begin{aligned} -2(\beta_j - x) + 3\mathrm{sgn}(x)\lambda_n^2|\alpha_j|x^2 =&0 \\ \mathrm{sgn}(x)(3\lambda_n^2|\alpha_j|x^2 + 2|x|) =&2\mathrm{sgn}(\beta_j)|\beta_j| \\ \mathrm{sgn}(x) =&\mathrm{sgn}(\beta_j) \\ |x| =&\frac{-1+\sqrt{1+6\lambda_n^2|\alpha_j||\beta_j|}}{3\lambda_n^2|\alpha_j|} \\ x =&\frac{2|\beta_j|\mathrm{sgn}(x)}{1+\sqrt{1+6\lambda_n^2|\alpha_j||\beta_j|}} \\ x =&\frac{2\beta_j}{1+\sqrt{1+6\lambda_n^2|\alpha_j||\beta_j|}}. \end{aligned}$$

So we have

$$\hat{\theta} = \sum_{j=1}^{n} \frac{2\beta_j}{1+\sqrt{1+6\lambda_n^2|\alpha_j||\beta_j|}}\psi_j.$$

However in the other cases such basic algorithm provides a solution even if it may be more efficient to use a gradient or conjugate gradient algorithm described by P. Ciarlet in [Cia82] to minimize the penalized loss function provided the penalty can be differentiated. We point out that choosing properly the coefficients α_j enables us to study norm-type penalties such as Sobolev and Besov pseudo-norm.
The following simulations are done using a wavelet basis ψ_{jk} with compactly supported functions with vanishing moments. We use Daubechies wavelet with 8 vanishing moments. From the observed data Y_i which is taken dyadic, we get wavelet coefficients from the Fast Wavelet Algorithm which can be found in [Mal98]. We present in the last section the results of some simulations: quadratic losses of different signals for 4 choices of penalty including the special choice of a Besov pseudo norm. The errors are taken either Gaussian or Laplacian.

However such methods require the differentiability of the loss function which is not the case if we consider the penalized least absolute deviation estimator whose behavior has been studied in [LvdG00]. Standard algorithms do not work so we have to find another way to minimize the functional. Such methods are close to the solution of the simplex algorithm proposed by Karmakar in [Kar92] or Barrodale and Roberts in [BR73].
The estimator is now defined by:

$$\hat{\theta}_n = \arg\inf_{\theta\in\Theta}\left(||Y-\theta||_{L^1} + \lambda_n^2\sum_{i=1}^{n}|\beta_{jk}|\right).$$

This problem is close to the one studied by Bruce and Sardy in [SS99]. I used two ways of solving the optimization problem: either modify slightly the loss function on order to gain some differentiability properties or use optimization theory to see the problem as a constrained minimization whose dual problem may be solved.

Penalized Least Absolute deviation estimator

1. Huber loss function:
   Previous authors have solved the minimization problem with a loss that approximates locally the behavior of the l^1 loss, the Huber loss function described by Huber in [DH81]. For a given cut point $\tau > 0$, this function ρ can be written:

   $$\rho(x) = \begin{cases} \frac{x^2}{2} & \text{if } |x| \leq \tau \\ \tau|x| - \frac{\tau^2}{2} & \text{elsewhere.} \end{cases}$$

   It is an hybrid between the l_1 loss and the l_2 loss since for large residuals, it behaves like the first one whereas for small residuals, it behaves like a quadratic loss.
   First of all, we point out that the minimization problem with a quadratic loss and a l_1 penalty

   $$\min_\beta \left(||Y - \Phi\beta||_2^2 + 2\lambda_n^2 ||\beta||_1 \right)$$

   has for solution the soft-thresholded estimator well studied by Donoho and Johnstone in [DJ95]

   $$\tilde{\theta} = \sum_{(j,k)} \delta_S(\hat{\beta}_{jk}, \lambda_n^2)\psi_{jk}$$

   where $\hat{\beta}_{jk} = \frac{1}{n}\sum_{i=1}^n Y_i\psi_{jk}(z_i)$ and $\delta_S(x, l) = \text{sgn}(x)(|x| - l)_+$ is the soft-thresholding operator. To compute this estimator, we can use the procedure Wave-Shrink implemented for MatLab.
   Sardy, Bruce and Tseng reformulate the optimization problem as shown:

   $$\min_{\beta,w} \left(||Y - (\Phi\beta + w)||_2^2 + \tau||w||_1 + \lambda_n^2||\beta||_1 \right).$$

   They propose to use a Block Coordinate Relaxation algorithm:

   (a) Choose a starting point for the algorithm $x = (\beta, w)$;

   (b) Partition $B = [\Phi, I]$ into two matrices B_1 an orthonormal matrix and B_2 the remaining matrix. Define x_1 and x_2 the corresponding vectors;

   (c) Set $r = Y - B_2x_2$ the residual vector;

   (d) Improve x_1 by solving the problem:

   $$x_1 = \arg\min_b \left(||r - B_1b||_2^2 + \lambda_n^2||b||_1 \right)$$

   using the WaveShrink procedure;

   (e) Test the approximation and if convergence criterion is not met, go to the first step.

The difficult point in this algorithm is finding the orthonormal matrix. Then if we let τ decrease to zero we get an approximation of our minimization problem.

2. Interior Point Algorithm:
Using convex duality theory, we want to transform the optimization problem into its dual problem in order to apply an Interior Point Algorithm developed by Chen, Donoho and Saunders in [CDS99]. We recall that we want to minimize

$$\min_{\beta} \left(||Y - \Phi\beta||_1 + \lambda_n^2 ||\beta||_1 \right).$$

We firstly rewrite the problem as:

$$\min_{\beta,w} \left(||w||_1 + \lambda_n^2 ||\beta||_1 \right)$$

$$\text{with } w = Y - \Phi\beta.$$

The optimization problem can be considered as a minimization problem with constraint. That is the reason why we make use of Lagrange multipliers to point out the dual problem, obtained by exchanging the order of the minimum and the maximum. Since the loss function is convex, and since the constraint is linear, the dual gap between the primal problem and the dual problem is zero as can be shown in a book from Rockafellar [Roc97].

$$\begin{aligned}
&\min_{\beta, w = Y - \Phi\beta} \left(||w||_1 + \lambda_n^2 ||\beta||_1 \right) \\
&= \min_{\beta,w} \left(\max_x (\sum_{i=1}^n |w_i| + \lambda_n^2 \sum_{i=1}^n |\beta_i| + (Y - \Phi\beta - w)^{'} x) \right) \\
&= \max_x \left(\min_{\beta,w} (\sum_{i=1}^n |w_i| + \lambda_n^2 \sum_{i=1}^n |\beta_i| + (Y - \Phi\beta - w)^{'} x) \right) \\
&= \max_x \left(\min_{\beta,w} (\sum_{i=1}^n (|w_i| - w_i x_i) + \sum_{i=1}^n (\lambda_n^2 |\beta_i| - \beta_i (\Phi x)^{'}) + Y^{'} x) \right) \\
&= \max_x \left(\sum_{i=1}^n \min_{w_i} (|w_i| - w_i x_i) + \sum_{i=1}^n \min_{\alpha_i} (\lambda_n^2 |\alpha_i| - \alpha_i (\Phi_i^{'} x)) + Y^{'} x \right)
\end{aligned}$$

where we have set Φ_i the i^{th} column of the matrix Φ. This problem is equivalent to

$$\min_x -Y^{'} x \text{ with : } \begin{cases} |\Phi_i^{'} x| & \leq \lambda_n^2 \\ |x_i| & \leq 1 \end{cases}$$

We have transformed the initial dual problem into a linear programming problem. More generally, if we had tried to minimize the Huber loss function, the problem would have been reduced to

$$\min_x \sum_{i=1}^n \frac{1}{2} x_i^2 - Y^{'} x \text{ with : } \begin{cases} |\Phi_i^{'} x| & \leq \lambda_n^2 \\ |x_i| & \leq \tau \end{cases}$$

Such a quadratic minimization issue is well handled using a Primal-Dual Log-Barrier Interior Point algorithm. The idea is the following: we add a log-penalty to the minimization problem:

$$\begin{aligned}\min_x -Y'x - \mu\sum_{i=1}^n \log(\lambda_n^2 - \Phi_i' x) - \mu\sum_{i=1}^n \log(\lambda_n^2 + \Phi_i' x)\\ - \mu\sum_{i=1}^n \log(\tau - x_i) - \mu\sum_{i=1}^n \log(\tau + x_i)\end{aligned}$$

So if we differentiate we find the following condition:

$$\begin{aligned}x - Y + \mu\sum_{i=1}^n \frac{1}{\lambda_n^2 - \Phi_i' x}\Phi_i - \mu\sum_{i=1}^n \frac{1}{\lambda_n^2 - \Phi_i' x}\Phi_i\\ + \mu\sum_{i=1}^n \frac{1}{\tau - x_i}e_i - \mu\sum_{i=1}^n \frac{1}{\tau + x_i}e_i = 0\end{aligned}$$

where e_i is the i^{th} canonical vector of the basis. If we set

$$\begin{array}{rclrcl} t_i^+ &=& \dfrac{\mu}{\lambda_n^2 - \Phi_i' x} & t_i^- &=& \dfrac{\mu}{\lambda_n^2 + \Phi_i' x}\\ r_i^+ &=& \dfrac{\mu}{\tau - x_i} & r_i^- &=& \dfrac{\mu}{\tau + x_i}\\ v_i^+ &=& \lambda_n^2 - \Phi_i' x & v_i^- &=& \lambda_n^2 + \Phi_i' x\\ u_i^+ &=& \tau - x_i & u_i^- &=& \tau + x_i \end{array}$$

and let

$$\begin{aligned} z &= (u^+, u^-, v^+, v^-)\\ s &= (r^+, r^-, t^+, t^-)\\ A &= [I, -I, \Phi, -\Phi]\\ c &= (c.\mathbf{1}, \boldsymbol{\lambda_n^2}.\mathbf{1}) \end{aligned}$$

the first order condition becomes:

$$\begin{cases} -A'x - z + c & = 0\\ Y - As - x & = 0\\ \mu\mathbf{1} - \text{diag}(\mathbf{s})\text{diag}(\mathbf{z})\mathbf{1} & = 0 \end{cases}$$

So the minimization problem can be solved by solving this non-linear system with a conjugate gradient for example and then decrease the log-barrier term μ to enforce convergence. This algorithm is a Primal-Dual algorithm, so at each Newton step, the three variables x, s, z are known and updated. s is the primal variable, giving a solution of the original minimization problem in α, x is the dual variable giving the solution of the maximization in the Lagrange multiplier. A stopping condition of the algorithm can be found when the three directions

$$\begin{aligned} d_1 &= -A^{'}x - z + c \\ d_2 &= Y - As - x \\ d_3 &= \mu\mathbf{1} - \text{diag}(\mathbf{s})\text{diag}(\mathbf{z})\mathbf{1} \end{aligned}$$

are small enough.
If we consider the l^1 minimization issue, we recall that the dual problem is a linear programming problem. So the minimization can be easily done with a small number of iterations.

Simulation Results

The following simulations have been done using MatLab and WaveLab. We have used four different signals: a function with regular oscillations $x \to \sin(8x)$, a regular function with a discontinuity Heavisine functional, a very irregular function: the Bump function and a high oscillating function: the Doppler function. These functions are observed with two different noise: a Gaussian white noise with variance 3 and a Laplacian noise with the same variance. In a first study, we consider the Heavisine function and the Doppler function with $n = 100$ observations, and decompose these two functions onto a wavelet basis using a Daubechies wavelet with 8 vanishing moments. As error distribution, we considered the standard Gaussian distribution, normalized with a noise-ratio equal to $1/3$, and also the Laplacian distribution, i.e ., double exponential distribution, with mean zero and variance 3. We have looked at 9 cases, corresponding to different values of the smoothing-parameter λ_n^2 including the theoretical optimal value for the Gaussian case 0.303 that corresponds to the 8^{th} line. The four tables summarize the performance of the LS and LAD estimators. In order to make comparison of LS and LAD relevant, we have put on a same line the results with $\sigma_0\lambda_n^2$ for the LS and λ_n^2 for the LAD. We also added a line where comparisons are made for the optimal cases, i.e., smallest $||\hat{\theta}_n - \theta_0||_n$ (corresponding to different smoothing parameters). In these simulations, we can see that LAD works better in the Laplacian case, and LS works better in th Gaussian case (as is to be expected). We note furthermore that the value $\lambda_n^2 = \sqrt{2\log n/n}$ is not optimal in the LAD case: it is too large. In the LS case, the corresponding value $\lambda_n^2 = \sigma_0\sqrt{2\log n/n}$ is also too large when the errors are Laplacian, but it is optimal when the errors are Gaussian.

Heavisine function with Gaussian errors

λ^2	MSE for LS	MSE for LAD
0.0303 (1)	0.7535	0.605
0.0607 (2)	0.5229	0.3994
0.1011 (3)	0.4782	0.3737
0.1517 (4)	0.4934	0.4507
0.2124 (5)	0.4749	0.4612
0.2427 (6)	0.3451	0.4828
0.2731 (7)	0.2821	0.5003
0.3034 (8)	0.2238	0.5601
0.6070 (9)	0.5852	0.6242
optimum	0.2238 at (8)	0.3737 at (3)

Heavisine function with Laplacian noise.

λ^2	MSE for LS	MSE for LAD
0.0303 (1)	1.7051	1.5157
0.0607 (2)	1.010	0.954
0.1011 (3)	.8201	0.6238
0.1517 (4)	0.7853	0.5896
0.2124 (5)	0.6021	0.4324
0.2427 (6)	0.5925	0.4654
0.2731 (7)	0.5896	0.5870
0.3034 (8)	0.6012	0.6925
0.607 (9)	0.6238	0.7021
optimum	0.5896 at (7)	0.4324 at (5)

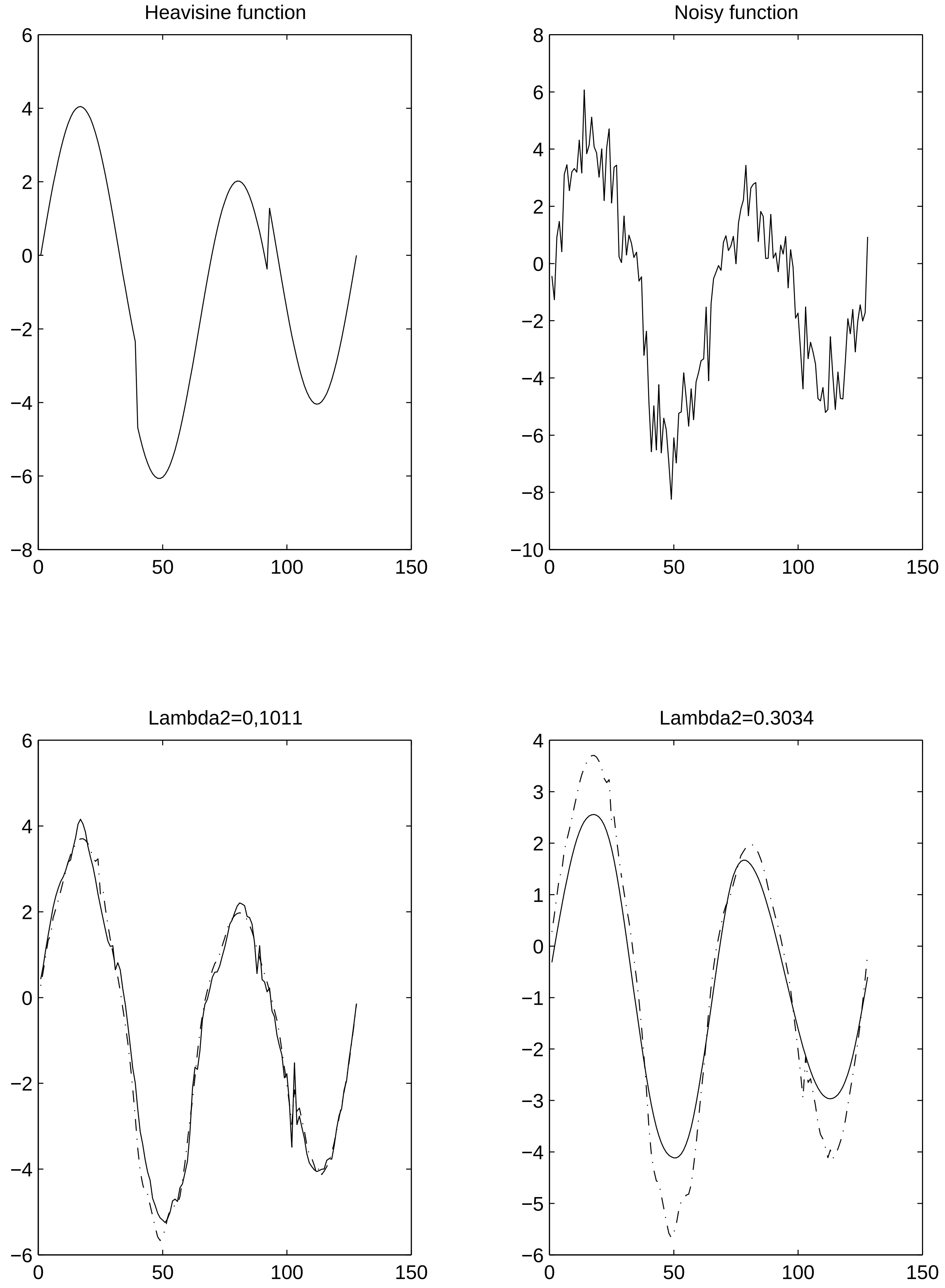
Doppler signal with Gaussian errors.

λ^2	MSE for LS	MSE for LAD
0.0303 (1)	0.5862	0.8103
0.0607 (2)	0.5210	0.7801
0.1011 (3)	0.3521	0.6610
0.1517 (4)	0.2625	0.5218
0.2124 (5)	0.2212	0.3451
0.2427 (6)	0.1521	0.2821
0.2731 (7)	0.1330	0.3299
0.3034 (8)	0.090	0.4445
0.6070 (9)	0.3928	0.5510
optimum	0.090 at (8)	0. 2821 at (6)

Doppler signal with Laplacian errors.

λ^2	MSE for LS	MSE for LAD
0.0303 (1)	0.736	0.901
0.0607 (2)	0.6260	0.7700
0.1011 (3)	0.5218	0.6101
0.1517 (4)	0.5680	0.5018
0.2124 (5)	0.6321	0.3451
0.2427 (6)	0.7081	0.2821
0.2731 (7)	0.8588	0.8229
0.3034 (8)	0.9097	0.8429
0.607 (9)	0.9254	0.9545
optimum	0.5218 at (3)	0.2521 at (6)

In a second study we use the four test functions to study the impact of the smoothness parameter. The following figures show the behavior of the risk function taken as a function of the parameter λ_n^2. In the first figure are displayed the quadratic risks for the four test functions where the estimator is the penalized least squares estimator with an l^1 penalty. In the second, the estimator studied is the penalized least square with a Besov type penalty.



Heavisine function

Noisy function

Lambda2=0,1011

Lambda2=0.3034

Figure 2.1:

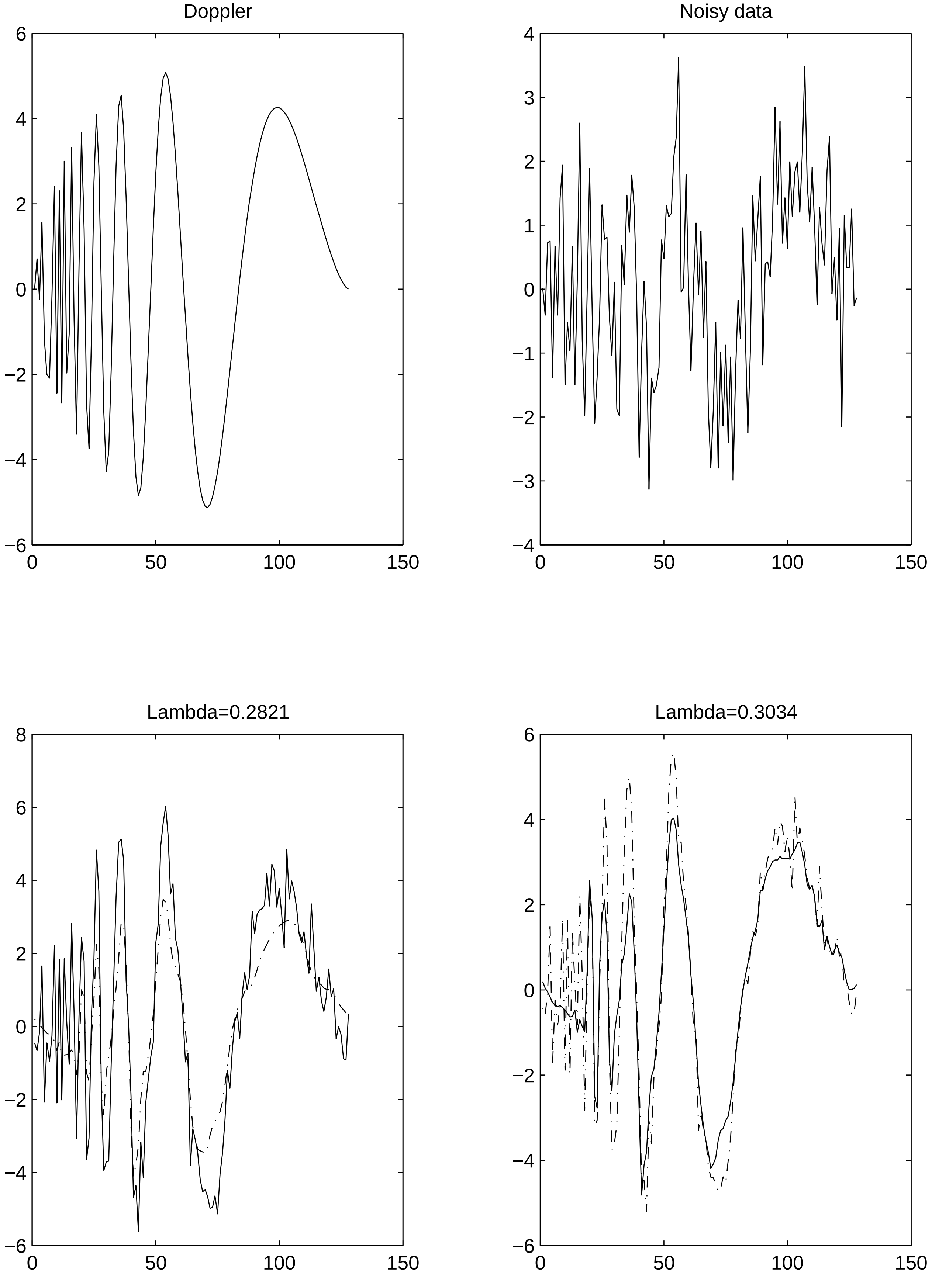
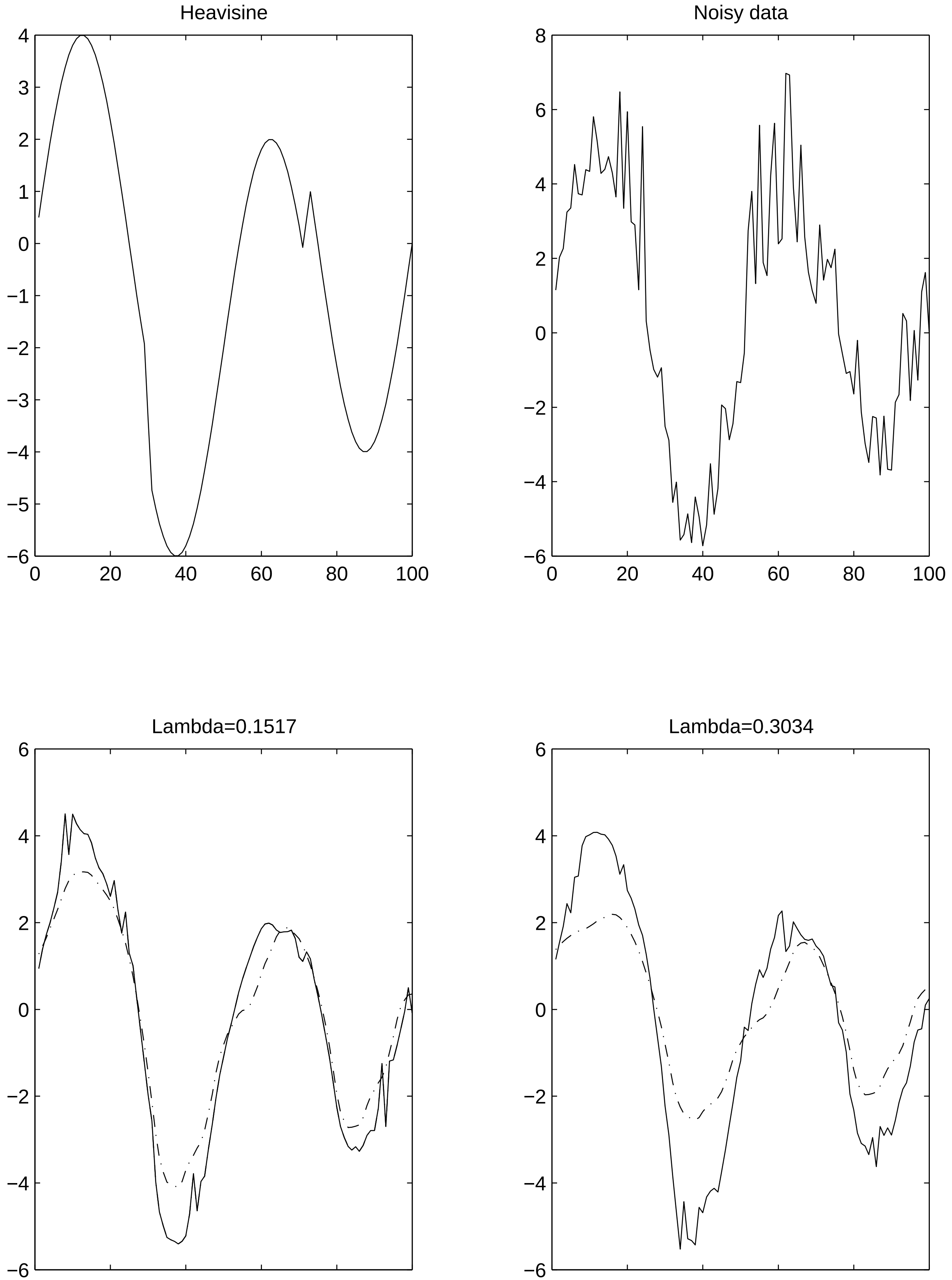

Doppler
6
4
2
0
−2
−4
−6
0
50
100
150
Noisy data
4
3
2
1
0
−1
−2
−3
−4
0
50
100
150
Lambda=0.2821
8
6
4
2
0
−2
−4
−6
0
50
100
150
Lambda=0.3034
6
4
2
0
−2
−4
−6
0
50
100
150

Figure 2.2:



Heavisine

Noisy data

Lambda=0.1517

Lambda=0.3034

Figure 2.3:

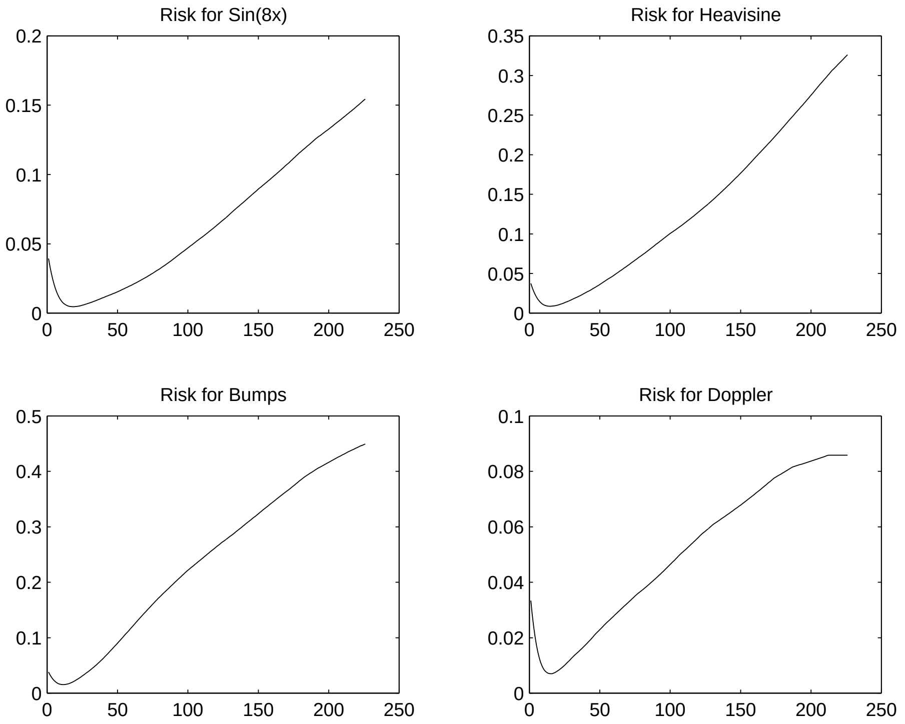


Figure 2.4: L1 penalized least-squares estimator with Gaussian errors

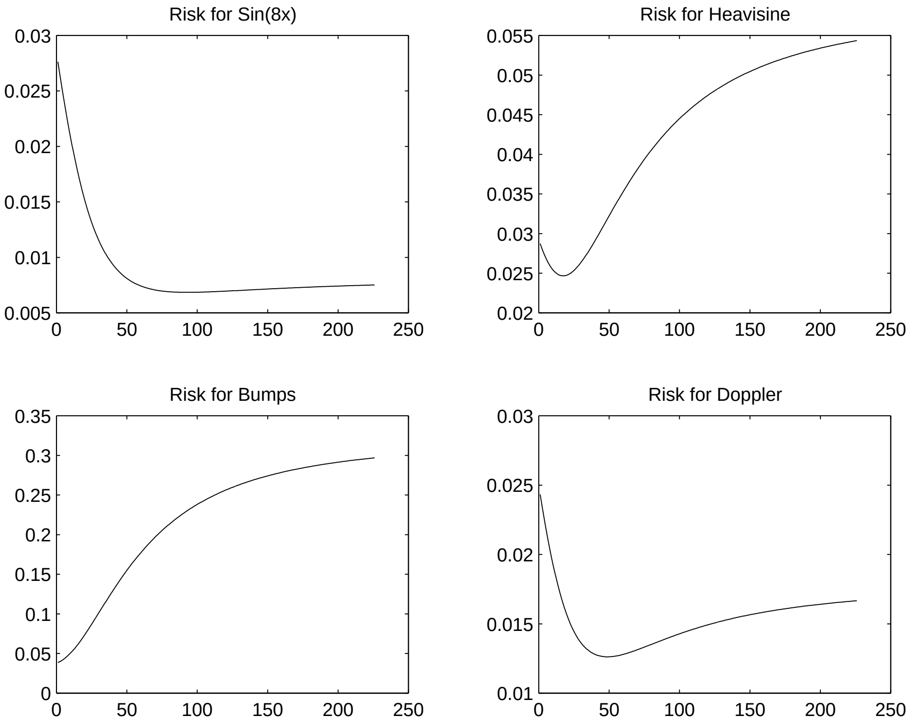


Figure 2.5: Besov penalized least-squares estimator with Gaussian errors

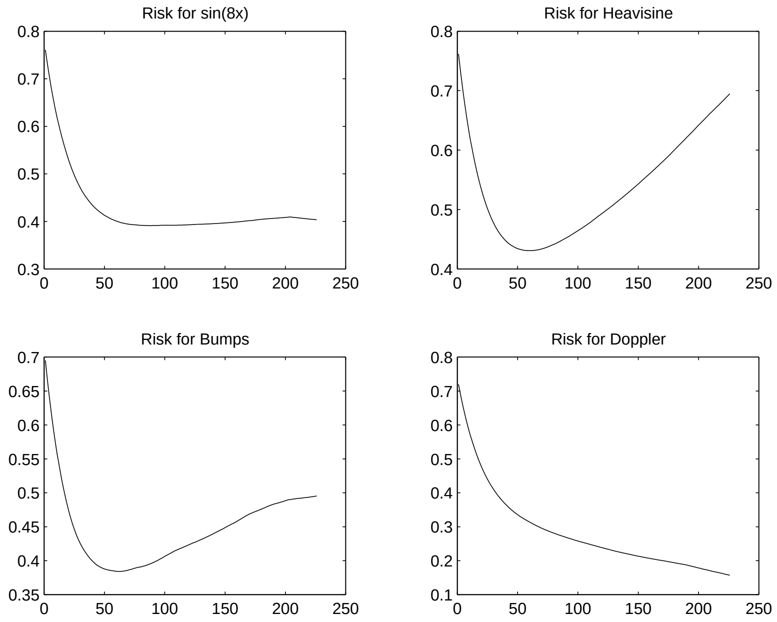


Figure 2.6: L1 penalized least-squares estimator with Laplacian errors

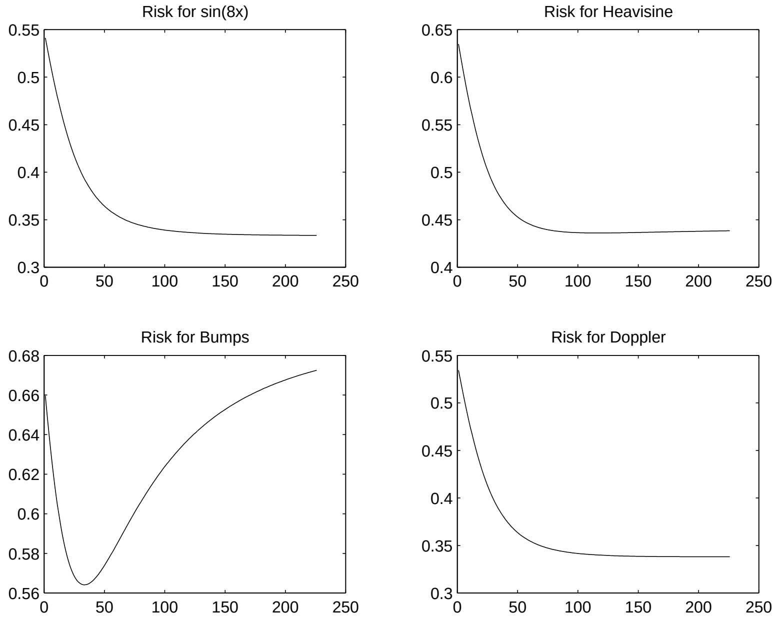


Figure 2.7: Besov penalized least-squares estimator with Laplacian errors

2.4.4 Example of entropy calculation

On the non-optimality of Ball and Pajor inequality

Abstract. We show that Ball and Pajor inequality is yet a powerful tool to calculate entropy of class of functions but does not provide sharper bounds as can be shown when using it to derive rates of convergence for M-estimator over a specific set.

Our model is the standard regression model

$$\begin{cases} Y_i = \theta_0(z_i) + W_i, \; i = 1, \dots, n, \\ \theta_0 \in \Theta, \end{cases}$$

where Θ is the set of functions

$$\Theta = \{\theta \in C^2([0,1],[0,1]), \; \theta^{''}(x) \leq 1, \; \forall x \in [0,1]\}.$$

We consider the M-estimator

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \sum_{\mathrm{i}=1}^{\mathrm{n}} \gamma_\theta(\mathrm{Y_i}).$$

where $\gamma_\theta : \mathbb{X} \to \mathbb{R}$ is a loss-function which has to be convex in θ. The rate of convergence of such estimators depends on the entropy of the set Θ, which can be found in [vdG00]. One can define entropy for general metric spaces but we shall restrict ourselves to classes of functions Θ. Let Q be a measure on a measurable space.

Definition 2.4.14. *Consider for each $\delta > 0$ a collection of functions $\theta_1, \dots, \theta_N$ such that, for each $\theta \in \Theta$ there exists $j \in \{1, \dots, N\}$ such that $||\theta - \theta_j||_Q \leq \delta$. Let $N(\delta, \Theta, ||.||_Q)$ the smallest value of N for which such a covering, with balls centered in θ_j and of radius δ, exists ($N = \infty$ if this value is not finite). $N(\delta, \Theta, ||.||_Q)$ is called the δ-covering number. $H(\delta, \Theta, ||.||_Q) = \log N(\delta, \Theta, ||.||_Q)$ is called the δ-entropy of the class of functions Θ*

In most cases the derivation of a good bound for the entropy of a class of functions is rather complicated. However Ball and Pajor in [BP90] prove the following theorem that enables us to calculate the entropy of a convex hull of a class K of functions from the δ-covering number of the set K.

Theorem 2.4.15. *Let Q be a probability and K be a set in $L^2(Q)$, so that there exists positive constants c and d so that*

$$N(\delta, K, Q) \leq c\delta^{-\frac{1}{d}}, \; \forall \delta > 0.$$

Then for $s = \frac{1}{2} + \frac{1}{d}$ there exists a finite constant $A > 0$ so that the following bound holds:

$$H(\delta, \mathrm{conv(K)}, \mathrm{Q}) \leq \mathrm{A}\delta^{-\frac{1}{\mathrm{s}}}.$$

This theorem provides an entropy bound but the problem that arises, is the optimality of the inequality provided. For the set

$$\Theta = \{\theta : \mathbb{R} \to [0,1],\ \theta \text{ is increasing } \}$$

the bound is optimal, i.e we can easily see that

$$\text{Extr}(\Theta) = \{1_{[\text{y},\infty)}\ , \text{y} \in \mathbb{R}\}.$$

The number of balls necessary to cover this set is

$$N(\delta, \text{Extr}(\Theta)) \leq \text{c}\delta^{-2},\ \forall \delta > 0.$$

So the entropy of the set Θ is:

$$H(\delta, \Theta, ||.||_\infty) \leq A\delta^{-1},$$

which corresponds to the bound one could find by direct calculations as it is done in [vdG00]

In this part, we want to calculate the entropy of

$$\Theta = \{\theta \in C^2([0,1],[0,1]),\ \theta^{''}(x) \leq 1,\ \forall x \in [0,1]\}.$$

Direct calculations are uneasy so, we consider this set as the convex hull of its extreme points, and then we make use of the theorem proved by Ball and Pajor in [BP90] to calculate its entropy. So the main difficulty is finding the minimal space of the extreme points.

Determination of the Extreme Points

With the next three lemmas, we will determine the nature of the extreme points of the set Θ. By the mere definition, an extreme point belongs to the border of the set defined as:

$$\partial\Theta = \left(\theta \in C^2([0,1],[0,1]), \exists x \in [0,1], \begin{cases} \theta(x) = 0 \\ \text{or } \theta(x) = 1 \\ \text{or } \theta^{''}(x) = 1 \end{cases} \right)$$

The three next lemmas characterize properties of extreme points of Θ. The first lemma describes the behavior of points located at the border of the set and drastically reduces the number of possible extreme points. Θ.

Lemma 2.4.16. *Let θ in $\partial\Theta$, if there exists $x_0 \in [0,1]$ such that $0 < \theta(x_0) < 1$ and $\theta^{''}(x_0) < 1$, then θ is not an extreme point.*

The next lemma describes a specific property of extreme points: except at the border of the interval, the extreme values $0, 1$ can not be reached simulteanously by a function in Extr(Θ).

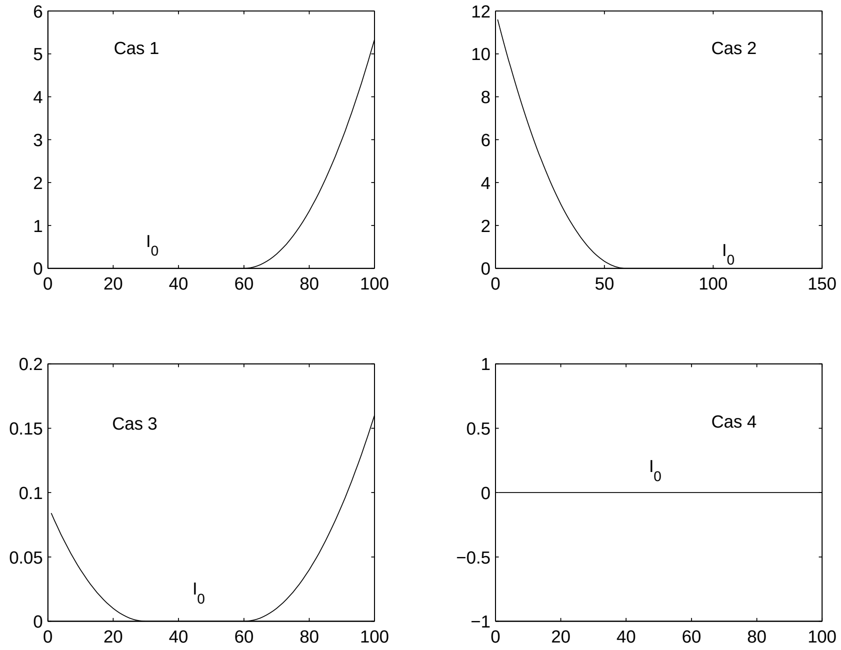


Figure 2.8:

Lemma 2.4.17. *For an extreme point θ, the following proposition is false:*

$$\exists x_1 \neq x_2,\ x_1 \notin \{0,1\},\ x_2 \notin \{0,1\}\ ,\ \theta(x_1) = 0 \,\text{and}\, \theta(x_2) = 1.$$

The last lemma describes the topology of the set of points where a function in $\mathrm{Extr}(\Theta)$ reaches the extremal value 0 or 1.

Lemma 2.4.18. *For an extreme point θ, if we set $I_0 = \{x, \theta(x) = 0\}$ and $I_1 = \{x, \theta(x) = 1\}$. then the two sets I_0 and I_1 are two intervals.*

Now we are able to describe the set of the extreme points.

1. If there exists an interior point x_0 such that $\theta(x_0) = 0$, we have two possibilities whether the interior of I_0 is empty or not.

   (a) if the interior of I_0 is not empty, we have 4 possibilities which is shown in figure 2.8:

   The three first cases are impossible due to the continuity of the function θ. Indeed on the one hand, we would have $\theta''(\tilde{x}_1) = 0$ and on the other hand $\theta''(\tilde{x}_1) = 1$ by continuity. So the only possibility is given by $\theta = 0$.

   (b) if the interior of I_0 is empty, then the only function possible is

   $$\theta = \frac{1}{2}(x - x_0)^2 \in \Theta,$$

   which is given in figure 2.9.

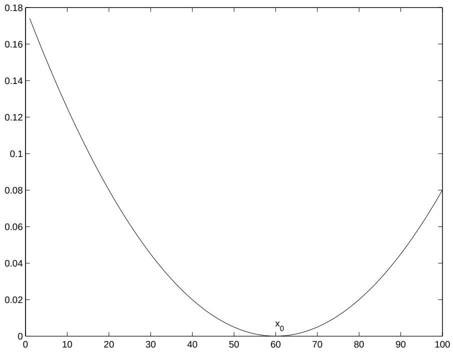


Figure 2.9:

2. If for all $x \in [0,1]$, $\theta(x) \neq 0$ and there exists x_1 such that $\theta(x_1) = 1$.

   (a) if the interior of I_1 is not empty, with the same arguments as above, we can conclude that $\theta = 1$.

   (b) if the interior of I_1 is empty, this case is impossible.

3. If for all x, $\theta(x) \notin \{0,1\}$, then an extreme point of Θ is a function of the following form

$$\theta : x \to \frac{1}{2}x^2 + ax + b,$$

where the constants a and b are properly chosen.

   (a) if $\theta(x) = \frac{1}{2}x^2 + ax + b$ takes its extrema in $]0,1[$. Then $\theta([0,1]) \subset [a,b]$, where $0 < a \leq b < 1$. Then define

$$\begin{cases} h_1(x) = \theta(x) + \epsilon \\ h_2(x) = \theta(x) - \epsilon. \end{cases}$$

   But $\forall 0 < \epsilon \leq \min(a, 1-b)$ we can see that the functions h_1 and h_2 belong to the set Θ. But we have again $\theta = \frac{h_1+h_2}{2}$, so such a function is not an extreme point.

   (b) if $\theta = \frac{1}{2}x^2 + ax + b$ has one extremum for a value corresponding to one of the extreme points of the interval 0 or 1, then, if we suppose that there exist two different functions in the set Θ, h_1 and h_2 such that, there exists a real $\lambda \in (0,1)$, such that $\theta = \lambda h_1 + (1-\lambda)h_2$. Let $\rho \in \{0,1\}$, if there exists $x_0 \in]0,1[$ such that $\theta(x_0) = \rho$, then since the functions h_1, h_2 take their value in the interval $[0,1]$, we have $h_1(x_0) = h_2(x_0) = \rho$. So we obtain $h_1^{'}(x_0) = h_2^{'}(x_0) = 0$. Moreover

$$1 = \theta^{''}(x) = \lambda h_1^{''}(x) + (1-\lambda)h_2^{''}(x),$$

   so $h_1^{''} = h_2^{''} = 1$. So the functions h_1 and h_2 are solutions of the differential equation:

$$\begin{cases} f^{''} = \theta^{''} \\ f(x_0) = \theta(x_0) \\ f^{'}(x_0) = \theta^{'}(x_0). \end{cases}$$

So the solution of this problem is $h_1 = h_2 = \theta$, and the function is an extreme point.

So we can conclude that the set of extreme points is composed of the following functions:

1. the two constant functions 0 and 1.

2. second order polynomals $\theta(x) = \frac{1}{2}x^2 + ax + b$ such that one of the three following condition is fulfilled:

   (a)
   $$\theta(0) \in \{0, 1\}$$
   which implies that $b \in \{0, 1\}$.

   (b)
   $$\theta(1) \in \{0, 1\}$$
   which implies that $|a + b| = \frac{1}{2}$.

   (c)
   $$\exists x_0 \in]0, 1[, \left\{ \begin{array}{l} \theta(x_0) = 0 \\ \theta'(x_0) = 0 \end{array} \right.$$
   which implies the following conditions:
   $$2b = a^2, \quad b \in]0, 1[.$$

Entropy Calculation

In the first section, we have determined the extreme points, $\mathrm{Extr}(\Theta)$, of the set Θ and so we can write that:
$$\Theta = \overline{\mathrm{Conv}(\mathrm{Extr}(\Theta))}.$$

Let us calculate the entropy of $\mathrm{Extr}(\Theta)$: consider a sequence of reals, (a_i), chosen such that balls centered in a_i form a $\frac{\delta}{2}$-covering sequence of the interval $[0, 1]$. So consider the sequence $a_i = i\frac{\delta}{2}$, $i = 0, \dots, \frac{2}{\delta}$. For all $a \in [0, 1]$, there exists $j = j(a)$ such that $|a_j - a| \leq \frac{\delta}{2}$, but for b_i defined by $b_i = \frac{1}{2}a_i^2$ or $|a_i + b_i| = \frac{1}{2}$, we have if we set $\theta_i = \frac{1}{2}x^2 + a_i x + b_i$:
$$\forall \theta \in \mathrm{Extr}(\Theta),\ \exists \theta_j,\ ||\theta - \theta_j||_\infty \leq \delta.$$

$$\begin{aligned} ||\theta - \theta_j||_\infty &\leq |a - a_j| + |b - b_j| \\ &\leq 2|a - a_j| \\ &\leq \delta. \end{aligned}$$

So we need $N = \frac{4}{\delta}$ δ-covering balls for $\mathrm{Extr}(\Theta)$. But for a set Λ, we know that taking the closure of a set has no impact on the entropy number, so
$$H(\Lambda, \delta) = H(\overline{\Lambda}, \delta)$$

and the theorem 2.4.15 from [BP90] enables us to compute the entropy of a set using the entropy of its extreme point.

So with this theorem we can conclude that the following proposition holds:

Proprosition 2.4.19. *For* $\Theta = \{\theta \in C^2([0,1],[0,1]),\ \theta''(x) \leq 1,\ \forall x \in [0,1]\}$

$$H_2(\Theta, \delta) \leq A\delta^{-\frac{2}{3}}.$$

But we know using results from Kolmogorov and Tikhomirov [KT59] that an optimal upperbound of the set Θ is given by:

$$H_\infty(\Theta, \delta) \leq A\delta^{-\frac{1}{2}}$$

for a positive finite constant A. As a result, the bound given by Ball and Pajor's inequality is not sharp enough.

Rates of convergence

In this section, we recall how rates of convergence for M-estimators such as the least squares estimator or the penalized least squares estimator can be established by an entropy argument. Such results are stated in [vdGW96], [vdG00] or [LvdG00].

Theorem 2.4.20. *Least-squares estimator.*
We consider the M-estimator

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} ||\mathrm{Y} - \theta||_{\mathrm{n}}^2.$$

If for $\eta > 0$,$A > 0$,$s \geq \frac{1}{2}$

$$H(\delta, \{\theta \in \Theta, ||\theta - \theta_0|| \leq \eta\}) \leq A\left(\frac{M_n}{\delta}\right)^{\frac{1}{s}}, \delta > 0$$

then

$$||\hat{\theta}_n - \theta_0|| = \begin{cases} O_{\mathbf{P}}(n^{-\frac{s}{2s+1}})M_n^{\frac{1}{2s+1}} & , s > \frac{1}{2} \\ O_{\mathbf{P}}(n^{-\frac{1}{4}})\sqrt{M_n \log(n)} & , s = \frac{1}{2}. \end{cases}$$

Theorem 2.4.21. *Consider the Penalized Least-Squares Estimator defined as the solution of the following optimization problem:*

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta}\left(\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{Y_i} - \theta(\mathrm{z_i}))^2 + \lambda_{\mathrm{n}}^2 \mathrm{I^p}(\theta)\right)$$

If for $\eta > 0$,$A > 0$,$s \geq \frac{1}{2}$, $M_n \geq I(\theta_0)$

$$H(\delta, \{\theta \in \Theta, ||\theta - \theta_0|| \leq \eta, I(\theta) \leq M_n\}) \leq A\left(\frac{M_n}{\delta}\right)^{\frac{1}{s}}, \delta > 0$$

then for

$$\lambda_n^{-1} = \begin{cases} O_{\mathbf{P}}(n^{\frac{s}{2s+1}})M_n^{\frac{p}{2}-\frac{1}{2s+1}} & , s > \frac{1}{2} \\ O_{\mathbf{P}}(n^{\frac{1}{4}})\sqrt{M_n^{p-1}\log(n)} & , s = \frac{1}{2}. \end{cases}$$

we have

$$||\hat{\theta}_n - \theta_0|| = O_{\mathbf{P}}(\lambda_n)M_n^{\frac{p}{2}}.$$

We are interested in the particular case

$$\Theta = \{\theta \in C^2([0,1],[0,1]),\ \theta^{''}(x) \leq 1,\ \forall x \in [0,1]\}$$

Using Ball and Pajor's inequality, we have found that $\exists C < \infty$ such that

$$H(\delta, \Theta, ||.||_2) \leq C\delta^{-\frac{2}{3}}.$$

So the M-estimators converge with a rate of convergence of $r_n = n^{-\frac{3}{8}}$ which is not optimal since the minimax rate of convergence is $n^{-\frac{4}{5}}$, which can be found using the optimal bound for the entropy.

Appendix

Proof of lemma 2.4.16:

Proof. To prove this result, we construct two functions belonging to the set Θ, such that θ is a barycenter of these two functions, which contradicts the definition of an extreme point. For this, consider a real $\epsilon > 0$. Due to the continuity of θ and its second derivative, there exists $\delta > 0$ such that for all x in a neighborhood of x_0, $]x_0 - \delta, x_0 + \delta[$ we have

$$\begin{cases} \theta(x) \in]\epsilon, 1-\epsilon[, \\ \theta^{''}(x) < 1 - \epsilon. \end{cases}$$

Define a cut-off function α with the following properties:

- $\alpha \in C^{\infty}([0,1],[0,1])$,
- $\alpha(x) = 1,\ \forall x \in [x_0 - \frac{\delta}{2}, x_0 + \frac{\delta}{2}]$,
- $\alpha(x) = 0,\ \forall x \in [0, x_0 - \delta] \cup [x_0 + \delta, 1]$.

Define for a real $\beta > 0$, the two following functions

$$\begin{cases} h_1 = \theta + \beta\alpha, \\ h_2 = \theta - \beta\alpha. \end{cases}$$

We want to prove that it is possible to choose β such that the two functions belong to Θ.

- $h_1(x) \leq 1$ because $h_1 = \theta$ on the set $[x_0 - \delta, x_0 + \delta]^c$ and on the interval $[x_0 - \delta, x_0 + \delta]$, $h_1(x) \leq \theta(x) + \beta$. But $\epsilon < \theta(x) < 1 - \epsilon$, so if we choose $\beta \leq 1 - \epsilon$, we have $h_1 \leq 1$.
- $h_1(x) \geq 0$ because $\beta > 0$.
- $h_1^{''}(x) \leq 1$. As a matter of fact, on the first interval $h_1^{''}(x) = \theta^{''}(x) \leq 1$. On the second interval we have $h_1^{''}(x) = \theta^{''}(x) + \beta\alpha^{''}(x)$. As soon as $\beta \leq \frac{\epsilon}{||\alpha^{''}||_\infty}$ we have $h_1^{''} \leq 1$.

So we have proven that $h_1 \in \Theta$.
Using similar arguments, we can prove that there exists β such that the two functions h_1 and h_2 belong to Θ. But since $\theta = \frac{h_1+h_2}{2}$ belongs also to Θ due to the convexity property, we have proven that θ does not belong to the set of the extreme points.

□

Proof of lemma 2.4.17:

Proof. We can state that $x_1 < x_2$, the other case $x_1 > x_2$ is handled in a similar way. Moreover we can take $x_2 = \inf\{x > x_1,\, \theta(x) = 1\}$, otherwise there exists a sequence of points $(x_n)_n$ such that $\theta(x_n) = 1$ and $x_n \to x_1$. So by continuity of the function θ we will have $\theta(x_n) = 1 \neq 0$. Using the same arguments we can take $x_1 = \sup\{x < x_2, \theta(x) = 0\}$. So for all $x \in]x_1, x_2[$, $\theta(x) \in]0, 1[$. But since θ is an extreme point and by lemma 2.4.16, we know that for every $x \in [x_1, x_2]$, $\theta^{''}(x) = 1$. So over this interval, $\theta(x) = \frac{1}{2}x^2 + ax + b$. But $\theta^{''} > 0$, so the function is a convex function, and by the property of extrema of convex functions, the function θ has extreme values at the points x_1 and x_2. So $\theta^{'}(x_1) = 0$ and $\theta^{'}(x_2) = 0$. But $\theta^{'}$ is a polynomal of the first degree with only one root. So the preceding statement is impossible since $x_1 \neq x_2$. □

Proof of lemma 2.4.18:

Proof. Let $x_1 < x_2 \in I_0$, and consider a point $x_3 \in]x_1, x_2[$, $\theta(x_3) \neq 0$. By lemma 2.4.17, we know that $\theta(x_3) \neq 1$. Now define

$$\tilde{x}_1 = \sup\{x \in]x_1, x_3[,\ \theta(x) = 0\}$$

$$\tilde{x}_2 = \inf\{x \in]x_3, x_2[,\ \theta(x) = 0\}$$

this two points are well defined (cf proof of previous lemma), and we have $\tilde{x}_1 < x_3 < \tilde{x}_2$, so $\forall x \in]\tilde{x}_1, \tilde{x}_2[, \theta(x) \in]0, 1[$. But $\theta^{''}(x) = 1$, for $x \in]\tilde{x}_1, \tilde{x}_2[$ and $\theta(\tilde{x}_1) = \theta(\tilde{x}_2) = 0$. So

$$\theta = \frac{1}{2}(x - \tilde{x}_1)(x - \tilde{x}_2).$$

But this gives that $\theta(x) < 0$ over $]\tilde{x}_1, \tilde{x}_2[$ which contradicts the definition of the function θ. So x_3 verifies $\theta(x_3) = 0$.
In a similar way for I_1: Let $x_1 < x_2$, such that $\theta(x_1) = \theta(x_2) = 1$. Suppose that there exists x_3, $\tilde{x}_1 < x_3 < \tilde{x}_2$ such that $\theta(x_3) \neq 1$. By lemma 2.4.17, we already know that $\theta(x_3) \neq 0$. θ is a polynomal such that $\theta^{'}(\frac{x_1+x_2}{2}) = 0$. But since $\tilde{x}_1$ and $\tilde{x}_2$ are also two extrema who are in the interior of the interval $[0, 1]$, we have $\tilde{\theta}(x_1) = \tilde{\theta}(x_2) = 0$. So the first derivative of θ has 3 roots, which is impossible, so I_1 is also an interval. □

Chapter 3

Penalization with smoothing penalties

In this chapter, we consider penalized M-estimators where the penalty is chosen in order to control the smoothness of the estimator. For this, we focus on semi-norm penalties such as, for instance, $f \to \int (f^{(s)}(t))^2 dt$ where s is the number of derivatives. Such estimation issue is well-known in the literature and has been studied by several authors such as Wahba in [Wah90] with spline smoothing techniques. Silverman in [Sil85] also tackled the problem. In our work, we will take for penalty, weighted sum of wavelet coefficients. Asymptotically, the penalty will be equivalent to a Besov semi-norm. We will investigate several choices of the penalty and show that, in all cases, the minimization problem has a solution. We will point out the cases where it is possible to give a direct explicit solution of the optimization program and, in the other cases, we will provide an approximation of the exact case. Moreover, the specific case B^s_{22} Besov space will be studied in particular, because of its specific Hilbert properties, and we will give the asymptotic properties of the penalized estimators. The main drawback of such smoothing techniques is that they require the prior knowledge of the function regularity in order to pick a smoothing parameter of the right order to reach the optimal rate of convergence. This major requirement prevents the method from being adaptive unless we try to let the data speak from themselves and automatically choose a parameter close enough to the theoretical value. Indeed, cross-validation method, made popular by G. Wahba in [Wah90], [XW96], [Wah85] or B. Drodge in [Dro96] or [Dro99], will give a partial answer to this question from a practical point of view.

In a second part, we give some details how the hard-thresholded estimator can be understood in terms of a penalized estimator. We will there recall model selection results from L. Birgé and P. Massart.

3.1 Pen(θ) = $||\theta||_{\mathcal{Y}}$

Our model is the standard regression model:

$$\begin{cases} Y_i &= \theta_0(z_i) + W_i, \ i = 1, \dots, n \\ \theta_0 &\in \Theta. \end{cases}$$

In the following part, the errors W_i, $i = 1, \ldots, n$ are assumed to be independent centered random variables with known variance σ^2 while the functions are defined on the interval $[0,1]$ and the $z_i = \frac{i}{n}$ are equispaced.
Take Q_n as the empirical measure of the covariables:

$$Q_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{z_i}.$$

We denote the $L_2(Q_n)$-norm of a function defined in a compact set $\mathcal{Z}$, namely $\theta : \mathcal{Z} \to \mathbf{R}$ as

$$||\theta||_{Q_n} = (\int \theta^2 dQ_n)^{1/2}.$$

The estimator is defined as a solution of the minimization problem:

$$\hat{\theta}_n = \arg \inf_{\theta\in\mathcal{Y}} \left(||Y-\theta||_n^2 + \lambda_n^2 I^p(\theta)\right), \tag{3.1.1}$$

where, for a space $\mathcal{Y}$ and a constant p depending of that choice, we take $I(\theta) = ||\theta||_{\mathcal{Y}}$. The function θ_0 may belong to $\mathcal{Y}$ but not necessarily. So, such an estimator approximates the data Y in the $\mathbb{L}^2$-sense with a smoothness constraint. The balance between these two contributions is determined by the trade-off parameter λ_n^2. The lesser this constant, the more weight is given to the data.

3.1.1 Properties of M-estimator with general Besov-norm

Consider a wavelet basis $(\psi_{jk})_{jk}$ with enough regularity of a Besov space B^s_{pq} with $1 \le p, q \le \infty$ and $s > \frac{1}{p} - \frac{1}{q}$. We consider the cases where the Besov spaces are embedded into $\mathbb{L}^2$ space. Every function can be decomposed onto this basis and we may write:

$$\theta = \sum_{jk} \beta_{jk}\psi_{jk}.$$

First of all, we recall that if $\theta_0 \in B^s_{pq} = \mathcal{Y}$, the pseudo-norm of this space can be described in term of wavelet coefficients:

$$||\theta||_{\mathcal{Y}} = \left(\sum_j (\sum_k 2^{jsp} 2^{\frac{j}{2}(p-2)} |\beta_{jk}|^p)^{q/p}\right)^{1/q}.$$

with the usual modifications when $p = \infty$ or $q = \infty$. Set $\mathcal{Y} = B^s_{pp}$ and $\hat{\alpha}_{jk} = \frac{1}{n}\sum_{i=1}^n Y_i \psi_{jk}(z_i)$ the estimated empirical wavelet coefficient. Then, to the minimization problem

$$\min_{\theta\in\mathcal{Y}} || \; Y - \theta \; ||^2 + \lambda_n^2 \; || \; \theta \; ||^p_{\mathcal{Y}}$$

corresponds the other minimization problem,

$$\inf_{d_{jk}} \sum_{j,k} \left(| \; \hat{\alpha}_{j,k} - d_{jk}|^2 + \lambda_n^2 \sum_j (\sum_k 2^{jsp} 2^{\frac{j}{2}(p-2)} |\beta_{jk}|^p)^{q/p} \right). \tag{3.1.2}$$

So if we want to continue our calculations we shall restrict ourselves to the cases where $p = q$. In that case, the problem decouples in term of wavelet coefficients and the minimization problem can be written as:

$$\inf_{d_{jk}} \sum_{j,k} \left(\mid \hat{\alpha}_{j,k} - d_{jk} |^2 + \lambda_n^2 2^{sjp} 2^{\frac{j}{2}(p-2)} \mid d_{jk} \mid^p \right), \tag{3.1.3}$$

So the penalized estimation problem is turned into an optimization problem: minimizing over x for a fixed t the quantity

$$E(x) = \mid x - t \mid^2 + \mu \mid x \mid^p \tag{3.1.4}$$

for $t = \hat{\alpha}_{jk}$, and $\mu = \lambda_n^2 2^{sjp} 2^{\frac{j}{2}(p-2)}$. This point of view is close to the K-functional introduced by Petree in [NP86] developed in a statistical point of view by DeVore in [DeV98]. Peetre's K-functional is originally used to measure the smoothness of functional spaces. It is defined as follows:

Definition 3.1.1. *Let A and B two Banach spaces and $\lambda_n^2 > 0$ a chosen parameter. Define:*

$$K_p(\lambda, f, A, B) = \inf_{f=a+b} \left(||a||_A^p + \lambda_n^2 ||b||_B^p \right)^{\frac{1}{p}}$$

where $\lambda_n^2 \in \mathbb{R}^+$ and $p < +\infty$; for a function $f \in A + B$ taken as the sum of vector spaces.

In statistical estimation, the K-functional can be used to quantify the quality of the approximation of a function g by a function f in the case where there exists a continuous injection of B into A, $B \to A$. The function represents a trade-off between smoothness and closeness to true data as the penalized estimator (4.2.8). If $B \to A$, then if $\forall a \in A + B = A$, $\exists b$ $(= b(a))$ such that:

$$K_p(\lambda, a, A, B) = \inf_{b \in B} \left(||a - b||_A^p + \lambda_n^2 ||b||_B^p \right)^{\frac{1}{p}}.$$

The function b corresponds to the approximating function.

Existence of the penalized estimator

We have seen that, when the problems decouples in terms of wavelet coefficients, the problem of finding a minimizer is equivalent to:

$$\min_x \left(|x - t|^2 + \lambda_n^2 |x|^p \right)$$

with the following notations:

$$t = \hat{\alpha}_{jk} = \frac{1}{n} \sum_{i=1}^{n} Y_i \psi_{jk}(x_i), \forall (j, k).$$

The following proposition shows that the estimator described in terms of wavelet coefficients is always defined.

Proprosition 3.1.2. *The solution $\bar{x}$ of the problem*

$$l(t) = \min_x \left(|x-t|^2 + \lambda_n^2 |x|^p \right).$$

for $p > 0$ exists and is bounded :

$$|\bar{x}| \leq |t|.$$

Remark 3.1.3. *We point out that although the solution exists for any positive p, we will only consider Besov spaces B_{pp}^s for $p \geq 1$. Nevertheless the case $p = 0$ covers the hard-thresholding case as we will explain it later in this paper, and is still of interest even if it is here out of context.*

Proof. The proof of the proposition relies on the existence of extrema of a continuous function over a compact set. Indeed, if we try to minimize the function over the compact set $[-|t|, |t|]$, by continuity, at least one minimum exists.
Now if x does not belong to this set, first consider $\tilde{x}$ in that set where

$$\tilde{x} = (x \wedge |t|) \vee (-|t|).$$

We can see that the minimizer $\tilde{x}$ has the following properties:

- $|\tilde{x}| \leq |x|$.
- $|\tilde{x}| \leq |t|$.
- $|\tilde{x} - t| \leq |x - t|$

So we have

$$\begin{aligned}(\bar{x}-t)^2 + \lambda_n^2 |\bar{x}|^p &\leq |\tilde{x}-t|^2 + \lambda_n^2 |\tilde{x}|^p \\ &\leq |x-t|^2 + \lambda_n^2 |x|^p.\end{aligned}$$

So the minimum over the compact set $[-|t|, |t|]$ is also minimum for the function over the whole set. □

So the M-estimator minimizing a quadratic loss with a B_{pp}^s-norm penalty is always defined. The choice of the penalty will determine the behavior of the estimator. The Besov spaces B_{pp}^s are characterized by two parameters p and s. The first one stands for an L^p-parameter while the other is a smoothness parameter similar to a number of weak derivatives of the function. We now look for the properties of the minimization problem as a function of the parameter p.

Asymptotic property of the minimizer for different $\mathbb{L}^p$-choices

We set, for a given $\lambda = \lambda_n^2$, and $\bar{x}(p)$ the solution of

$$\bar{x}(p) = \arg\min_x \left((t-x)^2 + \lambda|x|^p\right) \tag{3.1.5}$$

We begin with the sub-case $t \geq 0$. We will see that the other case can be handled with similar arguments. Since we have proven that $0 \leq \bar{x}(p) \leq t$, the solution belongs to a compact set for all real values of p, so the convergence of $\bar{x}(p)$, as a function of p, is equivalent to the existence of a unique adherent value.

- $p \to 1$: we have already seen that this case corresponds to the soft-thresholded estimator with a threshold value equal to $\frac{\lambda}{2}$.
- $p \to +\infty$: we have for all positive value t:

  $$2(t-x) + p\lambda x^{p-1} = 0.$$

  So if $t < 1$, the bounded property of t $\bar{x}(p)$ gives $p\bar{x}(p) \to 0$ so

  $$\lim_{p\to+\infty} \bar{x}(p) = t.$$

  If $t \geq 1$ then we distinguish two cases: $\bar{x}(p) < 1$: this case is impossible because this will lead to the contradiction $\bar{x}(p) = t$. If $\bar{x}(p) > 1$, then $p\bar{x}(p)^{p-1} \to +\infty$ but this is in contradiction with the fact that $t - \bar{x}(p)$ remains bounded. So we have $\bar{x}(p) = 1$.

The solution of the problem for positive t is $\min(1,t)$. By antisymmetry, the same arguments still hold and we obtain, for all real t:

$$\lim_{p\to+\infty} \bar{x}(p) = \min(\max(-1,t),1). \tag{3.1.6}$$

As a result, when p tends to infinity, only small coefficients will be properly estimated and it will provide a too rough estimator. So we will mainly pay attention in the following part to the cases where $p \leq 3$.

Remark 3.1.4. *To determine the asymptotic behavior of M-estimators, we could apply the general theorem for penalized M-estimation described in the first part in Section 2.1. As a matter of fact, entropy of Besov spaces can be calculated using the sequential definition with wavelet coefficients as it is done in Loubes and van de Geer in [LvdG00] for instance. However entropy methods does not always provide optimal asymptotic results due to the use of entropy upper bound and not exact calculations. That is the reason why we look into particular cases.*

Remark 3.1.5. *An explicit expression of the solution of the minimization problem does not often exist. But as it is done in [DL93] we can find an approximated solution. We can consider without any loss of generality that, for any positive t,* $x \in [0,t]$. *Indeed if not, it is obvious that by changing the value of* x *for a value in the interval* $[0,t]$, *we reduce either* $|x-t|^2$ *or*

$\mu|x|^p$. If $| x |\leq| t | /2$, we have $E(x)$, defined in (3.1.4), is no less than $t^2/4$. Moreover if $| x |\geq| t | /2$, this time $E(x)$ is no less than $\mu | t |^p /2^p$. If $x = 0$, $| t |^2\leq \mu | t |^p$. If $x = t$, then $| t |^2\geq \mu | t |^p$. so we find

$$E(x) = \min(| t |^2, \mu | t |^p)$$

and we have

$$x = t1_{|t|^2\geq\mu|t|^p}.$$

The minimum of the function differs from a factor 4 or 2^p. As a result hard-thresholded estimator with the proper thresholding level provides an approximation of the penalized estimator.

Remark 3.1.6. *We can point out some interesting cases where direct calculations lead to exact results. These special choices of p and Besov spaces, which have already been described in Chapter II, are the following:*

Particular cases with explicit expressions:

- *$p = 0$: we have already seen that this particular case does not correspond to a Besov norm, since Besov spaces B^s_{pq} are only defined for $p \geq 1$. But this case enables us to put together the two main types of thresholded estimators. Such estimators are used in estimation with wavelet basis and have been studied by D. Donoho, I. Johnstone, G. Kerkyacharyan and D. Picard in [DJKP95], [DJKP96b], [DJKP97], [DJ98], [DJ99], or B. Delyon and A. Juditsky in [DJ96a] for instance. Other reference is Antoniadis in [AGG97]. We can see that for $p = 0$ we obtain also a minimum of the type hard-thresholding since:*

  $$\lim_{p\to 0} |x|^p = 1_{|x|\neq 0}.$$

  So we find the well-known hard-thresholded estimator.

  $$\tilde{\theta}_n = \sum_{(j,k)} \hat{\alpha}_{jk} 1_{|\hat{\alpha}_{jk}|>\lambda_n^2} \psi_{jk}.$$

- *$p = 1$: we find the minimum by differentiating the constraint function and analyzing whether the minimum is in the open set, which corresponds to a zero of the differentiated loss-function, or in the boundaries of the set: it is given by*

  $$\tilde{x} = \mathrm{sgn}(x)(|x| - \lambda_n^2/2)_+.$$

  We can recognize the soft-thresholded estimator.

  $$\tilde{\theta} = \sum_{(j,k)} \mathrm{sgn}(\hat{\alpha}_{jk})(|\hat{\alpha}_{jk}| - \frac{\lambda_n^2}{2})_+\psi_{jk}.$$

- *$p = 2$: this time the result is a classical smoothed estimator obtained again by solving a first order condition.*

  $$\tilde{\theta} = \sum_{(j,k)} \frac{1}{1+\lambda_j^2} \hat{\alpha}_{jk} \psi_{jk},$$

  for $\lambda_j^2 = \lambda_n^2 2^{js} 2^{-\frac{j}{2}}$.

- *$p = 3$: the result is not obvious: if we differentiate the expression, we get:*

  $$\begin{aligned}
  2(x-t) + 3\mathrm{sgn}(x)\lambda_j^2 x^2 &= 0 \\
  \mathrm{sgn}(x)(3\lambda_j^2 x^2 + 2|x|) &= 2\mathrm{sgn}(t)|t| \\
  \mathrm{sgn}(x) &= \mathrm{sgn}(t) \\
  |x| &= \frac{-1 + \sqrt{1 + 6\lambda_j^2 |t|}}{3\lambda_j^2} \\
  x &= \frac{2|t|\mathrm{sgn}(x)}{1 + \sqrt{1 + 6\lambda_j^2 |t|}} \\
  x &= \frac{2t}{1 + \sqrt{1 + 6\lambda_j^2 |t|}}.
  \end{aligned}$$

  So we have

  $$\tilde{\theta}_n = \sum_{(j,k)} \frac{2\hat{\alpha}_{jk}}{1 + \sqrt{1 + 6\lambda_j |\hat{\alpha}_{jk}|}} \psi_{jk},$$

  with $\lambda_j = \lambda_n^2 2^{3js} 2^{\frac{j}{2}}$.

- *$\frac{1}{p} = \frac{s}{2} + \frac{1}{2}$. In that case, if we look at the approximated solution of the minimization problem, we can see that $\lambda = \mu$ and $s = t1_{|t| \geq \lambda^{\frac{1}{2-p}}}$. The threshold does not depend on the position of the coefficients but is the same for all. This subcase deals with the Besov space of minimal smoothness to be embedded in an L^2- space.*

Asymptotic Behavior of approximate solution

In this section we study the asymptotic behavior of the approximate estimator of the problem, which corresponds to the hard-thresholded estimator

$$\tilde{\theta}_n = \sum_{(j,k)} \hat{\alpha}_{jk} 1_{|\hat{\alpha}_{jk}| > \lambda_n^{\frac{2}{2-p}}} \psi_{jk},$$

where we have set $\hat{\alpha}_{jk} = \frac{1}{n}\sum_{i=1}^n Y_i \psi_{jk}(z_i)$ the estimated empirical wavelet coefficient. Moreover we assume that our data are dyadic: $Y_1, \ldots, Y_{2^m}$ and set for simplicity reasons $n = 2^m$. The unknown function is supposed to belong to Besov spacs $B_{pp}^s \cup H^s$. This estimator is an threshold estimator. We will describe the asymptotic behavior of both the hard-thresholded and soft-thresholded estimator. For this set $L = \lambda_n^{\frac{2}{2-p}}$.

1. Linear estimator:
   We consider that the true signal function θ_0 can be written as follows:

$$\theta_0 = \sum_{(j,k)} \alpha_{jk} \psi_{jk},$$

with $\alpha_{jk} = \frac{1}{n} \sum_{i=1}^n \theta_0(z_i) \psi_{jk}(z_i)$. If we do not take into account every coefficients but truncate at a resolution level J, we define θ_J the projection of the estimator onto V_J defined as the approximation space with resolution J by:

$$\theta_J = \sum_{j<J} \sum_k \alpha_{jk} \psi_{jk}.$$

Then the error between θ_0 and is projection onto the spave V_J is:

$$\begin{aligned} ||\theta_0 - \theta_J||^2 &\leq \sum_{j>J} \sum_k \frac{2^{2sj}}{2^{2sJ}} \alpha_{jk}^2 \\ &\leq 2^{-2sJ} \sum_{jk} (2^{js} |\alpha_{jk}|)^2 \\ &\leq 2^{-2Js} ||\theta_0||^2_{H^s} \end{aligned}$$

and

$$||\theta_0||^2_{H_s} = \sum_{jk} (2^{js} |\alpha_{jk}|)^2.$$

Now we study the error with the linear estimator of θ_0 when we replace the wavelet coefficient by the estimated coefficient based on the observations and when thresholding the coefficients at a fixed resolution level J. The estimated coefficient is

$$\begin{aligned} \hat{\alpha}_{jk} &= \frac{1}{n} \sum_{i=1}^n y_i \psi_{jk}(z_i) \\ &= \alpha_{jk} + \eta_{jk} \end{aligned}$$

where $\alpha_{jk} = \frac{1}{n} \sum_{i=1}^n \theta(z_i) \psi(z_i)$, and $\eta_{jk} = \frac{1}{n} \sum_{i=1}^n \epsilon_i \psi_{jk}(z_i)$ is an error of zero-mean and of variance $\sigma^2 = \frac{\sigma_0^2}{n}$. So we have

$$\begin{aligned} E||\theta_0 - \hat{\theta}_J||_2^2 &\leq \sum_{j<J} \sum_k E|\alpha_{jk} - \hat{\alpha}_{jk}|^2 + \sum_{j\geq J} \sum_k |\alpha_{jk}|^2 \\ &\leq \sum_{j<J} \sum_k E|\eta_{jk}|^2 + 2^{-2sJ} ||\theta_0||^2_{H_s}. \end{aligned}$$

Using the first inequality we can deduce that

$$E||\theta - \hat{\theta}_J||_2^2 \leq 2^{J-m} \sigma_0^2 + 2^{-2Js} ||\theta||^2_{H_s}.$$

To minimize this quantity, we choose J such that the two terms have the same order, which implies

$$2^J = \left(\frac{n||\theta_0||^2_{H_s}}{\sigma_0^2}\right)^{1/(s+1)},$$

so the rate of convergence is in $n^{-\frac{2s}{2s+1}}$, the classical Besov's rate of convergence for a linear estimator.

2. Non-linear estimator:
Now we calculate the error when using non-linear estimator, based on the hard and soft-thresholding of the coefficients in the wavelet decomposition. Write δ the thresholding operator and L the threshold level, the estimator is

$$\hat{\theta}_n = \sum_{j=1}^{J}\sum_k \delta(\hat{\alpha}_{jk})\psi_{jk}.$$

The estimation error for the quadratic norm is then:

$$\begin{aligned} E\|\hat{\theta}_n - \theta_0\|^2 &\leq 2(E\|\hat{\theta}_n - \theta_J\|^2 + E\|\theta_J - \theta_0\|^2) \\ &\leq 22^{-2Js} + 2(E\|\hat{\theta}_n - \theta_J\|^2). \end{aligned}$$

For J large enough, the first approximation term goes to zero faster than the second term which determines the rate of convergence. To study the term $E\|\hat{\theta}_n - \theta_J\|^2$ we point out that

$$\begin{aligned} E\|\hat{\theta}_n - \theta_J\|^2 &= \sum_{j\leq J}\sum_k |\delta(\hat{\alpha}_{jk}) - \alpha_{jk}|^2 \\ &= \sum_{j\leq J}\sum_k |\delta(\alpha_{jk} + \epsilon_{jk}) - \alpha_{jk}|^2 \end{aligned}$$

and we use the following lemma depending on the type of the threshold operator. For this set the hard thresholded operator with level L

$$\delta_h(x) = x1_{|x|\geq L} \tag{3.1.7}$$

and the soft thresholded operator with level L

$$\delta_s(x) = \text{sgn}(x)(|x| - L)_+ \tag{3.1.8}$$

Lemma 3.1.7. • *For $\epsilon \geq 0$, the following inequality holds:*

$$|\delta_h(t+\epsilon, L) - t| \leq \begin{cases} 2|\epsilon| & , |t| > 2L, \\ \max(|t|, |\delta_h(\epsilon, \frac{L}{2})|) & , |t| \leq 2L. \end{cases}$$

- *For $\epsilon \geq 0$, we have:*

$$|\delta_s(t+\epsilon, L) - t| \leq \begin{cases} |\epsilon - L| & , t > L, \\ \max(|t|, \delta_s(\epsilon, L)) & , -L - t \leq \epsilon \leq L - t, \\ |\epsilon + L| & , t < -L. \end{cases}$$

Moreover, we have to control the two main quantities, for a given resolution level L

$$\mathrm{Card}\{\alpha_{jk},\ |\alpha_{jk}| > L\} \tag{3.1.9}$$

$$\sum_{j \leq J} \sum_{|\alpha_{jk}| \leq L} \alpha_{jk}^2 \tag{3.1.10}$$

For this we use the two following lemmas in the general case for a function in a Besov space B^s_{pq}. We recall that, for a function θ on $[0,1]$ in the Besov space B^s_{pq} with smoothness $s > \frac{1}{2}$, and with $p \geq 1$, $q \geq 1$, there are orthonormal basis functions $\psi_{j,k}$ in L_2(Lebesgue measure), such that

$$\theta = \sum_{j=1}^{\infty} \sum_{k=0}^{2^j - 1} \alpha_{j,k} \psi_{j,k},$$

where the coefficients $\{\alpha_{j,k}\}$ satisfy

$$\left(\sum_{j=1}^{\infty} 2^{j((2s+1)\frac{p}{2} - 1)\frac{q}{p}} \left\{ \sum_{k=1}^{2^j} |\alpha_{j,k}|^p \right\}^{\frac{q}{p}} \right)^{\frac{1}{q}} \leq 1 \tag{3.1.11}$$

We shall simply take (3.1.11) as a starting point. This is called Condition 1.

Condition 1.

$$\sum_{j=1}^{J} 2^{jr} \left\{ \sum_{k=1}^{2^j} |\alpha_{j,k}|^p \right\}^m \leq 1.$$

In Condition 1, it is assumed that $p \geq 1$, $mp \geq 1$ and $r \geq 0$. In the Besov B^s_{pq}-case, one has $m = \frac{q}{p}$, and $r = ((2s+1)\frac{p}{2} - 1)\frac{q}{p}$, and $J = \infty$.

Lemma 3.1.8. *Suppose that Condition 1 holds with $m + r \geq 1$ and $J < \infty$. Then*

$$\sum_{j=1}^{J} \sum_{k=1}^{2^j} |\alpha_{j,k}|^{\frac{mp}{m+r}} \leq J^{\frac{m+r-1}{m+r}}.$$

Corollary 3.1.9. *Suppose Condition 1 holds, with $m + r \geq mp$ and $J < \infty$. Then*

$$\sum_{j \leq J} \sum_{|\alpha_{j,k}| \leq \epsilon} |\alpha_{j,k}|^2 \leq J^{\frac{m+r-1}{m+r}} \epsilon^{2-\frac{mp}{m+r}},$$

and moreover

$$\{\#\alpha_{j,k},\ 1 \leq j \leq J :\ |\alpha_{j,k}| > \epsilon\} \leq J^{\frac{m+r-1}{m+r}} \epsilon^{-\frac{mp}{m+r}}.$$

The following lemma shows that is is possible to extend the last result to the case $J = \infty$.

Lemma 3.1.10. *Suppose that Condition 1 holds, with $m + r > 1$ and $J = \infty$. Then for all $0 < \epsilon \leq 1$,*

$$\{\#\alpha_{j,k} : |\alpha_{j,k}| > \epsilon\} \leq (\frac{mp}{r} \log_2 \frac{1}{\epsilon})^{\frac{m+r-1}{m+r}} \epsilon^{-\frac{mp}{m+r}}.$$

Lemma 3.1.11. *Suppose that Condition 1 holds with $m + r > mp$ (J may be infinite). Then for a constant c depending on m, p and r, and for all $\epsilon \leq 1$,*

$$\sum_{j=1}^{J} \sum_{|\alpha_{j,k}| \leq \epsilon} |\alpha_{j,k}|^2 \leq c\epsilon^{2-\frac{mp}{m+r}}.$$

These preliminary results enable us to prove the consistency of both soft and hard-thresholded estimator.

(a) Hard-Thresholding:
Using previous lemma, we have

$$|\delta_h(s + \epsilon, L) - s| = \begin{cases} |\epsilon|, |s + \epsilon| > 2L \\ |s|, |s + \epsilon| \leq 2L. \end{cases}$$

So, using this inequality, we obtain:

$$\begin{aligned} E||\theta_J - \hat{\theta}_n||^2 &\leq \sum_{|\alpha_{jk}| > 2L} 4\frac{\sigma_0^2}{n} + \sum_{|\alpha_{jk}| \leq 2L} (\alpha_{jk})^2 + E(t^2_{L/2}(\epsilon_{jk})) \\ &\leq (1) + (2) + (3) \end{aligned}$$

Using lemma 3.1.10, we have if we set $L = \frac{\sigma_0 t}{\sqrt{n}}$:

$$\begin{aligned} (1) &= \sum_{|\alpha_{jk}| > 2L} 4\frac{\sigma_0^2}{n} \\ &\leq \frac{4\sigma_0^2}{n} \mathrm{Card}\{|\alpha_{\mathrm{jk}}| > 2\mathrm{L}\} \\ &\leq c_1 \frac{4\sigma_0^2}{n} (\frac{n}{t})^{-\frac{2s}{2s+1}} (\log \frac{n}{t})^{\frac{r}{1+r}} \end{aligned}$$

for a positive finite constant c_1. For the second term, use lemma 3.1.11:

$$\begin{aligned}(2) &= \sum_{|\alpha_{jk}|\leq 2L} (\alpha_{jk})^2 \\ &\leq c_2 \frac{n^{-\frac{2s}{2s+1}}}{t}\end{aligned}$$

for c_2 a finite positive constant. For the last term we have:

$$\begin{aligned}(3) &= \sum_{|\alpha_{jk}|\leq 2L} E\delta_h(\epsilon_{jk}, L/2)^2 \\ &\leq 2n \int_{L/2}^{\infty} x^2\theta_\sigma(x)\,dx \\ &\leq 2\sigma_0^2 \int_{\frac{L\sqrt{n}}{2\sigma_0}}^{\infty} x^2\theta_1(x)\,dx \\ &\leq \underbrace{2\sigma_0^2 \int_{t/2}^{\infty} x^2\theta_1(x)\,dx}_{\asymp t\exp(-t^2/2)}\end{aligned}$$

So we can choose t such that the two terms are of the same order and we obtain that there exists a constant c such that

$$E||\hat{\theta}_n - \theta_J||_2^2 \leq cn^{-\frac{2s}{2s+1}}(\log n)^{\frac{2s}{2s+1}} \tag{3.1.12}$$

(b) Soft-Thresholding:
If we follow the same proof and if we use the other upper bound of lemma 19, we have this time:

$$|t - \delta_s(t+\epsilon)| = \begin{cases} |\epsilon - L| & , L - t \leq \epsilon \\ |t| & , -L - t \leq \epsilon \leq L - t \\ |\epsilon + L| & , \epsilon \leq -L - t. \end{cases}$$

So we have the following inequality:

$$E||\theta_J - \hat{\theta}_n||_2^2 \leq \begin{cases} L^2 + \sigma^2 & , |t| > L \\ t^2 + E(\delta_s(\epsilon))^2 & , |t| \leq L. \end{cases}$$

Using this result, we can calculate the square risk of the soft-thresholded estimator:

$$\begin{aligned}E||\theta_J - \hat{\theta}_n||_2^2 &\leq \sum_{|\alpha_{jk}|>L} (L^2 + \sigma^2) + \sum_{|\alpha_{jk}|\leq L} (\alpha_{jk}^2 + E(\delta_s(\epsilon_{jk})^2)) \\ &\leq (1) + (2) + (3)\end{aligned}$$

Using lemma 3.1.10 where we have $m = 1$ and $r = p(s + \frac{1}{2} - \frac{1}{p})$, we have

$$\begin{aligned}
(1) &= \sum_{|\alpha_{jk}|>L} (L^2 + \sigma^2) \\
&\leq (L^2 + \sigma^2)\mathrm{Card}\{|\alpha_{\mathrm{jk}}| > \mathrm{L}\} \\
&\leq c_1 L^{2 - \frac{p}{1+r}} (\log(L))^{\frac{r}{1+r}} \\
&\leq c_1 L^{\frac{4s}{2s+1}} (\log L)^{\frac{r}{1+r}} \\
&\leq c_2 \left(\frac{n}{t}\right)^{\frac{-2s}{2s+1}} (\log \frac{\sqrt{n}}{t})^{\frac{r}{1+r}}
\end{aligned}$$

where we have set $L = \frac{\sigma_0 t}{\sqrt{n}}$ and c_1 a finite positive constant. Using lemma 3.1.11, we can give an upper bound for the second term.

$$\begin{aligned}
(2) &= \sum_{|\alpha_{jk}|\leq L} \alpha_{jk}^2 \\
&\leq c_2 L^{2 - \frac{p}{1+r}} \\
&\leq c_2 n^{\frac{-2s}{2s+1}}
\end{aligned}$$

for a finite positive constant c_2. The last term can be upper bounded as follows:

$$\begin{aligned}
(3) &= \sum_{|\alpha_{jk}|\leq L} E(\delta_s(\epsilon_{jk})^2) \\
&\leq 2n \int_L^\infty (x - L)^2 \theta_\sigma(x)\, dx \\
&\leq 2\sigma_0^2 \int_{\frac{L\sqrt{n}}{\sigma_0}=t}^\infty (x - \frac{L\sqrt{n}}{\sigma_0})^2 \theta_1(x)\, dx \\
&\leq \underbrace{2\sigma_0^2 \int_t^\infty (x - t)^2 \theta_1(x)\, dx}_{= \frac{4\sigma_0^2}{t^3} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2) + O(1/t^5)}.
\end{aligned}$$

So we choose t such that the two terms have the same order:

$$\exp(-t^2/2) = \frac{{\sigma_0}^{-p}}{\sqrt{n}^{2-p}} ||\theta||^p_{B^s_{pp}},$$

which gives a threshold

$$t = \sqrt{(2 - p)\log(n) - 2p\log(\frac{||\theta||^s_{B^s_{pp}}}{\sigma_0})}.$$

If we let p decrease to 0, we can recognize the universal threshold proposed by D.Donoho and I.Johnstone in [DJ95]. This gives a rate of convergence of

$$n^{-\frac{2s}{2s+1}} (\log n)^{\frac{2s}{1+2s}}$$

which is the minimax rate of convergence up to logarithmic factors.

3.1.2 Penalty $I(\theta) = ||\theta||_{\mathcal{Y}}$ where $\mathcal{Y} = B^s_{22}$

This sub-case is close to an easier case where we consider functions lying in a Sobolev space, since the Besov space B^s_{22} is isomorphic to the Sobolev space H^s. So we can show the convergence of the smoothed estimator as in the first part of our study but this time in the inner norm of the space. We begin by studying this easier case.

Sobolev space

$I(\theta) = \sum_{j=1}^n w_j^2 \beta_j^2$
We make the assumption that the signal function θ_0 belongs to a Sobolev space H^s and is bounded. As a matter of fact, there exists a real s such that for all $t \leq s$, $f \in \mathbb{L}^2$ and $f^{(t)} \in \mathbb{L}^2$. Our aim is to reconstruct the signal function using this information, so we define the penalized estimator as the solution of

$$\hat{\theta}_n = \arg\min_{\theta \in H^s} \left(||Y - \theta||_n^2 + \lambda_n^2 \int (f^{(s)})^2(t)dt \right) \tag{3.1.13}$$

We aim at considering a sequential version of Sobolev semi-norm and we would like to take for penalty a sum of weighted coefficients with properly chosen coefficients. For this, we use results that have been described by Silverman in [Sil82]: there exists a basis (ψ_i) such that, if we consider a decomposition of any θ onto this basis we get: $\theta = \sum_{j=1}^n \alpha_j \psi_j$ and if we consider weights $w_i = i^s \ \forall i \in [1, n]$ then the equivalent semi-norm is given by: $\sum_{i=1}^n w_i^2 \alpha_i^2$. As a consequence the estimator $\hat{\theta}_n = \sum_{j=1}^n \tilde{\alpha}_j \psi_j$ is defined as :

$$\hat{\theta}_n = \arg\min_{\beta_j \in \mathbb{R}^n} \left(\sum_{j=1}^n |\hat{\alpha}_j - \beta_j|^2 + \lambda_n^2 \sum_{j=1}^n i^{2s} \beta_j^2 \right) \tag{3.1.14}$$

To minimize the penalized loss function we write first order condition and in that case the minimum is reached at the point where the differentiated function has 0 for value, so we just differentiate the expression and find for a given smoothing parameter λ_n^2:

$$\left\{ \begin{array}{l} \tilde{\alpha}_j = \frac{1}{1+w_j^2\lambda^2} \hat{\alpha}_j \\ \hat{\alpha}_j = \frac{1}{n} \sum_{i=1}^n Y_i \psi_j(z_i). \end{array} \right. \tag{3.1.15}$$

This estimator is a linear smoothed estimator, similar as the one described by Devore in [DeV98]. Although it is not a robust estimator, this estimator achieves the minimax rate of convergence for a good smoothing parameter as it is shown in the following theorem. This result has been stated for instance by Ibragimov and Nemirovski in [INK86] and [HI86] using kernel estimators.

Theorem 3.1.12. *The penalized least square estimator with a Sobolev norm penalty is a consistent estimator. If $\lambda_n^2 = n^{-\frac{2s}{1+2s}}$ it achieves the minimax rate of convergence for a quadratic loss over Sobolev balls $\{\theta \in \Theta, \in (f^{(s)}(t))^2 dt \leq M\}$. There exists a finite constant c depending only on M and σ^2 the variance errors such that*

$$E||\theta_0 - \hat{\theta}_n||^2 \leq c n^{-\frac{2s}{2s+1}} \tag{3.1.16}$$

Proof. The proof of this result is postponed in the appendix. We decompose the mean square errors into a bias term and a variance term. Using properties of the basis and of Sobolev spaces, we find an asymptotic development for these two terms. So we have

$$\begin{aligned} E||\hat{\theta}_n - \theta_0||_n^2 &= \text{Bias}^2 + \text{Variance} \\ &\approx \lambda_n^2 + \frac{\sigma^2}{n}\left(\frac{1}{\lambda_n^2}\right)^{\frac{1}{s}} \end{aligned}$$

Choosing the optimal smoothing parameter λ_n^2 in a way that it tends to zero but not too quickly such that there is an effective smoothing effect, leads to the minimax rate of convergence $n^{-\frac{2s}{2s+1}}$. □

This result could have been viewed as consequence of the theory of the regression over a Besov ball using entropy conditions. Indeed we recall the main theorem for penalized least squares estimators:

Theorem 3.1.13. *Suppose that for some fixed constants $\eta > 0$, $A > 0$ and $s \geq \frac{1}{2}$, one has*

$$H(\delta, \{\theta \in \Theta : \ I(\theta) \leq M, \ ||\theta - \theta_0|| \leq \eta\}) \leq A\left(\frac{M}{\delta}\right)^{\frac{1}{s}}, \text{ forall}\delta > 0, \ n \geq 1,$$

for all $M \geq M_n$, where $M_n \geq I(\theta_0)$. Then for

$$\lambda_n^{-1} = \begin{cases} O_{\mathbf{P}}(n^{\frac{s}{2s+1}} M_n^{\frac{p}{2} - \frac{1}{2s+1}}), & \text{if } s > \frac{1}{2}, \\ O_{\mathbf{P}}(n^{\frac{1}{4}}) M_n^{\frac{p}{2} - \frac{1}{2}} (\log n)^{-\frac{1}{2}}, & \text{if } s = \frac{1}{2}, \end{cases}$$

the estimator $\hat{\theta}_n$ satisfies

$$||\hat{\theta}_n - \theta_0|| = O_{\mathbf{P}}(\lambda_n) M_n^{\frac{p}{2} - \frac{1}{2s+1}}.$$

Indeed, consider a ball of a Sobolev space:

$$\Theta = \{(\alpha) \in \mathbb{R}^{\mathbb{N}}, \sum_j w_j^2 \alpha_j^2 \leq 1 \, , ||\alpha|| \leq R\}$$

the entropy of this set can be easily calculated in the following lemma:

Lemma 3.1.14. *For every positive δ, there exists a finite constant C such that:*

$$H(\delta, \Theta, ||.||) \leq C.\delta^{-\frac{1}{s}} \log\left(\frac{R}{\delta}\right) \tag{3.1.17}$$

Proof. The proof relies on elementary combinatorics and can be found in the appendix. □

A more refined proof that can be found in Birman and Solomjak in [BS67] shows that the entropy of the set can be bounded by:

$$H(\delta, \Theta, ||.||) \leq C\delta^{-\frac{1}{s}} \tag{3.1.18}$$

As a consequence, the theorem applies for the estimation problem with a Sobolev norm penalty: the hypothesis over the entropy of the set are fulfilled with $M_n = 1$. So the theorem gives the rate of convergence in $n^{\frac{-2s}{2s+1}}$.
However, the rate of convergence depends on an optimal choice of the smoothing parameter. The optimal parameter exists but still depends on the data and on the smoothness coefficient s, indeed $\lambda_{\text{optimal}} = n^{-\frac{2s}{2s+1}}$. Hence the estimation method is not adaptive.

3.1.3 Besov space B^s_{22}.

In this part, we consider B^s_{22} Besov spaces with $s > \frac{1}{2}$. We assume that our data are dyadic, and decompose the function onto the approximation space V_{j_0} for j_0 properly chosen. We consider the M-estimator:

$$\hat{\theta}_n = \arg \min_{\theta = \sum_{(j,k)} \beta_{jk}\psi_{jk}} \left(||Y - D_{01}\theta||^2 + \lambda_n^2 ||\theta||^2_{B^s_{22}} \right) \tag{3.1.19}$$

where $D_{01}\theta = \sum_{j=j_0}^{j_1} \sum_{k=0}^{2^j} \beta_{jk}\psi_{jk}$ is the projection of the function onto the spaces with resolution between j_0 and j_1. We decompose any θ onto a wavelet basis and with V_{j_0} an approximation space we have:

$$\theta = \sum_k \alpha_k \psi_k + \sum_{j \geq j_0} \sum_k \beta_{jk}\psi_{jk}$$

Define also

$$\begin{cases} \hat{\alpha}_k & = \frac{1}{n} \sum_{i=1}^n \phi_k(z_i) Y_i \\ \hat{\beta}_{jk} & = \frac{1}{n} \sum_{i=1}^n \psi_{jk}(z_i) Y_i \end{cases}$$

Then the original problem (3.1.19) can be rewritten as follows:

$$\hat{\theta}_n = \arg \min_{(\alpha_k, \beta_{jk})} \left(\sum_k |\alpha_k - \hat{\alpha}_k|^2 + \sum_{j=j_0}^{j_1} \sum_k |\hat{\beta}_{jk} - \beta_{jk}|^2 + \lambda_n^2 \sum_{j \geq j_0} \sum_k 2^{2js} \beta_{jk}^2 \right)$$

With this norm-type penalty, the problem has an explicit solution :

$$\hat{\theta}_n = \sum_k \tilde{\alpha}_{j_0k}\phi_{j_0k} + \sum_{j_1 \geq j \geq j_0, k} \tilde{\beta}_{jk}\psi_{jk}$$

where

$$\begin{cases} \tilde{\alpha}_{j_0k} & = \hat{\alpha}_{j_0k}, \\ \tilde{\beta}_{jk} & = \frac{\hat{\beta}_{jk}}{1 + \lambda^2 2^{2js}}, \quad j_1 \geq j \geq j_0. \end{cases}$$

So there exists always a solution which depends on the smoothing parameter λ_n^2.

Optimal choice of smoothing parameter

Since the asymptotic behavior of the estimator depends on λ_n^2, our aim is to adjust the parameter level in order to find the best estimator possible, in terms of quadratic norm. The choice of the parameter λ_n is quite difficult. It must be chosen not too large, but still large enough such that the trade off may occur. We can wonder whether an optimal parameter exists, optimal in the sense that the risk function, taken as a function of λ, has a minimum at that value. Indeed we look for :

$$\tilde{\lambda}_n = \arg\min_l E||\hat{\theta}(l) - \theta_0||_n^2.$$

Theorem 3.1.15. *The best choice in terms of quadratic loss is always defined under the hypothesis (I) defined as follows:*
Hypothesis (I): $j_0 = 0(1)$ *and* $2^{j_1} = 0\left(\frac{n}{\log(n)}\right)$. *Indeed there exists* $\tilde{l}_n$ *such that*

$$\tilde{l}_n = \arg\min_l E||\hat{\theta}(l) - \theta_0||_n^2.$$

This theorem shows that there exists a smoothing parameter such that the corresponding penalized estimator minimizes the quadratic norm $\text{MSE}(\lambda_n^2) = E\|\hat{\theta}_n - \theta_0\|^2$. However, this theoretical value depends on the unknown function θ_0. So such criterion must be replaced by another one that does not use the expression of θ_0.
Cross validation technics are usually used to tackle this issue of choosing a data-dependent smoothing parameter in various fields: selecting a bandwidth in kernel estimation, a threshold level in thresholding methods or a smoothing parameter in penalized estimation. The main idea is to minimize a quantity which is equivalent to the mean square errors but only involve known quantities. Here we will consider Generalized Full Cross Validation as it is described by Droge in [Dro96]. Consider $\tilde{f}^{(i)}$ the estimator constructed without using data Y_i:

$$\tilde{f}^{(i)}_{\lambda_n^2} = \sum_k \tilde{\alpha}_k^{(i)} \phi_k + \sum_{j \geq j_0} \sum_k \tilde{\beta}_{jk}^{(i)} \psi_{jk} \tag{3.1.20}$$

with

$$\begin{cases} \tilde{\alpha}_k^{(i)} & = \frac{1}{n-1} \sum_{l \neq i} Y_l \psi_k(z_l) \\ \tilde{\beta}_{jk}^{(i)} & = \frac{1}{1+\lambda_n^2 2^{2js}} \frac{1}{n-1} \sum_{l \neq i} Y_l \psi_{jk}(z_l) \end{cases}$$

Then define the criterion as:

$$\text{GFCV}(\lambda_n^2) = \frac{1}{n} \sum_{i=1}^n (Y_i - \tilde{f}^{(i)}_{\lambda_n^2}(z_i))^2 (1 + H(\lambda_n^2))^2$$

for

$$H(\lambda_n^2) = \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{n} \sum_k \phi_k^2(z_i) + \sum_{j=j_0}^{j_1} \sum_k \frac{1}{1 + \lambda_n^2 2^{2js}} \psi_{jk}^2(x_i) \right).$$

The following theorem shows the asymptotic equivalence between the two quantities: mean square errors and cross validation criterion. As a consequence, the smoothing parameter numerically found by minimizing the criterion gives an estimator which is likely to have optimal rate of convergence among the class of penalized estimators.

Theorem 3.1.16. *If the following conditions are fulfilled*

$$2^{j_1} = o(n) \quad 2^{j_0} = o(n)$$

then for n large we have

$$E\mathrm{GFCV}(\lambda_n^2) \approx \mathrm{MSE}(\lambda_n^2) + \sigma^2 \tag{3.1.21}$$

Proof. The proof of this result implies some calculations which have been postponed to the appendix □

Remark 3.1.17. *S. van de Geer in [vdG01] consider penalized estimators with general norms. Define for two functions θ_1 and θ_2 of Θ a quasi-distance between them by*

$$\tau^2(\theta_1|\theta_2) = ||\theta_1 - \theta_2||^2 + \lambda_n^2||\theta_1||_\Theta^2 \tag{3.1.22}$$

With a theorem similar to the theorem in the first part, it is proven that, if there exists a control over the entropy of the class $\{\theta \in \Theta,\ \tau^2(\theta|\theta_0) \leq \delta^2\}$ then the estimator minimizing this pseudo distance where θ_0 is replaced by the observations Y_i, $i = 1, \dots, n$ convergences for the pseudo distance $\tau^2(.|\theta_0)$. The problem of the choice of the smoothing parameter is yet open. That is why we modify the last property and in order to find the optimum parameter for the penalized norm. As a matter of fact, we look for

$$\tilde{\lambda}_n = Arg \inf_{\lambda} \left(||\theta_0 - \hat{\theta}_n||_n^2 + \lambda_n^2 ||\hat{\theta}_n||_{B_{22}^s} \right).$$

We can prove that under the same assumptions the optimum parameter exists. This proof can be found in the appendix. A slight modification of cross validation criterion would also lead to a data dependent optimal choice of the smoothing parameter.

Corollary 3.1.18. *Under the conditions (I) there exists $\tilde{\lambda}_n$ such that:*

$$\tilde{\lambda}_n = arg \min_{\lambda} (E||\theta_0 - \hat{\theta}_n||_n^2 + \lambda^2 ||\hat{\theta}_n||_{B_{22}^s}^2).$$

Convergence in Besov norm

We begin by studying the relations between the true wavelet coefficient β_{jk} and the empirical wavelet coefficient $\bar{\beta}_{jk} = \frac{1}{n}\sum_{i=1}^n \theta(z_i)\psi_{jk}(z_i)$.

Lemma 3.1.19. *We know by the integrability of $\theta\psi_{jk}$ that*

$$\bar{\beta}_{jk} - \beta_{jk} = o(1).$$

But we can be more precise in our inequality:

$$|\bar{\beta}_{jk} - \beta_{jk}| \leq ||\theta||_{B_{\infty\infty}^s} n^{-s} 2^{-\frac{j}{2}} ||\psi||_1 + ||\theta||_\infty Var(\psi) \frac{2^{\frac{j}{n}}}{n}.$$

With this lemma, we can prove consistency of the penalized estimator in B^σ_{22}, $\forall \sigma < s$ as soon as the smoothing parameter tends to zero.

Theorem 3.1.20. *Under the conditions (I), and if we consider $\sigma < s$ such that moreover*

$$2^{j_1} = o(n^{\frac{s}{\sigma} \wedge \frac{1}{1+2\sigma}})$$

then a necessary and sufficient condition to have the consistency of the penalized estimator in norm $\|.\|_{B^\sigma_{22}}$ is given by $\lambda_n^2 \to 0$.

$$E||\hat{\theta}_n - \theta_0||^2_{B^\sigma_{22}} \to 0 \Longleftrightarrow \lambda_n \to 0.$$

Moreover if $\lambda_n^2 = n^{-\frac{2s}{2s+1}}$ then there exists a positive constant $C < \infty$ such that:

$$E||\hat{\theta}_n - \theta_0||^2_{B^\sigma_{22}} \leq Cn^{-\frac{2(s-\sigma)}{2s+1}} \tag{3.1.23}$$

A similar method is used for estimation in nonparametric mixed effect models by Angelini, De Canditis and Leblanc in [ADCL00] or Huang and Lu in [HL00] or [HL01].

The main drawback of such methods is that they require the knowledge of the functional space where the true function θ_0 lies. We could be tempted to try to estimate the regularity of the space where the function lies. For Sobolev spaces H^s or Besov spaces B^s_{pq}, define $\hat{s}_n$ such an estimator. Then we could use for penalty the pseudo-norm $\|.\|_{H^{\hat{s}_n}}$ or $\|.\|_{B^{\hat{s}_n}_{pq}}$. A natural idea is to consider

$$\hat{s}_n = \arg\min_t \min_{H^t} \left(\|Y - \theta\|^2_n + n^{-\frac{2t}{2t+1}} I_t(\theta) \right) \tag{3.1.24}$$

where $I_t(\theta)$ is Sobolev (or Besov) pseudo-norm. In the case of Besov spaces, the minimum is taken over the spaces B^t_{pq}. Though this method does not help constructing an adaptive approach since, on the one hand, the estimator $\hat{s}_n$ is uncomputable, and on the other hand, to prove the convergence of the estimator, an extra penalty term over the indexes t should be added. Nevertheless this method enlights that estimating the smoothness parameter can be the starting point for an adaptive penalized estimation. That is the reason why it becomes natural to adopt a Bayesian point of view and consider this key parameter as an hyper parameter of the model. If we choose a good prior on it, we hope that the posterior law will concentrate on models with the good smoothness parameter. Such method will be investigated in the next chapter.

3.1.4 Appendix

Proof of lemma 3.1.7

Proof.

$$|\delta_h(t+\epsilon,L)-t| = \begin{cases} |\epsilon| & ,|t+\epsilon|>L, \\ |t| & ,|t+\epsilon|\leq L. \end{cases}$$

- If $|t|>2L$:
  then if
  $$|\delta_h(t+\epsilon,L)-t| = |\epsilon| \leq 2|\epsilon|,$$
  else we have both $|t+\epsilon|\leq L$, and $|t|>2L$, so
  $$L\leq |\epsilon|,$$
  and
  $$\begin{aligned} |t| &= |t+\epsilon-\epsilon| \\ &\leq 2|\epsilon|. \end{aligned}$$

- If $|t|\leq 2L$:
  then
  - If $|t+\epsilon|\leq L$, then
    $$|\delta_h(t+\epsilon,L)-t| = |t| \leq \max(|t|,|\delta_h(\epsilon,\frac{L}{2})).$$
  - If $|t+\epsilon|>L$, then
    $$|\delta_h(t+\epsilon,L)-t| = |\epsilon|.$$
    In that case if $|\epsilon|>\frac{L}{2}$, we have $\delta_h(\epsilon,\frac{L}{2})=\epsilon$, so
    $$|\delta_h(t+\epsilon,L)-t| \leq |\delta_h(\epsilon,\frac{L}{2})|.$$
    On the other hand, if $|\epsilon|\leq\frac{L}{2}$, combined with $|t+\epsilon|>L$ we obtain
    $$|t|>\frac{L}{2}\geq|\epsilon|.$$
    so
    $$|\delta_h(t+\epsilon,L)-t| \leq |\delta_h(\epsilon,\frac{L}{2})|.$$

□

Proof.

$$|\delta_s(t+\epsilon,L)-t| = \begin{cases} |\epsilon - L| & , L-t \leq \epsilon, \\ |t| & , -L-t \leq \epsilon \leq L-t, \\ |\epsilon + L| & , \epsilon \leq -L-t. \end{cases}$$

The proof is analogous to the one of the first lemma looking at the inequality in the three main cases

$$t > L \quad , |t| \leq L \quad , t < -L.$$

□

Proof of lemma 3.1.8

Proof. By Hölders inequality, for a sequence $a_1, \dots, a_L$, and for $t \geq 1$,

$$\sum_{l=1}^{L} |a_l| \leq L^{\frac{t-1}{t}} \left(\sum_{l=1}^{L} |a_l|^t \right)^{\frac{1}{t}} \tag{3.1.25}$$

Apply this first with $L = J$, $|a_j| = \sum_{k=0}^{2^j-1} |\alpha_{j,k}|^{\frac{mp}{m+r}}$, and $t = m+r$. Then we find

$$\sum_{j=1}^{J} \left\{ \sum_{k=0}^{2^j-1} |\alpha_{j,k}|^{\frac{mp}{m+r}} \right\} \leq J^{\frac{m+r-1}{m+r}} \left(\sum_{j=1}^{J} \left\{ \sum_{k=0}^{2^j-1} |\alpha_{j,k}|^{\frac{mp}{m+r}} \right\}^{m+r} \right)^{\frac{1}{m+r}}.$$

Next, apply (3.1.25) with $L = 2^j$, $|a_{j,k}| = |\alpha_{j,k}|^{\frac{mp}{m+r}}$, and $t = (m+r)/m$. This yields

$$\left\{ \sum_{k=0}^{2^j-1} |\alpha_{j,k}|^{\frac{mp}{m+r}} \right\} \leq \left\{ 2^{\frac{jr}{m+r}} (\sum_{k=0}^{2^j-1} |\alpha_{j,k}|^p)^{\frac{m}{m+r}} \right\}.$$

Do this for each $j = 1, \dots J$, and insert the result in :

$$\begin{aligned} & J^{\frac{m+r-1}{m+r}} \left(\sum_{j=1}^{J} \left\{ \sum_{k=0}^{2^j-1} |\alpha_{j,k}|^{\frac{mp}{m+r}} \right\}^{m+r} \right)^{\frac{1}{m+r}} \\ & \leq J^{\frac{m+r-1}{m+r}} \left(\sum_{j=1}^{J} \left\{ 2^{\frac{jr}{m+r}} (\sum_{k=0}^{2^j-1} |\alpha_{j,k}|^p)^{\frac{m}{m+r}} \right\}^{m+r} \right)^{\frac{1}{m+r}} \\ & = J^{\frac{m+r-1}{m+r}} \left(\sum_{j=1}^{J} 2^{jr} \left\{ \sum_{k=0}^{2^j-1} |\alpha_{j,k}|^p \right\}^m \right)^{\frac{1}{m+r}} \end{aligned}$$

□

Proof of lemma 3.1.10

Proof. Condition 1 implies that

$$\sum_{k=1}^{2^j} |\alpha_{j,k}|^p \leq 2^{-j\frac{r}{m}},$$

and hence

$$|\alpha_{j,k}| \leq 2^{-j\frac{r}{mp}}.$$

So $|\alpha_{j,k}| \leq \epsilon$ for all $j \geq \frac{mp}{r}\log_2 \frac{1}{\epsilon} = J_\epsilon$ (say). Complete the proof by taking J equal to the largest integer less than J_ϵ in the previous proof. □

Proof of lemma 3.1.11

Proof. This follows from the application of Hölder's inequality (twice). Let M be the smallest integer such that $2^M \geq \epsilon^{-\frac{mp}{m+r}}$. We have

$$\sum_{j=1}^{J} \sum_{|\alpha_{j,k}|\leq\epsilon} |\alpha_{j,k}|^2 \leq \epsilon \sum_{j=1}^{M} \sum_{k=0}^{2^j-1} 1 + \sum_{j>M} \sum_{k=0}^{2^j-1} |\alpha_{j,k}|^2,$$

where we assume $M < J$ (if not, the second term in the right hand side of the inequality vanishes, and the result follows immediately). But

$$\begin{aligned}
\sum_{j>M} \sum_{k=0}^{2^j-1} |\alpha_{j,k}|^2 &= \sum_{j>M} \left(2^{-j(\frac{m+r-mp/2}{mp/2})}\right) \left(2^{j(\frac{m+r-mp/2}{mp/2})} \sum_{k=0}^{2^j-1} |\alpha_{j,k}|^2\right) \\
&\leq \left(\sum_{j>M} 2^{-j(\frac{m+r-mp/2}{mp/2})(\frac{mp/2}{mp/2-1})}\right)^{\frac{mp/2-1}{mp/2}} \left(\sum_{j=1}^{J} 2^{j(\frac{m+r-mp/2}{mp/2})mp} (\sum_{k=0}^{2^j-1} |\alpha_{j,k}|^2)^{mp/2}\right)^{\frac{2}{mp}} \\
&\leq \left(\sum_{j>L} 2^{-j(\frac{m+r-mp/2}{mp/2-1})}\right)^{\frac{mp/2-1}{mp/2}} \left(\sum_{j=1}^{J} 2^{j(m+r-mp/2)} (2^{j(\frac{p/2-1}{p/2})} \{\sum_{k=0}^{2^j-1} |\alpha_{j,k}|^{p/2}\}^{\frac{2}{p}})^{mp/2}\right)^{\frac{2}{mp}} \\
&\leq C_0 2^{-M(\frac{m+r-mp/2}{mp/2})} \left(\sum_{j=1}^{J} 2^{jr} \sum_{k=0}^{2^j-1} |\alpha_{j,k}|^p\right)^{\frac{2}{mp}},
\end{aligned}$$

for some constant c_0. So we arrive at

$$\begin{aligned}
\sum_{j=1}^{J} \sum_{|\alpha_{j,k}|\leq\epsilon} |\alpha_{j,k}|^2 &\leq \epsilon 2^{M+2} + C_0 2^{-M(\frac{2(m+r)-mp}{mp})} \\
&\leq c\epsilon^{2-\frac{mp}{m+r}},
\end{aligned}$$

for some constant c. □

Proof of theorem 3.1.12

Proof. 1. The variance term :

$$
\begin{aligned}
&\sum_{j=1}^{n} \text{Variance}(\tilde{\alpha}_j) \\
&= \sum_{j=1}^{n} \frac{1}{n(1+\lambda^2 w_j^2)^2}\sigma^2 \\
&= \frac{\sigma^2}{n}\left(\sum_{j=1}^{n} \frac{1}{(1+\lambda^2 j^{2s})^2}\right) \\
&= \frac{\sigma^2}{n}(\sum_{j, j\leq (1/\lambda)^{1/s}} \frac{1}{(1+\lambda^2 j^{2s})^2} + \sum_{j, j > (1/\lambda)^{1/s}} \frac{1}{(1+\lambda^2 j^{2s})^2}) \\
&\leq \frac{\sigma^2}{n} \sum_{j, j\leq (1/\lambda)^{1/s}} 1 + \frac{\sigma^2}{n} \sum_{j, j > (1/\lambda)^{1/s}} \frac{1}{(1+\lambda^2 j^{2s})^2} \\
&\leq \frac{\sigma^2}{n}(1/\lambda)^{1/s} + \frac{\sigma^2}{n} \sum_{j, j > (1/\lambda)^{1/s}} \frac{1}{\lambda^4 j^{4s}}) \\
&\leq \frac{\sigma^2}{n}(1/\lambda)^{1/s} + \frac{\sigma^2}{n\lambda^4} \int_{x > (1/\lambda)^{1/s}} \frac{1}{x^{4s}}\, dx \\
&= \frac{\sigma^2}{n}(1/\lambda)^{1/s} + \frac{\sigma^2}{n\lambda^4} \int_{(1/\lambda)^{1/s}}^{n} \frac{1}{x^{4s}}\, dx \\
&= \frac{\sigma^2}{n}(1/\lambda)^{1/s} + \frac{\sigma^2}{n\lambda^4} \left[\frac{x^{-(4s-1)}}{-(4s-1)}\right]_{(1/\lambda)^{1/s}}^{n} \\
&\approx \frac{\sigma^2}{n}(1/\lambda)^{1/s} \text{ as soon as } 4s-1>0.
\end{aligned}
$$

Moreover :

$$
\begin{aligned}
\sum_{j=1}^{n} \text{Variance}(\tilde{\alpha}_j) &\geq \frac{\sigma^2}{n} \sum_{j, j > (1/\lambda)^{1/s}} \frac{1}{(1+\lambda^2 j^{2s})^2} \\
&\geq \frac{\sigma^2}{4n} \sum_{j, j > (1/\lambda)^{1/s}} \frac{1}{\lambda^4 j^2 s} \\
&\approx \frac{\sigma^2}{n}(1/\lambda)^{1/s}
\end{aligned}
$$

So the variance term is of order :

$$\frac{\sigma^2}{n}(1/\lambda)^{1/s}$$

2. The Bias term :
   We have to calculate it :

$$\begin{aligned}\sum_{j=1}^{n}(E(\tilde{\alpha_j})-\alpha_j)^2 &= \sum_{j=1}^{n}\frac{\lambda^4 w_j^4}{(1+\lambda^2 w_j^2)^2}\alpha_j^2\\ &= \sum_{j=1}^{n}\frac{\lambda^4 j^{4s}}{(1+\lambda^2 j^{2s})^2}\alpha_j^2\end{aligned}$$

$$\begin{aligned}\sum_{j=1}^{n}(E(\tilde{\alpha_j})-\alpha_j)^2 &\geq \sum_{j>(1/\lambda)^{1/s}}\frac{\lambda^4 j^{2s-1-2\epsilon}}{(1+\lambda^2 j^{2s})^2}\\ &\geq \sum_{j>(1/\lambda)^{1/s}} 1/4\lambda^4 j^{2s-1-2\epsilon}\end{aligned}$$

   which is a term of order λ^2.

   Since we have considered a normalized function, i.e if $||\theta||^2_{H_s}=\sum_{i=1}^n w_i\alpha_i^2\leq 1$, then

$$\begin{aligned}\sum_{j=1}^{n}(E(\tilde{\alpha_j})-\alpha_j)^2 &\leq \sum_{j=1}^{n}\frac{\lambda^4 w_j^4}{1+\lambda^2 w_j^2}\alpha_j^2\\ &\leq \lambda^2\sum_{j=1}^{n}\frac{\lambda^2 w_j^2}{1+\lambda^2 w_j^2}w_j^2\alpha_j^2\\ &\leq \lambda^2\sum_{j=1}^{n}w_j^2\alpha_j^2\\ &\leq \lambda^2.\end{aligned}$$

   So we have the same the same order for the bias term when $\mathcal{F}=\{\theta\in H_s, ||\theta||_{H_s}\leq 1\}$. As a result, the error due to the bias term is of order λ^2.

3. Rate of convergence
   The error in the two terms (bias and variance) must be of the same order, which gives :

$$\lambda^2 \approx \frac{1}{n}\left(\frac{1}{\lambda}\right)^{1/s}$$

$$\lambda \approx n^{-s/(2s+1)}$$

   So the rate of convergence for the estimator is $r_n = n^{-\frac{2s}{2s+1}}$.

□

Proof of lemma 3.1.14

Proof. The norm is the quadratic norm $:||\alpha||^2 = \sum_{j\geq 1} \alpha_j^2$. Since by definition of Θ the sum $\sum_j w_j^2 \alpha_j^2$ converges, there exists N such that

$$\begin{aligned} 1 &\geq \sum_{j\geq N} w_j^2 \alpha_j^2 \\ &\geq w_N^2 \sum_{j \geq N} \alpha_j^2. \end{aligned}$$

Then for $w_j = j^s$, and for $N = \delta^{-1/s}$, we have the following inequality

$$\sum_{j\geq N} \alpha_j^2 \leq \delta^2.$$

So the δ-entropy for the norm specified of the set $\mathcal{F}$ is the δ-entropy of a ball of radius R in a N dimensional space. This entropy is well-known, see for instance van de Geer [vdGW96] :

$$H(\delta, B_R(\mathbb{R}^N)) \leq C.N \log(\frac{R}{\delta}).$$

So for $N = \delta^{-1/s}$, we have

$$H(\delta, \Theta, ||.||) \leq C.\delta^{-\frac{1}{s}} \log\left(\frac{R}{\delta}\right).$$

□

Proof of theorem 3.1.15

Proof. The risk function $E||\hat{\theta}_n - \theta||_n^2$ is here taken as a function of the variable λ_n^2, called $\Phi(\lambda_n^2)$. We want to show that under some conditions this function has a minimum since the function is continuously increasing for large values of the variable.

$$\begin{aligned} \frac{d}{d\lambda_n^2}\Phi(\lambda_n^2) &= \frac{d}{d\lambda_n^2}\left(\frac{1}{n}\sum_{i=1}^n |\tilde{f}_n(z_i) - \theta(z_i)|^2\right) \\ &= \frac{2}{n}\sum_{i=1}^n (\tilde{f}_n(z_i) - \theta(z_i)) \frac{d}{d\lambda_n^2}\left(\tilde{f}_n(z_i)\right) \\ \frac{d}{d\lambda_n^2}\tilde{f}_n(z_i) &= \sum_{j_0}^{j_1}\sum_k \frac{d}{d\lambda_n^2}\left(\tilde{\beta_{jk}}\right)\psi_{jk}(z_i) \\ &= \sum_{j_0}^{j_1}\sum_k \frac{-2^{2js}}{(1+\lambda_n^2 2^{2js})^2}\hat{\beta}_{jk}\psi_{jk}(z_i). \end{aligned}$$

So we have

$$\frac{d}{d\lambda_n^2}E||\hat{\theta}_n - \theta_0||_n^2 = E\left(\frac{2}{n}\sum_{j=j_0}^{j_1}\sum_k \frac{-2^{2js}}{(1+\lambda_n^2 2^{2js})^2}\hat{\beta}_{jk}\sum_{p=j_0}^{j_1}\sum_q(\tilde{\beta}_{pq}-\beta_{pq}) * \sum_{i=1}^n \psi_{jk}(z_i)\psi_{pq}(z_i)\right)$$

$$= E\left(\frac{2}{n}\sum_{j=j_0}^{j_1}\sum_k \frac{-2^{2js}}{(1+\lambda_n^2 2^{2js})^3}\hat{\beta}_{jk}((1+\lambda_n^2 2^{2js})\beta_{jk}\bar{\beta}_{jk} - E(\hat{\beta}_{jk}^2))\right).$$

But

$$\bar{\beta}_{jk} = \frac{1}{n}\sum_{i=1}^n \theta_0(z_i)\psi_{jk}(z_i) = E\hat{\beta}_{jk}.$$

It is the discrete wavelet coefficient. Since the function $\theta_0\psi_{jk}$ is Riemann integrable, and since we have chosen $z_i = \frac{i}{n}$, the empirical coefficient converges to the true wavelet coefficient:

$$\bar{\beta}_{jk} - \beta_{jk} = o(1).$$

For the same reason, we have the following convergence:

$$\frac{1}{n}\sum_{i=1}^n \psi_{jk}(z_i)^2 - 1 = o(1).$$

Moreover:

$$E\hat{\beta}_{jk}^2 = \bar{\beta}_{jk}^2 + \frac{\sigma^2}{n}\frac{1}{n}\sum_{i=1}^n \psi_{jk}^2(z_i).$$

So if we make the first approximation we get for n large enough:

$$\frac{d}{d\lambda_n^2}E||\hat{\theta}_n - \theta_0||_n^2 \approx \frac{2}{n}\sum_{j=j_0}^{j_1}\sum_k \frac{-2^{2js}}{(1+\lambda_n^2 2^{2js})^3}((1+\lambda_n^2 2^{2js})\beta_{jk}\bar{\beta}_{jk} - \bar{\beta}_{jk}^2 - \frac{\sigma^2}{n}\frac{1}{n}\sum_{i=1}^n \psi_{jk}^2(z_i))$$

$$= \frac{2}{n}\sum_{j=j_0}^{j_1}\sum_k \frac{-2^{2js}}{(1+\lambda_n^2 2^{2js})^3}(\lambda_n^2 2^{2js}\bar{\beta}_{jk} - \frac{\sigma^2}{n}).$$

Then we need to find a lower bound for this quantity by a strictly positive constant to prove the strict growth of the function for large values of λ_n^2. So, assume we have proven this result, there exists a constant c such that, for all $|\lambda_n^2| \geq c$, $\frac{d}{d\lambda_n^2}E||\tilde{\theta}_{\lambda_n} - \theta_0|| > 0$. As a consequence, by continuity, the minimum of the function lies and exists in the compact set $[-c, c]$. So it suffices to prove the first statement.
First of all we remark than if we define the number of non zero coefficients in the wavelet decomposition

$$N_j = \sharp\{|\beta_{jk}| > 0\}$$

then there exists a constant c depending of the function θ and of the wavelet ψ

$$N_j \leq 2^j.$$

Moreover we suppose that there exists $k^{'}$ and a constant $c_0 > 0$ such that $|\beta_{j_0k'}| \geq c_0 2^{-j_0 s}$, hence we obtain the following upper bound:

$$|\bar{\beta}_{j_0k'}| \geq |\beta_{j_0k'}| - |\bar{\beta}_{j_0k'} - \beta_{j_0k'}|$$

But we have

$$|\bar{\beta}_{j_0k'} - \beta_{j_0k'}| \leq ||\theta||_{B^s_{22}} 2^{-j/2} n^{-s} ||\psi||_1 + \frac{2^{j/2}||\theta||_\infty}{n}\mathrm{Var}(\psi),$$

where $\mathrm{Var}(\theta)$ is the variation of a function θ ,i.e

$$\mathrm{Var}(\theta) = \sup_{(x_i)\subset[0,1]^n} |\theta(x_i) - \theta(x_{i-1})|.$$

Then we obtain

$$|\bar{\beta}_{j_0k'}| \geq \frac{c_0}{2} 2^{-j_0 s}.$$

So,

$$\begin{aligned}
\frac{1}{2}\frac{d}{d\lambda_n^2} E||\hat{\theta}_n - \theta_0||_n^2 &\geq |\sum_{j=j_0}^{j_1} \frac{2^{2js}}{(1+\lambda_n^2 2^{2js})^3} \sum_k (\lambda_n^2 2^{2js} \bar{\beta}_{jk}^2 - \frac{\sigma^2}{n})| \\
&> C_1 \frac{2^{4j_0 s}\lambda_n^2 2^{-2j_0 s}}{(1+\lambda_n^2 2^{2j_0 s})^3} - \sum_{j=j_0}^{j_1} \frac{\sigma^2}{n(\lambda_n^2)^3 2^{4js}} N_j \\
&> \frac{1}{\lambda_n^4 2^{4j_0 s}} (-C_3 \frac{2^{j_1}}{\lambda_n^2 n} + C_1 \frac{1}{(1+\lambda_n^{-2} 2^{-2j_0 s})^3}) \\
&> \frac{1}{\lambda_n^4 2^{4j_0 s}} (C \frac{1}{(1-o(1))^2} - \frac{1}{l} o(1)) \\
&> 0 \quad \text{for } n \to +\infty.
\end{aligned}$$

So the statement is proven. □

Proof of theorem 3.1.16

Proof.

$$\begin{aligned}
E\mathrm{GFCV}(\lambda_n^2) &= (1+H(\lambda_n^2))^2 E\frac{1}{n}\sum_{i=1}^n |Y_i - \tilde{f}^{(i)}_{\lambda_n^2}|^2 \\
&= (1+H(\lambda_n^2))^2 \left(\mathrm{MSE}(\lambda_n^2) + \frac{1}{n}\sum_{j=1}^n E|W_j|^2 + E\frac{2}{n}\sum_{j=1}^n W_j(\theta_0(z_j) - \tilde{\theta}_{\lambda_n^2}(z_j)) \right) \\
&= (1+H(\lambda_n^2))^2 \left(E||\hat{\theta}_n - \theta_0||^2 + \sigma^2 + E\frac{2}{n}\sum_{j=1}^n W_j(\theta_0(z_j) - \tilde{\theta}_{\lambda_n^2}(z_j)) \right)
\end{aligned}$$

But we have as soon as $2^{j_0} = o(n)$ and $2^{j_1} = o(n)$

$$H(\lambda_n^2) \to 0 \tag{3.1.26}$$

Indeed by easy calculations we can write:

$$\begin{aligned} H(\lambda_n^2) &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n}\sum_k \phi_k^2(z_i) + \sum_{j_0}^{j_1}\sum_k \frac{1}{1+\lambda_n^2 2^{2js}}\psi_{jk}^2(z_i)\right) \\ &\leq \frac{2^{j_0}}{n} + \frac{2^{j_1}}{n} \end{aligned}$$

which goes to zero with the appropriate assumptions. Moreover, we can see that

$$\begin{aligned} & E\frac{2}{n}\sum_{j=1}^{n} W_j(\theta_0(z_j) - \tilde{\theta}_{\lambda_n^2}(z_j)) \\ &= -\frac{2}{n}E\sum_{j=1}^{n} W_j\tilde{\theta}_{\lambda_n^2}(z_j) \\ &= -\frac{2}{n}\sum_{i=1}^{n}\left(\sum_k \phi_k^2(z_i)E(Y_iW_i) + \sum_{j=j_0}^{j_1}\sum_k \frac{1}{1+\lambda_n^2 2^{2js}}\psi_{jk}^2(z_i)E(Y_iW_i)\right) \\ &= 2\sigma^2 H(\lambda_n^2) \end{aligned}$$

which goes to zero as n increases. This two inequalities show that for n large, Cross validation is asymptotically equivalent to mean square error. □

Proof of corollary 3.1.18

Proof. If we use the wavelet decomposition the penalty can be described as a function of the coefficients as usual

$$\text{Pen}(\theta_0) = \sum_j\sum_k 2^{2js}|\beta_{jk}|^2.$$

So for the penalized estimator the penalty is:

$$\text{Pen}(\hat{\theta}_n) = \sum_j 2^{2js}\sum_k \frac{1}{(1+\lambda_n^2 2^{2js})^2}\hat{\beta}_{jk}^2.$$

$$\frac{d}{d\lambda_n^2}\text{Pen}(\hat{\theta}_n) = -2\sum_j 2^{4js}\sum_k \frac{1}{(1+\lambda_n^2 2^{2js})^2}\hat{\beta}_{jk}^2.$$

$$\begin{aligned} \frac{d}{d\lambda_n^2}\left(\lambda_n^2\text{Pen}(\hat{\theta}_n)\right) &= \sum_j 2^{2js}\sum_k \frac{1}{(1+\lambda_n^2 2^{2js})^2}\hat{\beta}_{jk}^2 - 2\sum_j 2^{4js}\sum_k \frac{1}{(1+\lambda_n^2 2^{2js})^2}\hat{\beta}_{jk}^2 \\ &= \sum_{j=j_0}^{j_1} 2^{2js}\frac{1-\lambda_n^2 2^{2js}}{(1+\lambda_n^2 2^{2js})^3}\sum_k |\hat{\beta}_{jk}|^2. \end{aligned}$$

$$
\begin{aligned}
E\frac{d}{d\lambda_n^2}\left(\lambda_n^2 \mathrm{Pen}(\hat{\theta}_n)\right) &= \sum_{j=j_0}^{j_1} 2^{2js}\frac{1-\lambda_n^2 2^{2js}}{(1+\lambda_n^2 2^{2js})^3}\sum_k(\beta_{jk}^2+\frac{\sigma^2}{n}) \\
&= \sum_{j=j_0}^{j_1} 2^{2js}\frac{1-\lambda_n^2 2^{2js}}{(1+\lambda_n^2 2^{2js})^3}(\sum_k(\beta_{jk}^2)+\frac{\sigma^2 2^j}{n}) \\
&\geq \sum_{j=j_0}^{j_1} 2^{2js}\frac{1-\lambda_n^2 2^{2js}}{(1+\lambda_n^2 2^{2js})^3}\left(\frac{\sigma^2 2^j}{n}+C_0^2 2^{j_0}2^{-2j_0 s}\right).
\end{aligned}
$$

So if we go on using the same previous proof we can prove that for n large enough:

$$
\frac{d}{d\lambda_n^2}E\left(E||\hat{\theta}_n-\theta_0||_n^2+\lambda_n^2\mathrm{Pen}(\hat{\theta}_n)\right)>0
$$

and with the same compacity argument the corollary is proved □

Proof of lemma 3.1.19

Proof.

$$
\begin{aligned}
|\bar{\beta}_{jk}-\beta_{jk}| &= |\int\theta_0(x)\psi_{jk}(x)\,dx-\frac{1}{n}\sum_{i=1}^n\theta_0(z_i)\psi_{jk}(z_i)| \\
&= |\sum_{i=1}^n\int_{z_{i-1}}^{z_i}\left(\theta_0(x)\psi_{jk}(x)-\theta_0(z_i)\psi_{jk}(z_i)\right)\,dx| \\
&\leq |\sum_{i=1}^n\int_{z_{i-1}}^{z_i}(\theta_0(x)-\theta_0(z_i))\psi_{jk}(x)\,dx|+|\sum_{i=1}^n\int_{z_{i-1}}^{z_i}\theta_0(z_i)(\psi_{jk}(x)-\psi_{jk}(z_i))\,dx| \\
&\leq \sum_{i=1}^n\int_{z_{i-1}}^{z_i}||\theta_0||_{B^s_{\infty\infty}}|x-z_i|^s|\psi_{jk}(x)|\,dx+ \\
&|\sum_{i=1}^n\int_{z_{i-1}}^{z_i}\theta_0(z_i)(\psi_{jk}(x)-\psi_{jk}(z_i))\,dx| \\
&\leq ||\theta_0||_{B^s_{\infty\infty}}n^{-s}2^{-\frac{j}{2}}||\psi||_{L_1}+\sum_{i=1}^n\int_{z_{i-1}}^{z_i}||\theta_0||_\infty|\psi_{jk}(x)-\psi_{jk}(z_i)|\,dx.
\end{aligned}
$$

But by the second mean theorem there exists $t_i\in(z_{i-1},z_i)$ such that

$$
\int_{z_{i-1}}^{z_i}|\psi_{jk}(x)-\psi_{jk}(z_i)|\,dx=(z_i-z_{i-1})(\psi_{jk}(t_i)-\psi_{jk}(z_{i-1})).
$$

$$
\begin{aligned}
|\bar{\beta}_{jk}-\beta_{jk}| &\leq ||\theta||_{B^s_{\infty\infty}}n^{-s}2^{-\frac{j}{2}}||\psi||_{L_1}+||\theta||_\infty 2^{\frac{j}{2}}\frac{1}{n}\sum_{i=1}^n|\psi(2^jt_i-k)-\psi(2^jz_{i-1}-k)| \\
&\leq ||\theta||_{B^s_{\infty\infty}}n^{-s}2^{-\frac{j}{2}}||\psi||_{L_1}+||\theta_0||_\infty 2^{\frac{j}{2}}\frac{1}{n}\mathrm{Var}(\psi).
\end{aligned}
$$

□

Proof of theorem 3.1.20

Proof. 1. Necessary Condition:

If the penalized estimator is consistent, then it implies by the property of the wavelet basis that

$$E(\tilde{\beta}_{jk} - \beta_{jk})^2 \to 0.$$

But direct calculations give

$$E(\tilde{\beta}_{jk} - \beta_{jk})^2 = \frac{1}{(1+\lambda_n^2 2^{2js})^2}(\bar{\beta}_{jk} - \beta_{jk})^2 + \frac{\sigma^2}{n}\frac{1}{n}\sum_{i=1}^{n}\psi_{jk}^2(z_i) - \\ 2\lambda_n^2 2^{2js}\beta_{jk}(\bar{\beta}_{jk} - \beta_{jk}) + \lambda_n^4 2^{4js}\beta_{jk}^2$$

Now if we use the upper bound of the lemma we obtain:

$$0 \leq E(\tilde{\beta}_{jk} - \beta_{jk})^2 = (o(1) + \lambda_n^4 2^{4js}\beta_{jk}^2).$$

So we must have

$$\lambda_n^2 \to 0.$$

2. Sufficient Condition:

We must study the term

$$E||\hat{\theta}_n - \theta_0||^2_{B^\sigma_{22}}$$

and we use the bias-variance decomposition :

$$\begin{aligned} E||\hat{\theta}_n - \theta_0||^2_{B^\sigma_{22}} &= E||\hat{\theta}_n - E\hat{\theta}_n||^2_{B^\sigma_{22}} + ||E\hat{\theta}_n - \theta_0||^2_{B^\sigma_{22}} \\ &\leq \frac{\sigma^2}{n}\sum_{j=j_0}^{j_1}\sum_k \frac{2^{2j\sigma}}{(1+\lambda_n^2 2^{2js})^2}(\frac{1}{n}\sum_{i=1}^{n}\psi_{jk}^2(z_i) + \frac{2n}{\sigma^2}(\bar{\beta}_{jk} - \beta_{jk})^2 + \\ & 2\sum_{j=j_0}^{j_1}\sum_k \frac{2^{2j\sigma}}{(1+\lambda_n^2 2^{2js})^2}\lambda_n^4 2^{4j\sigma}\beta_{jk}^2 + \sum_{j>j_1} 2^{2j\sigma}\sum_k \beta_{jk}^2 + (P_{j_0})). \end{aligned}$$

Where the term P_{j_0} denotes the difference between the function and the projected function onto the space V_{j_0}.

The first term to be bounded is the part of the Besov norm of the true function which does not belong to the projected space:

$$\begin{aligned} \sum_{j>j_1} 2^{2j\sigma}\sum_k \beta_{jk}^2 &= \sum_{j>j_1} 2^{2j\sigma - s + s}\sum_k \beta_{jk}^2 \\ &\leq 2^{(j_1+1)(\sigma - s)}\sum_j 2^{2js}\sum_k \beta_{jk}^2 \\ &\leq 2^{(j_1+1)(\sigma-s)}||\theta_0||^2_{B^s_{22}}. \end{aligned}$$

For the second main term we firstly note that we have the following inequality:

$$\frac{\lambda_n^4 2^{4js}}{(1+\lambda_n^2 2^{2js})^2} \le \lambda_n^{2\frac{s-\sigma}{s}} 2^{2j(s-\sigma)}.$$

$$\begin{aligned} 2\sum_{j=j_0}^{j_1}\sum_k \frac{2^{2j\sigma}}{(1+\lambda_n^2 2^{2js})^2}\lambda_n^4 2^{4j\sigma}\beta_{jk}^2 &\le 2\sum_{j=j_0}^{j_1}\sum_k 2^{2j\sigma}2^{2j(s-\sigma)}\lambda_n^{2\frac{s-\sigma}{s}}\beta_{jk}^2 \\ &\le 2\lambda_n^{2\frac{s-\sigma}{s}}||\theta_0||^2_{B^s_{22}}. \end{aligned}$$

So, using again the upper bound of the lemma, there exists a constant C non infinite such that

$$E||\hat{\theta}_n-\theta_0||^2_{B^\sigma_{22}} \le C\left(\lambda_n^{2\frac{s-\sigma}{s}} + 2^{2j_1(\sigma-s)} + \frac{2^{j_0}}{n} + \frac{2^{j_1(1+2\sigma)}}{n} + \frac{2^{2j_1\sigma}}{n^{2s}}\right).$$

This quantity becomes an $o(1)$ as soon as we have

$$\begin{cases} 2^{j_1} = o(n)\, j_1 \to \infty \\ \lambda_n^2 \to 0 \end{cases}$$

So under assumptions (I), the estimator converges as soon as $\lambda_n^2 \to 0$. Moreover, if we choose j_1 to minimize the quantity we find $\frac{2^{j_1}}{n} \approx 2^{-2j_1 s}$ condition which leads to the optimal choice:

$$2^{j_1} = n^{\frac{1}{1+2s}}.$$

Up to this point we find that

$$E||\hat{\theta}_n-\theta||^2 = O(\lambda_n^{2(1-\frac{\sigma}{s})}) \qquad (3.1.27)$$

So the choice $\lambda_n^2 = n^{-\frac{2s}{2s+1}}$ leads to the optimal rate of convergence.

□

3.2 Sieves Methods

For sake of completeness in our study, we want to present the method used by L.Birgé and P.Massart in [BM97] or [BM98] to estimate a density. This method introduces the hard-thresholded estimator but in a different way than the one chosen by Donoho and al. in [DJ96b]. Indeed the main idea is the following : instead of choosing a level and thresholding the coefficients up to that level, the authors determine the optimal number of coefficients to be kept in the decomposition of the estimator, and this number corresponds to an adequate thresholding level.

In the context of density estimation, we observe n i.i.d random variables $X_1, \dots, X_n$ with common density θ_0 with respect to Lebesgue measure μ. Consider $(\psi_\lambda)_{\lambda\in\Lambda}$ an orthonormal system in $L^2(\mu)$ with $|\Lambda_n| = N_n$ and $S_n = \mathrm{Vect}\{\psi_\lambda,\ \lambda \in \Lambda\}$. Define $m \in \mathcal{M}_n$ a subset of Λ_n, $D_m = |m|$ and $S_m = \mathrm{Vect}\{\psi_\lambda,\ \lambda \in m\}$. To each m, we associate the penalty $\mathrm{pen}(m) = \frac{L_n D_m}{n}$ where $L_n \geq 1$ is a weight. Set $\gamma_n(\theta) = \frac{1}{n}\sum_{i=1} -2\theta(X_i) + \|\theta\|_2^2$ an empirical contrast function. Then, the penalized estimator, $\tilde{\theta}$, is defined as the solution of the following minimization problem:

$$\begin{cases} \gamma_n(\tilde{\theta}) + \mathrm{pen}(\tilde{m}) = \inf_{m\in M_n}(\inf_{\theta\in S_m}\gamma_n(\theta) + \mathrm{pen}(m)), \\ \tilde{\theta} \in S_{\tilde{m}}. \end{cases} \tag{3.2.1}$$

Theorem 3.2.1. *If we choose for penalty function the function* $\mathrm{pen}(m) = \frac{L_n D_m}{n}$*, the solution of the minimization problem* (3.2.1) *is the set of the* λ*'s such that* $\hat{\beta}_\lambda^2 > L_n/n$*, which corresponds to the hard-threshold wavelet estimator:*

$$\tilde{\theta} = \sum_{\lambda\in\Lambda_n} \hat{\beta}_\lambda 1_{\hat{\beta}_\lambda^2 > L_n/n}\psi_\lambda.$$

Proof. The empirical coefficient $\hat{\beta}_\lambda$ is defined by

$$\hat{\beta}_\lambda = \frac{1}{n}\sum_{i=1}^{n}\psi_\lambda(X_i).$$

Minimizing the contrast over S_m leads to the projection estimator $\hat{\theta} = \sum_{\lambda\in\Lambda_m}\hat{\beta}_\lambda\psi_\lambda$. We point out that $\gamma_n(\hat{\theta}) = -\sum_{\lambda\in\Lambda_m}\hat{\beta}_\lambda^2$. As a result we get:

$$\begin{aligned} &\inf_{m\in\mathcal{M}_n}\left(\inf_{\lambda\in\Lambda_m}\gamma_n(\theta) + \mathrm{pen}(m)\right) \\ &= \inf_{m\in\mathcal{M}_n}\left(\gamma_n(\hat{\theta}) + \mathrm{pen}(m)\right) \\ &= \inf_{m\in\mathcal{M}_n}\left(-\sum_{\lambda\in\Lambda_m}\hat{\beta}_\lambda^2 + \frac{L_n D_m}{n}\right) \\ &= \inf_{m\in\mathcal{M}_n} -\sum_{\lambda\in\Lambda_n}\left(\hat{\beta}_\lambda^2 - \frac{L_n}{n}\right) \end{aligned}$$

which gives the final expression of the estimator. □

In order to study the asymptotic behaviour of such an estimator, we denote by B_0 the space of functions $\theta = \sum_\lambda \beta_\lambda\psi_\lambda$, so that

$$\Sigma_\infty(\theta) = \sum_{j\geq 0} 2^{j/2}\sup_{\lambda\in\Lambda(j)} \mid \beta_\lambda \mid < +\infty.$$

Theorem 3.2.2. *If we assume that θ_0 belongs to some Besov space $B^s_{p\infty}$ and that the regularity of the wavelet verifies $r+1 > s > 1/p$ and of course under the asumption that*

$$| \Sigma_\infty(\theta_n) - \Sigma_\infty(\theta_0) | \leq C/\Phi,$$

then we have the following rate of convergence for any $\mathbb{L}^q$-loss,

$$\mathbb{E} \mid\mid \tilde{\theta} - \theta_0 \mid\mid^q = O\left(\frac{\log(n)}{n}\right)^{qs/(1+2s)}.$$

We have obtained the classical adaptive rate of convergence for the estimation in a Besov set.

The proof relies on an important theorem by Ledoux and Talagrand, which is derived from a concentration inequality [LT91]:

Theorem 3.2.3. *Let X_i, $i \in [1,n]$ independent random variables and let u_i be n Rademacher variables independent of the firsts random variables, and $\{f_t,\ t \in T\}$ a family of functions uniformly bounded by b. Let $v = \sup_{t\in T} Var(f_t(X_1))$., there exists universal constants so that for any positive ζ*

$$\mathbf{P}\left(\sup\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n f_t(X_i) - \mathbb{E}f_t(X_i)\right) \geq \kappa_2 E\left(\sup_{t\in T} \mid \frac{1}{\sqrt{n}}\sum_{i=1}^n u_i f_t(X_i) \mid\right) + \zeta\right)$$

$$\leq \exp\left(-\kappa_1(\frac{\zeta^2}{v} \wedge \frac{\zeta\sqrt{n}}{b})\right).$$

where κ_1 and κ_2 are two constants.

Chapter 4

Bayesian adaptation and penalization

Our aim in this section is first to establish connections between Bayesian estimators, obtained by considering the maximum of the posterior law, and penalized M-estimators, defined as solution of optimization problems. In a second part, we will try to construct an estimator, based on a well chosen Bayesian-type penalized loss function, that will be adaptive in the minimax sense. Then, we will pay a special attention to the case where the functions are decomposed onto wavelet bases. Finally, we will consider the model of super smooth functions and see if previous results can be extended to this case.
The scheme of our study is the Bayesian framework: we observe independent random variables represented by an infinite dimensional random vector $X = (X_1, \dots, X_n, \dots)$ with probability $\mathbf{P}_0 = \frac{dp_0}{d\mu}$ where μ is the Lebesgue measure. The distribution function is indexed by a parameter $\theta = (\theta_i)_{i\geq 1}$ lying in a space Θ, and so we will write $p = p_\theta$ its density. The true parameter θ_0 is the unknown parameter of interest to be estimated. We willstudy the case where θ is a location parameter.

4.1 Bayesian estimation and M-estimation

The idea of Bayesian estimation is the following: we define a prior probability $\pi(\theta)$ on Θ. Given the parameter θ, the data follow a law with density $p_\theta(X) = p(X|\theta)$. Thanks to Bayes rule, the posterior distribution is defined as:

$$p(\theta|X) = \frac{p(X|\theta)\pi(\theta)}{p(X)}$$

where $p(X) = \int p(X|\theta)\pi(\theta)\, d\theta$.
We observe the sub-model where we only consider a finite number of observations, say n, so we observe:

$$\left\{ \begin{array}{l} X_i = \theta_i + W_i, \quad i = 1, \dots, n \\ \theta = (\theta_i)_{i\geq 1} \in \Theta \end{array} \right. \tag{4.1.1}$$

where the random variables W_i stand for a noise with known variance $\frac{\sigma^2}{n}$ and θ_i is the parameter of interest. The number of observations n, growing to infinity will give the asymptotics

of the estimation problem. The law of $W_i, i = 1, \dots, n$ is given by the problem and satisfies some conditions. In Section 4.1, we will take Gaussian errors, or Laplacian errors while in the following sections, we will only look at the first Gaussian case.
Define the Maximum A Posteriori (MAP) estimator, which maximizes the posterior distribution for all values of the parameter, as:

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} \log p(\theta|X). \tag{4.1.2}$$

The definition of the MAP (4.1.2) is equivalent to :

$$\begin{aligned}\hat{\theta} =&\arg\max_{\theta\in\Theta} \log p(\theta|X)\\ =&\arg\min_{\theta\in\Theta} - \log p(\theta|X)\\ =&\arg\min_{\theta\in\Theta} (-\log p(X|\theta) - \log p(\theta))\end{aligned}$$

Now set,

$$\rho(\theta, X) = -\log p(\theta|X),$$

and $I(\theta)$ and λ_n^2 such that

$$-\log p(\theta) = \lambda_n^2 I(\theta)$$

or equivalently

$$p(\theta) = \exp(-\lambda_n^2 I(\theta)),$$

we obtain an other equivalent definition of the Maximum A Posteriori estimator:

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} \left(\rho(\theta, X) + \lambda_n^2 I(\theta)\right) \tag{4.1.3}$$

As a result, the Bayesian estimator can be viewed as a penalized M-estimator minimizing a loss function $\rho(X, \theta)$ with a penalty $I(\theta)$ while λ_n^2 is a smoothing parameter which determines the influence of both terms. Provided the loss function is convex in θ and the errors have good concentration properties, the behaviour of such estimators is a consequence of the complexity of the sets $\{\theta \in \Theta,\ I(\theta) \leq M\}$ for an appropriate M. Penalized M-estimators have been studied by van de Geer in [vdG00] in a general framework. So, with this analogy, we can deduce rates of convergence for Bayes estimators from results for penalized M-estimators.
For instance, if we consider the Gaussian case, i.e the unknown parameter is the mean of an infinite Gaussian vector, the errors are such that $W_i \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$, and the loss is quadratic. As a consequence, the maximum a posteriori estimator is defined by:

$$\begin{aligned}\hat{\theta}_n =&\arg\min_{\theta\in\Theta} \left(\frac{n}{2\sigma^2}\sum_{i=1}^n |X_i - \theta_i|^2 - \log p(\theta)\right)\\ =&\arg\min_{\theta\in\Theta} \left(\frac{1}{n}\sum_{i=1}^n |X_i - \theta_i|^2 - \frac{2}{n^2}\log p(\theta)\right)\\ =&\arg\min_{\theta\in\Theta} \left(||X - \theta||_n^2 + \lambda_n^2 I(\theta)\right) \end{aligned}\tag{4.1.4}$$

where we have set

$$\lambda_n^2 I(\theta) = -\frac{2}{n^2}\log p(\theta).$$

This problem is linked with the following functional estimation problem: given fixed points on an interval called (z_i), $i = 1, \dots, n$, set the empirical measure $\mathbf{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ and if there exists a basis of the space $\mathbb{L}^2(\mathbf{P}_n)$, ψ_i, $i = 1, \dots, n$, we recognize a penalized Least-Squares estimator of a function $\theta_{\mathbf{0}} = \sum_{i=1}^n \theta_i \psi_i$ when the observations are the X_i.
As it is often the case in Bayesian estimation, the choice of the prior will determine the behavior of the estimator.
If we choose a normal prior with mean 0 and variance $\lambda_n^{-2}\tau_i^{-2}$, for all $i = 1, \dots, n$ we have $\theta_i \sim \mathcal{N}\left(0, \lambda_n^{-2}\tau_i^{-2}\right)$, the estimator is of the following form:

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta}\left(||X-\theta||_n^2 + \lambda_n^2 \sum_{i=1}^n \tau_i^2\theta_i^2 \right) \tag{4.1.5}$$

We recognize here a penalized least-squares estimator with smoothing l^2-penalty.
In the same way, if we consider a Laplacian prior for the parameter of interest: for a positive sequence $\tau = (\tau_i)_{i\geq 1}$, set a Laplacian prior for each θ_i with parameter $\lambda_n^{-2}\tau_i$:

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta}\left(||X-\theta||_n^2 + \lambda_n^2 \sum_{i=1}^n \tau_i|\theta_i| \right) \tag{4.1.6}$$

which is a least squares estimator with a soft-thresholding type penalty. In both cases, the theory of penalized M-estimators gives rates of convergence for these estimators, rates depending on the entropy of the set of functions $\{\theta = \sum_j \theta_j\psi_j, \theta \in \Theta,\ I(\theta) < \infty\}$.
If the errors are not Gaussian but Laplacian, the general problem is to estimate the mean of a Laplacian random variable. Even if we have no L^2 equivalence with the functional estimation, we can still consider the MAP estimator. It can be written as:

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta}\left(||X-\theta||_{1,n} + \lambda_n^2 I(\theta) \right) \tag{4.1.7}$$

with, as usual, a choice of the penalty depending on the prior put on the parameter of interest. We have used the following notation $\|X-\theta\|_{1,n} = \frac{1}{n}\sum_{i=1}^n |X_i - \theta_i|$. We recognize a penalized least absolute deviation estimator.
In all cases, provided a smoothness condition holds over the set $\Theta = \{\theta_i, i = 1, \dots\}$, and some additional entropy conditions on that space, such estimators have been studied by Loubes and van de Geer in [LvdG00] and asymptotic rates of convergence are given. In that point of view, Bayesian estimation appears naturally as a subcase of penalized M-estimation for a proper choice of the penalization.
Following similar ideas, L. Birgé and P. Massart in [BM01] showed that it is possible to give another interpretation using a model selection approach. They proved that the penalized projection estimator is the mode of the posterior distribution with an improper uniform prior distribution on the collection of models and mixing the prior over the parameter on each

model. It could be of interest to investigate the property of such prior and to wonder whether it selects, a posteriori, the good model.

However, such generalization provides general theorems which do not always give an optimal majoration of the error term. Indeed, in most cases, a direct approach leads to better bound for mean squares error.

4.2 Adaptive Bayesian Estimation

AbstractWe present an empirical Bayes estimator of a function. Using a prior on the number of derivatives, we study the behavior of the posterior distribution. Using empirical risk minimization, we construct an adaptive estimator and give a Bayesian interpretation.

AMS 1991 subject classifications. Primary 62G05, secondary 62G20.
Key words and phrases. Adaptive estimation, Bayes estimation, penalization.

4.2.1 Introduction

Consider the infinite dimensional model:

$$X_i = \theta_i + W_i, \; i = 1, 2, \dots \tag{4.2.1}$$

where $\theta = (\theta_1, \theta_2, \dots) \in l^2(\mathbb{N})$ is an infinite dimensional vector and the random variables W_i are independent identically distributed with law $\mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$. The parameter θ is unknown and the goal of our study is to make inference about θ. One of the interest with this model is due to its equivalence with the white noise model

$$\left\{ \begin{array}{l} dX(t) = f(t)dt + \frac{1}{\sqrt{n}} dW(t), \; 0 \leq t \leq T \\ f \in L^2[0, T] \end{array} \right. \tag{4.2.2}$$

Given an orthonormal bases $(\psi_i)_i$ of $L^2[0, T]$, the problem of estimating an unknown signal is reduced to the problem of estimating the mean of an i.i.d Gaussian sample (4.2.1) with

$$\begin{aligned} X_i &= \int_0^T \psi_i(t) dX(t) \quad W_i = \frac{1}{\sqrt{n}} \int_0^T \psi_i(t) dW(t) \\ \theta_i &= \int_0^T f(t) \psi_i(t) dt \end{aligned}$$

The signal is reconstructed from the coefficients using $\theta = \sum_j \theta_j \psi_j$ and the map between θ and f is an isometric isomorphism. Such equivalence has been stated by Nussbaum in [Nus96] in density estimation and by Brown and Low in [BL96] in the regression model.

The quality of the estimation depends on smoothness properties of f. For a fixed basis, assuming f to belong to a specific class of signals, it implicitly yields corresponding assumptions on its coefficients (θ_j). So the Bayesian approach seems natural: one assigns a prior to θ and looks at its posterior distribution. Such issue has been tackled by Diaconis and Freedman in [DF97] or Freedman in [Fre99]. They consider an independent Gaussian prior for each θ_i and studied the Bayes estimator $\hat{\theta}_i$ as well as the quadratic loss $||\hat{\theta} - \theta||_2$. The variance of the Gaussian prior depends on a parameter which measures the smoothness of the function $\sum_j \theta_j \psi_j$, which is unknown in practice. But the use of an informative prior prevents adaptive estimation. So this approach can not be used practically. Later, Beltiser and Ghosal in [BG00] assume that θ satisfies a smoothness condition, which is characterized by a smoothness index, say s_0, which describes the order of decay of the coefficients to zero. They assume that θ belongs to the space defined as follows:

$$\Theta(s_0) = \{\theta : \sum_{i=1}^{\infty} i^{2s_0}\theta_i^2 < \infty\}.$$

For instance if the function θ is in a Sobolev space H^{s_0}, the inequality is satisfied with $s = s_0$. Since this parameter is unknown, it is viewed, in a Bayesian point of view as an hyperparameter of the model, which means that this quantity is considered as a realization of an unknown random variable. That is the reason why, these authors define a collection of models and construct a prior over the whole range of models. The models depend on the parameter s which belongs to a set of indexes $\mathcal{S}$. They show that the hierarchical prior selects the best model in the sense that the posterior distribution converges at the rate of convergence given by the true model. Such a technique is also used by Ghosal and van der Vaart in [GvdV00] in the context of density estimation. However the resulting adaptive estimator is not very natural. The problem of adaptive estimation in this setting was originally studied by Efromovich and Pinsker in [EP80] who give minimax rates of convergence.

In this work, we propose to consider generalized maximum posterior estimator, which maximizes the posterior likelihood over all the models and to extend results over the posterior distribution to this estimator. The organization of the paper is as follows. In the next section, we study the case when the smoothness parameter is known. Then in section 4.2.3, we construct a Gaussian prior over a collection of models and define the posterior maximum estimators $(\hat{\theta}_n, \hat{s}_n)$ which estimates the quantities (θ, s_0). We show that the estimator $\hat{\theta}_n$ is an adaptive estimator of θ in the sense that, without prior knowledge of the regularity of the function, it achieves the optimal rate of convergence. The proof of the main result is based on a selection lemma which describes the asymptotic behavior of $\hat{s}_n$. The prior plays the role of a penalty over the smoothness of the model, and the estimators minizes a penalized quadratic criterion with this smoothing penalty. All the proofs are postponed to Section 4.2.4.

4.2.2 Estimation with known smoothness

We recall that we observe an infinite dimensional vector $X = (X_1, X_2, \dots)$ such that

$$X_i \sim \mathcal{N}\left(\theta_i, \frac{\sigma^2}{n}\right)$$

which corresponds to the statistical model described in (4.2.1). We only observe the i.i.d sample X_i, $i = 1, \dots, n$. Our aim is to estimate the infinite sequence $\theta = (\theta_i)_i \in \Theta$. Actually, this parameter is the coefficients of a function $\theta = \sum_i \theta_i \psi_i$. For this, we consider a parameter of the space that describes the regularity of the function, and we define the following smoothness assumption: the coefficients satisfy

$$(\theta_i)_i \in l^2(\mathbb{N}) \tag{4.2.3}$$

and there exists a smoothness coefficient s_0 and a finite positive constant M such that:

$$s_0 = \arg\max_{s \in \mathbb{R}} \left(\sum_{i=1}^{\infty} i^{2s} \theta_i^2 < \infty \right).$$

This parameter can be viewed as the maximal number of derivatives of the function lying in the L^2 space. It is linked with the Sobolev structure of the functional space since the parameter space becomes:

$$\theta \in \Theta(s_0) = \{\theta = (\theta_i)_i, \sum_{i=1}^{\infty} i^{2s_0} \theta_i^2 \leq M\} \tag{4.2.4}$$

From now and throughout all this section, we assume that we know the true regularity s_0. Such a set can be viewed as a Pinsker ellipsoid [Pin80]. Indeed, we consider the equivalent functional estimation problem: given an orthonormal family of $\mathbb{L}^2$ namely, $\psi_j, j \geq 1$, consider, as in section 4.2.1, the function whose decomposition in the basis (ψ_j) is given by:

$$\theta(t) = \sum_{i=1}^{\infty} \theta_i \psi_i(t)$$

So the problem turns to estimate a function $\boldsymbol{\theta_0}$ in Pinsker's ellipsoid described in [Pin80]

$$\boldsymbol{\theta_0} \in \Theta(s) = \{\theta = \sum_{i=1}^{\infty} \theta_i \psi_i, \sum_{i=1}^{\infty} i^{2s} \theta_i^2 \leq M\}.$$

Pinsker showed that the exact asymptotics of the quadratic risk over this ellipsoid is given by:

$$E||\hat{\theta}_n - \theta||^2 \leq C_s(M) n^{-\frac{2s}{2s+1}}$$

where $C_s(M) = M^{\frac{1}{2s+1}} \gamma(s)$ and $\gamma(s) = (2s+1)^{\frac{1}{2s+1}} \left(\frac{s}{s+1}\right)^{\frac{2s}{2s+1}}$ is the Pinsker constant.
We use Bayes methods to construct an estimator which achieves this minimax rate of convergence. To perform Bayes estimation, we consider a Gaussian prior over the coefficients

$$\theta_i \sim \mathcal{N}(0, \tau_i^2(s_0)),\ i \geq 1,\ \tau_i^2 = i^{-2s_0}.$$

The coefficients θ_i are chosen independent and independent from the additional noise. The prior information means that the coefficients are normally distributed but since the variance goes to zero as i increases, we assume that the coefficients tends to zero which is a common assumption for coefficients of a regular function onto basis, which provide spare enough representation. Indeed, wavelet bases, introduced by Meyer in [Mey87], have the property that only the few largest wavelet coefficients contain the information conveyed by the signal θ. Such property gives sense to the prior. We can point out that some work has been recently done in that direction and special behavior of wavelet coefficients has been more precisely

modelized by Sapatinas and al. in [ASS98] or Johnstone and Silverman in [JS01].
As a result, the distribution of the data given the parameter, say $p(X|\theta)$ is such that

$$X_i|\theta_i \sim \mathcal{N}(\theta_i, \tau_i^2).$$

Using Bayes rule, we obtain the posterior distribution $p(\theta_i|X_i)$:

$$\theta_i|X_i \sim \mathcal{N}(\frac{\tau_i^2}{\sigma_n^2+\tau_i^2}X_i, \frac{\sigma_n^2\tau_i^2}{\sigma_n^2+\tau_i^2}).$$

So the estimator, maximum of the posterior law is given by:

$$\begin{aligned}(\hat{\theta}_i) &= \arg\max_{(\theta_i)} \log p(\theta|X)\\ &= \arg\min_{(\theta_i)} - \log p(\theta|X)\\ &= (\frac{\tau_i^2}{\sigma_n^2+\tau_i^2}X_i) \qquad (4.2.5)\end{aligned}$$

The Maximum A Posteriori estimator $\hat{\theta}_n$ defined in (4.2.5) converges if $\tau_i^2 = \lambda_n^2 i^{-2s_0}$ with $s_0 > \frac{1}{2}$ and $\theta_0 \in \Theta$ defined above (4.2.4) for a quadratic loss to the true parameter θ_0 at a rate of convergence $n^{-\frac{2s_0}{2s_0}+1}$ as soon as the smoothing sequence λ_n^2 decreases to zero but not too quickly: indeed the following theorem describes the asymptotic behavior of the maximum a posteriori estimator.

Theorem 4.2.1. *The Maximum a posteriori estimator of $\theta_0 \in \Theta(s_0)$ for a given prior $\theta_i \sim \mathcal{N}(0, \lambda_n^2 i^{-2s_0})$, $s_0 > \frac{1}{2}$ is consistent for a good choice of the smoothing sequence λ_n^2. Moreover for $\lambda_n^2 = n^{-\frac{1}{2s_0+1}}$, the estimator achieves the minimax rate of convergence. As a matter of fact, there exists a finite constant $C = C(M)$ such that*

$$E||\hat{\theta}_n - \theta_0||^2 \leq Cn^{-\frac{2s_0}{2s_0+1}}. \qquad (4.2.6)$$

This result can be easily connected to the theory of M-estimation. As a matter of fact, this estimator can be written in the following way:

$$(\hat{\theta}_i) = \arg\min_{\theta\in\Theta(s_0)} \left(||Y - \theta(t)||_n^2 + \tilde{\lambda}_n^2 \sum_{i=1}^n i^{2s_0}\theta_i^2 \right)$$

where we have set $\tilde{\lambda}_n^2 = \frac{1}{n\lambda_n^2} = n^{-\frac{2s}{2s+1}}$. The estimator minimizes a quadratic loss function with a penalty and is therefore a penalized M-estimator. Penalized M-estimators have been intensely studied by S. van de Geer in [vdG90], [vdGW96] or [vdG00]. Using such results, we know that this estimator converges at a rate in $O_{\mathbf{P}}(\tilde{\lambda}_n)$. The proof of this statement relies on the behavior of empirical process in a ball whose entropy can be upper bounded. So, if the optimal parameter λ_n has been correctly chosen, we find the rate of convergence in $n^{\frac{-2s_0}{2s_0+1}}$, which corresponds to the minimax rate of convergence for a quadratic loss.
Since the choice of the smoothing parameter requires the knowledge of the smoothness coefficient s_0, previous method can not be used in practice. Adaptive estimation imply to let the data speak from themselves and pick automatically the right order.

4.2.3 Adaptive estimation

In our Bayesian framework, we have, in a first step, considered the unknown function as a realization of a random variable: its prior law heavily depends on an unknown parameter s_0, which can be interpretated as a smoothness parameter. It seems rather natural to consider this parameter also as a random variable. For this, we make the assumption that the smoothness coefficient belongs to a finite class of indices $s \in \mathcal{S}$. This class is finite but its size may depend on n. We assume that $s_{\max}$ is finite. We also assume that the true coefficient s_0 belongs to that set, which can be written as

$$\mathcal{S} = \{s_m, m \in \mathcal{M} \subset \mathbb{Z}\}$$

where the s_m are sorted in an increasing order and $\mathcal{M}$ is a subset of $\mathcal{S}$. Over $\mathcal{S}$, we construct a prior probability q and define $q_m = q(s = s_m)$ which is assumed to be positive. So we have $\sum_{m\in\mathcal{M}} q_m = 1$ and $\forall m \in \mathcal{M}$, $q_m > 0$. The prior has to be chosen such that the posterior law will tend to choose a model close to the true one $s = s_0$. As a matter of fact, to each parameter $s \in \mathcal{S}$, we associate the corresponding model in the sequence space

$$\Theta(s) = \{\theta \in l^2(\mathbb{N}), \sum_{i=1}^{\infty} i^{2s}\theta_i^2 \leq M\}$$

and the set $\Theta = \cup_{s\in\mathcal{S}}\Theta(s)$ is the set of all possible models.
In this framework, the unknown parameter is a couple of variables:

$$(\theta_0, s_0) \in \Theta \times \mathcal{S}$$

and the prior distribution is given by $p(\theta, s) = p(\theta|s)q(s)$. The prior over a single model $\Theta(s)$ is the same prior considered in section 4.2.2: $\theta_i \sim \mathcal{N}(0, \lambda_n^2(s)\tau_i^2(s))$ with $\tau_i^2(s) = i^{-2s}$ and $\lambda_n^2(s) = n^{-\frac{1}{2s+1}}$. Indeed, if s were the true regularity, this choice of prior would lead to an estimator with optimal rate of convergence. As a result, the natural prior over Θ is a mixture of all priors on each model $\Theta(s)$. The framework of our study is the following: we observe $X_i = \theta_i + W_i$, $i = 1, \ldots, n$ (4.2.1) with two unknown parameters $(\theta_0, s_0) \in \Theta \times \mathcal{S}$ over which we define a prior. So the laws of the parameters are the following:

$$\begin{aligned} X_i|(\theta, s) &\sim \mathcal{N}(\theta_i, \sigma_n^2)\\ \theta_i|s &\sim \mathcal{N}(0, \lambda_n^2(s)\tau_i^2)\\ X_i|s &\sim \mathcal{N}(0, \lambda_n^2(s)\tau_i^2 + \sigma_n^2)\\ s &\sim q(s)\\ W_i &\sim \mathcal{N}(0, \sigma_n^2)\end{aligned} \tag{4.2.7}$$

The estimator we consider maximizes the posterior distribution:

$$(\hat{\theta}_n, \hat{s}_n) = \arg\max_{(\theta,s)\in\Theta\times\mathcal{S}} \log p(\theta, s|X) \tag{4.2.8}$$

$\hat{\theta}_n$ is an estimator of the parameter of interest θ while $\hat{s}_n$ should approximate the auxiliary smoothness parameter s_0. We can find a more explicit version of the estimators using Bayes rule:

$$p(\theta, s|X) = \frac{p(X|\theta, s)p(\theta|s)q(s)}{p(X)}.$$

So (4.2.8) becomes:

$$\begin{aligned}(\hat{\theta}_n, \hat{s}_n) &= \arg \min_{(\theta,s)\in\Theta\times\mathcal{S}} - \log p(\theta, s|X) \\ &= \arg \min_{(\theta,s)\in\Theta\times\mathcal{S}} (- \log p(X|\theta, s) - \log p(\theta|s) - \log q(s)) \qquad (4.2.9)\end{aligned}$$

defining the estimator has the minimizer of the global penalized empirical risk.

We now can see that such an expression (4.2.9) can be written into two different ways whether we decide to minimize the expression first on s then on θ or the contrary. On the one hand we can write:

$$\begin{aligned}(\hat{\theta}_n, \hat{s}_n) &= \arg \min_{(\theta,s)\in\Theta\times\mathcal{S}} \left(\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \theta_i)^2 + \frac{1}{n} \sum_{i=1}^n \log \lambda_n^2(s)\tau_i^2 + \frac{1}{\lambda_n^2(s)n} \sum_{i=1}^n \tau_i^{-2}\theta_i^2 - \frac{2}{n} \log q(s) \right) \\ &= \arg \min_{\theta\in\Theta} \left(\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \theta_i)^2 + \min_{s\in\mathcal{S}} \left[\frac{1}{n} \sum_{i=1}^n \log \lambda_n^2(s)\tau_i^2 + \frac{1}{\lambda_n^2(s)n} \sum_{i=1}^n \tau_i^{-2}\theta_i^2 - \frac{2}{n} \log q(s) \right] \right) \\ &= \arg \min_{\theta\in\Theta} \left(\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \theta_i)^2 + \min_{s\in\mathcal{S}} \left[n^{-\frac{2s}{2s+1}} \sum_{i=1}^n i^{2s}\theta_i^2 + I(s) \right] \right) \qquad (4.2.10)\end{aligned}$$

where we set

$$I(s) = \frac{1}{n} \sum_{i=1}^n \log(n^{-\frac{1}{2s+1}} i^{-2s}) - \frac{2}{n} \log(q(s)) \qquad (4.2.11)$$

We have put the stress on the minimization over the coefficients $\theta = (\theta_i)_i$. We recognize a penalized least squares estimator with a complexity penalty. Indeed the estimator of θ_0 can be written in the following form:

$$\hat{\theta}_n = \arg \min_{\theta\in\Theta} \left(\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \theta_i)^2 + \text{pen}(\theta)_{\hat{s}} \right) \qquad (4.2.12)$$

$$\text{pen}(\theta)_{\hat{s}} = \min_{s\in\mathcal{S}} \left[n^{-\frac{2s}{2s+1}} \sum_{i=1}^n i^{2s}\theta_i^2 + I(s) \right] \qquad (4.2.13)$$

Provided some additional hypothesis hold, the asymptotic behavior of such estimator can be found using ideas from empirical risk minimization that can be found in [vdG01]. The estimator minimizes over Θ an l^2 loss balanced by a penalty term (4.2.13). This penalty minimizes over all possible indexes $s \in \mathcal{S}$ a sum of two terms, one depending on the smoothness of the function while the second depends on the choice of the prior $q(s)$ on $\mathcal{S}$. For a convenient choice

of this prior, we get $I(s) = \lambda_0^2 n^{-\frac{2s}{2s+1}}$ for a given constant λ_0^2. Van de Geer in [vdG01] showed in a different context, the convergence of the estimator at the minimax rate of convergence. Our method is close to model selection methods since, with a selection lemma, we first pick an optimal smoothness index $\hat{s}_n$ and then reconstruct the estimator using this value: $\hat{\theta}(\hat{s}_n)$. The major difference with methods inspired by Birgé and Massart see for instance [BBM99], [BM98], [BM97] or Lugosi and Nobel in [LN99] lies in the fact that we penalize the roughness of the model and not its dimension nor its complexity.
So, on the other hand, if we minimize the quantity (4.2.10) first over θ and afterwards over s, we find after some calculations and taking for simplicity reasons $\sigma^2 = 1$, the following form of the estimator:

$$\begin{cases} \hat{s}_n & = \arg\min_{s\in\mathcal{S}} \left(\sum_{i=1}^{n} \frac{1}{\frac{1}{n} + n^{-\frac{1}{2s+1}} i^{-2s}} X_i^2 + nI(s) \right) \\ \hat{\theta}_i(\hat{s}_n) & = \frac{1}{1 + n^{-\frac{2\hat{s}_n}{2\hat{s_n}+1}} i^{2\hat{s}_n}} X_i,\ i = 1, \dots, n \end{cases} \tag{4.2.14}$$

The first estimator $\hat{s}_n$ can be considerated as a biased estimator of the smoothness of the model. The second estimator is the solution of a quadratic penalized loss in the estimated model $\Theta(\hat{s}_n)$. In fact, for a particular choice of the penalty $I(s) = n^{-\frac{2s}{2s+1}}\lambda_0^2$ where λ_0^2 will be made precise later, i.e for a good choice of the prior q over the set $\mathcal{S}$, we will show that the estimator $\hat{s}_n$ always overestimates the regularity of the model. As a matter of fact, the following lemma shows that the probability of underestimating the regularity of the model tends to zero in probability.

Lemma 4.2.2. *Assume that $\mathcal{S}_- = \{s_m,\ m < 0\}$ is finite, with cardinality that may depend polynomaly on n, and choose the prior q in the following way: there exists a positive constant λ_0^2 such that: $I(s) = \lambda_0^2 n^{-\frac{2s}{2s+1}}$. For $s < s_0$, there exist constants $c_1 > 0$, c_2 and an integer N such that for every $n \geq N$*

$$\mathbf{P}(\hat{s}_n = s) \leq c_1 \exp(-c_2^2 n^{\frac{2s_0}{2s_0+1}\frac{1}{s_0+s}}) \tag{4.2.15}$$

So as a result we have

$$\mathbf{P}(\hat{s}_n < s_0) \overset{n\to\infty}{\longrightarrow} 0. \tag{4.2.16}$$

The proof of this selection lemma involvess some calculations and is postponed to the appendix.

Remark 4.2.3. *This lemma is the frequentist version of the Bayesian selection lemma in [BG00], over the concentration of the prior around smoother models. With a prior over $\mathcal{S}$ which may be taken infinite, such that $\forall m \in \mathbb{Z}$, $q(s = s_m) > 0$, Belitser and Ghosal, have proved that the posterior distribution converges at the optimal rate of convergence $n^{-\frac{2s_0}{2s_0+1}}$. More precisely, there exist an integer N and a constant $c > 0$ such that for any $m < 0$ and $n > N$,*

$$E\,\mathbf{P}(s = s_m | X) \leq \frac{q_m}{q_0} \exp\left(-cn^{\frac{1}{s_0+s_m+1}}\right) \tag{4.2.17}$$

As a consequence we have

$$\mathbf{P}(s < s_0|X) \to 0. \tag{4.2.18}$$

in probability.

Using this pseudo estimator of the smoothness index, the estimator turns to be the solution of a minimization program in the space $\Theta(\hat{s}_n)$. Direct calculations lead to a smoothed estimator with a smoothing coefficient depending on the data, since it is a function of $\hat{s}_n$. Even if $\hat{s}_n$ does not approximate the true value s_0, it selects smoother models such that the estimator $\hat{\theta}(\hat{s}_n)$ achieves the optimal rate of convergence. Since the estimator is constructed without any knowledge of the regularity, the estimator is adaptive. This property is a direct consequence of the Bayesian point of view, which takes more into account the data, and provides adaptivity. Moreover, it links the smoothing effect of a global penalty such as $\text{pen}(\theta) = \min_{s\in\mathcal{S}} \left[n^{-\frac{2s}{2s+1}} \left(\sum_{i=1}^n i^{2s}\theta_i^2 + \lambda_0^2\right)\right]$, and the behavior of the underlying posterior law. So penalized M-estimation with smoothing penalties acts as model selection, which advantages smooth enough models. The following theorem describes the optimality of such estimator.

Theorem 4.2.4. *The estimator obtained by maximizing the posterior distribution for a Gaussian prior distribution depending on an hyperparameter $s \in \mathcal{S}$ is convergent provided the collection of regularity indexes is finite (possibly depending polynomaly on n). There exists a finite constant C, depending on M, such that*

$$E||\hat{\theta}(\hat{s}_n) - \theta_0||_n^2 \leq C n^{-\frac{2s_0}{2s_0+1}} \tag{4.2.19}$$

The MAP estimator is adaptive for L^2 losses.

Remark 4.2.5. *Using the first form of the estimator* (4.2.10) *with the special choice $I(s) = \lambda_0^2 n^{-\frac{2s}{2s+1}}$, we can apply results from van de Geer in [vdG01], which rely on the asymptotic behavior of the estimator with the complexity of the set $\Theta^*(\delta)$ defined by*

$$\Theta_*(\delta) = \{\theta \in \Theta : \tau^2(\theta|\theta^*) \leq \delta^2\}$$

where we have set a pseudo distance

$$\tau^2(\theta|\theta_0) = ||\theta - \theta_0||_2^2 + \text{pen}^2(\theta)$$

for $\text{pen}(\theta)^2 = \min_{s\in\mathcal{S}} n^{-\frac{2s}{2s+1}} \left(\sum_{i=1}^n i^{2s}\theta_i^2 + \lambda_0^2\right)$. *Then define* $\theta^* = \arg\min_{\theta\in\Theta} \tau^2(\theta|\theta_0)$ *which has an interpretation as trade-off between bias and variance. As soon as the entropy $H(\delta, \Theta_*)$, i.e the logarithm of the minimum number of balls necessary for a δ-covering of the set Θ^* with respect to quadratic norm, is such that*

$$\int_0^1 H(u, \Theta_*)du \leq c\sqrt{n}\delta_n^2 \tag{4.2.20}$$

for a given constant c then theorem 2.1 gives the rate of convergence of the estimator. In this case, δ_n^2 can be chosen equal to $\delta_n = \sqrt{\frac{s_{\max}\log n}{n}}$. The rate of convergence is deduced from this entropy upper bound.

Remark 4.2.6. *Bayesian estimator and penalized estimators are deeply linked. In our work, we have enlighted the correspondance between smoothing penalties and the posterior mode with a Gaussian prior. Following the same ideas, L. Birgé and P. Massart in [BM01] showed that it is also possible to give another interpretation using a model selection approach. They proved that the penalized projection estimator is the mode of the posterior distribution with an improper uniform prior distribution on each model and, for each model, a prior corresponding to the prior for the true model. It could be of interest to investigate the property of such prior and to wonder whether it selects a posteriori the good model.*

Remark 4.2.7. *We have only used linear estimator with an adaptive Bayes-type smoothing effect. It could be also meaningfull to extend previous results to non linear estimators such as thresholded estimators, studied in the wavelet case by Donoho and al. in [DJKP96a], [DJKP96b], [DJKP97].*

4.2.4 Proofs

<u>Proof of theorem 5.2.2</u>:

Proof. We have found that the estimator $\hat{\theta}_i$ can be written for all $i \geq 1$:

$$\hat{\theta}_i = \frac{\tau_i^2}{\sigma_n^2 + \tau_i^2} X_i.$$

So since $E||\theta - \hat{\theta}||^2 = \sum_{i=1}^n |\theta_i - \hat{\theta}_i|^2$, we calculate the difference between the true coefficient and its estimator:

$$\theta_i - \hat{\theta}_i = (1 - \frac{\tau_i^2}{\sigma_n^2 + \tau_i^2})\theta_i - \frac{\tau_i^2}{\sigma_n^2 + \tau_i^2} W_i$$

Finally we find

$$\theta_i - \hat{\theta}_i \sim \mathcal{N}\left((1 - \frac{\tau_i^2}{\sigma_n^2 + \tau_i^2})\theta_i, \left(\frac{\tau_i^2}{\sigma_n^2 + \tau_i^2}\right)^2 \sigma_n^2\right). \tag{4.2.21}$$

$$\begin{aligned} E||\hat{\theta} - \theta||^2 &= \left(\sum_{i=1}^{\infty} \frac{\sigma_n^4 \theta_i^2}{(\sigma_n^2 + \tau_i^2)^2} + \sum_{i=1}^{\infty} (\frac{\tau_i^2}{\sigma_n^2 + \tau_i^2})^2 \sigma_n^2\right) \\ &= T_1 + T_2. \end{aligned}$$

The first term T_1 can be majorated by:

$$\begin{aligned} T_1 &= \sum_{i=1}^{\infty} \frac{\sigma_n^4 \theta_i^2}{(\sigma_n^2 + \tau_i^2)^2} \\ &\leq \sigma^2 \sum_{i=1}^{\infty} \frac{\sigma^2 i^2 s}{n + \sigma^2 i^{2s}} \frac{1}{n + \sigma^2 i^{2s}} i^{2s} \theta_i^2 \\ &\leq M \frac{\sigma^2}{n}. \end{aligned}$$

The second term T_2 can be controlled by the following inequality due to a comparison between a sum and an integral:

$$\begin{aligned} T_2 &= \sum_{i=1}^{\infty} \frac{\sigma_n^2 \tau_i^4}{(\sigma_n^2 + \tau_i^2)^2} \\ &\leq (\frac{\sigma^2}{n})^{1-\frac{1}{2s}} \int_0^{\infty} \frac{du}{(1+u^{2s})^2} \\ &\leq Cn^{\frac{1}{2s}-1} \end{aligned}$$

for C a finite constant, where we have set s such that the last integral is finite, so $4s > 1$. So we find that as soon as $s > \frac{1}{4}$,

$$E||\theta - \hat{\theta}||_n^2 \leq Cn^{-1+\frac{1}{2s}} + M\frac{\sigma^2}{n}$$

which proves that the estimator converges with a rate of convergence of $n^{1-\frac{1}{2s}}$ as soon as $s > \frac{1}{2}$.

As in the previous proof we start from the expression of the estimator and replace i^{-2s} by $\lambda_n^2 i^{-2s}$. This gives the following bound:

$$\begin{aligned} E||\hat{\theta} - \theta||^2 &= \sum_{i=1}^{\infty} |\theta_i - \hat{\theta}_i|^2 \\ &\leq \sum_{i=1}^{\infty} \frac{\sigma_n^4 \theta_i^2}{(\sigma_n^2 + \lambda_n^2 i^{-2s})^2} + \sum_{i=1}^{\infty} \frac{\sigma_n^2}{(1 + \sigma_n^2 \frac{i^{2s}}{\lambda_n^2})^2} \\ &\leq \frac{M\sigma^2}{n\lambda_n^2} + (\sigma^2)^{1-\frac{1}{2s}} (\frac{1}{n})^{1-\frac{1}{2s}} (\frac{1}{\lambda_n^2})^{-\frac{1}{2s}} \qquad (4.2.22) \end{aligned}$$

where we have used the regularity property of the sequence of coefficients and the comparison between integral and sum. We have also chosen s such that the integral is finite, which gives raise to the condition $s > \frac{1}{4}$. Now choose the smoothness parameter λ_n^2 to minimize the upper bound, we find

$$\lambda_n^2 = n^{-\frac{1}{2s+1}}$$

and find the result:

$$E||\hat{\theta} - \theta||^2 \leq 2\sigma^2 \sup(M, (\sigma^2)^{-\frac{1}{2s}}) n^{-\frac{2s}{2s+1}}.$$

□

Proof of Lemma 5.1.12

Proof. In this proof, for sake of simplicity we consider that the errors have variance equal to 1. The estimator of the smoothness coefficient maximizes the posterior law, so we have by definition of the estimator of the smoothness index

$$\hat{s}_n = \arg\min_{s\in\mathcal{S}} n^{\frac{1}{2s+1}} \left(\sum_{i=1}^{n} \frac{1}{n^{-\frac{2s}{2s+1}} + i^{-2s}} X_i^2 + \lambda_0^2 \right)$$

We make the assumptions that the set of indexes $\mathcal{S} = \{s_m,\, m \in \mathcal{Z}\}$ is a finite set, possibly depending on n and write $s_{\max}$ the maximal value of the parameter. Since the set of indexes is finite we can write:

$$\mathbf{P}(\hat{s}_n < s_0) = \sum_{m<0} \mathbf{P}(\hat{s}_n = s_m)$$

We want to give an upper bound of the last quantity using the property of the estimator:

$$\begin{aligned}
&\mathbf{P}(\hat{s}_n = s) = \\
&\mathbf{P}\left(n^{\frac{1}{2s+1}} (\sum_{i=1}^{n} \frac{X_i^2}{(n\lambda_n^2(s))^{-1} + i^{-2s}} + \lambda_0^2) \le n^{\frac{1}{2s_m+1}} (\sum_{i=1}^{n} \frac{X_i^2}{(n\lambda_n^2(s_m))^{-1} + i^{-2s_m}} + \lambda_0^2),\ \forall m \in \mathbb{Z} \right) \\
&\le \mathbf{P}\left(\sum_{i=1}^{n} \frac{X_i^2}{(n\lambda_n^2(s))^{-1} + i^{-2s}} + \Phi(s) \le \sum_{i=1}^{n} \frac{X_i^2}{(n\lambda_n^2(s_0))^{-1} + i^{-2s_0}} + \Phi(s_0) \right) \\
&\le \mathbf{P}\left(\sum_{i=1}^{n} a_i X_i^2 \le \Phi(s_0) - \Phi(s) \right)
\end{aligned}$$

where we have set

$$\begin{aligned}
a_i &= \frac{1}{(n\lambda_n^2(s))^{-1} + i^{-2s}} - \frac{1}{(n\lambda_n^2(s_0))^{-1} + i^{-2s_0}} \\
\Phi(s) &= \lambda_0^2 n^{\frac{1}{1+2s}}
\end{aligned}$$

We put ourselves in the particular case where $s < s_0$ such that the quantity above is negative. So we obtain:

$$\begin{aligned}
\mathbf{P}(\hat{s}_n = s) &\le \mathbf{P}\left(\exp(-\frac{1}{2}\sum_{i=1}^{n} a_i X_i^2) \ge \exp(\frac{\Phi(s) - \Phi(s_0)}{2}) \right) \\
&\le \exp\left(\frac{\Phi(s_0) - \Phi(s)}{2} \right) E(\exp(-\frac{1}{2}\sum_{i=1}^{n} a_i X_i^2)) \\
&\le \exp\left(\frac{\Phi(s_0) - \Phi(s)}{2} \right) \prod_{i=1}^{n} E(\exp(-\frac{1}{2} a_i X_i^2)).
\end{aligned}$$

Where we have used independence of the random variables X_i and Chebychev's inequality. Now recall the following calculation: if $X \sim \mathcal{N}(\mu, \sigma^2)$ then, for every real $a > -\sigma^{-2}$ we have

$$Ee^{-\frac{a}{2}X^2} = \frac{1}{\sqrt{1 + a\sigma^2}} \exp(-\frac{\mu^2 a}{2(1 + a\sigma^2)}).$$

In our case where $\sigma^2 = \frac{1}{n}$ the condition is equivalent to $a_i > -n$, which is equivalent to

$$i^{-2s} - i^{-2s_0} < \frac{1}{\lambda_n^2}(i^{-2s} + i^{-2s_0}) \to +\infty$$

condition that is fulfilled for a n large enough. So we have:

$$E\left(e^{-\frac{1}{2}a_i X_i^2}\right) = \frac{1}{\sqrt{1+\frac{a_i}{n}}}\exp(-\frac{a_i}{2(1+\frac{a_i}{n})}\theta_i^2).$$

Write

$$\begin{aligned} a_i &= \frac{1}{(n\lambda_n^2(s))^{-1} + i^{-2s}} - \frac{1}{(n\lambda_n^2(s_0))^{-1} + i^{-2s_0}} \\ &= \frac{1}{n^{-\frac{2s}{2s+1}} + i^{-2s}} - \frac{1}{n^{-\frac{2s_0}{2s_0+1}} + i^{-2s_0}} \\ &= \frac{i^{-2s_0} - i^{-2s} + n^{-\frac{2s_0}{2s_0+1}} - n^{-\frac{2s}{2s+1}}}{(n^{-\frac{2s}{2s+1}} + i^{-2s})(n^{-\frac{2s_0}{2s_0+1}} + i^{-2s_0})} \end{aligned}$$

So we have:

$$\begin{aligned} \frac{a_i}{1+\frac{a_i}{n}} &= \frac{i^{-2s_0} - i^{-2s} + n^{-\frac{2s_0}{2s_0+1}} - n^{-\frac{2s}{2s+1}}}{(n^{-\frac{2s}{2s+1}} + i^{-2s})(n^{-\frac{2s_0}{2s_0+1}} + i^{-2s_0}) + \frac{i^{-2s_0} - i^{-2s} + n^{-\frac{2s_0}{2s_0+1}} - n^{-\frac{2s}{2s+1}}}{n}} \\ &= \frac{i^{-2s_0} - i^{-2s} + n^{-\frac{2s_0}{2s_0+1}} - n^{-\frac{2s}{2s+1}}}{R_{i,n}}. \end{aligned}$$

So using the last result we have, if we have set $r(s) = n^{-\frac{2s}{2s+1}}$:

$$\mathbf{P}(\hat{s}_n = s) \leq e^{\frac{\Phi(s_0)-\Phi(s)}{2}} \prod_{i=1}^{n} \frac{1}{\sqrt{1+a_i/n}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \frac{i^{-2s_0} - i^{-2s} + r(s_0) - r(s)}{R_{i,n}}\theta_i^2\right).$$

But since

$$\exp(\frac{\Phi(s_0)-\Phi(s)}{2}) = \exp(\lambda_0^2(1 - \frac{n^{\frac{1}{2s+1}}}{n^{\frac{1}{2s_0+1}}}))$$

we have for large n:

$$\exp(\frac{\Phi(s_0)-\Phi(s)}{2}) \leq \sqrt{1 + a_i i^{-2s_0} + \frac{a_i}{n}}.$$

As a consequence:

$$e^{\frac{\Phi(s_0)-\Phi(s)}{2}} \prod_{i=1}^{n} \frac{1}{\sqrt{1+a_i/n}} \leq \prod_{i=1}^{n}\left(1 + \frac{a_i i^{-2s_0}}{1+a_i/n}\right).$$

So

$$\begin{aligned}
\mathbf{P}(\hat{s}_n = s) &\leq \prod_{i=1}^{n}(1+\frac{a_i i^{-2s_0}}{1+a_i/n})^{-\frac{1}{2}} \exp -\frac{a_i\theta_i^2}{2(1+a_i/n)} \\
&\leq \exp\left(\frac{1}{2}\sum_{i=1}^{n}\frac{a_i}{1+a_i/n}(i^{-2s_0}-\theta_i^2)\right) \\
&\leq \exp\left(\frac{1}{2}\sum_{i=1}^{n}\frac{(i^{-2s_0}-i^{-2s})+(r(s_0)-r(s))}{R_{i,n}}(i^{-2s_0}-\theta_i^2)\right) \\
&\leq \exp(\frac{S_1+S_2}{2}).
\end{aligned}$$

where we have set

$$\begin{aligned}
S_1 &= \sum_{i=1}^{n}\frac{i^{-2s_0}-i^{-2s}+r(s_0)-r(s)}{R_{i,n}} i^{-2s_0} \\
S_2 &= -\sum_{i=1}^{n}\frac{i^{-2s_0}-i^{-2s}+r(s_0)-r(s)}{R_{i,n}}\theta_i^2,
\end{aligned}$$

with

$$R_{i,n} = (r(s)+i^{-2s})(r(s_0)+i^{-2s_0}) + \frac{i^{-2s_0}-i^{-2s}+r(s_0)-r(s)}{n}.$$

Lemma 4.2.8. *there exists an integer N such that for every $n \geq N$ we have*

$$\begin{aligned}
S_1 &\leq -\frac{1}{12}(n\lambda_n^2)^{\frac{2}{s_0+s}} \\
S_2 &\leq \frac{1}{24}(n\lambda_n^2)^{\frac{2}{s_0+s}}.
\end{aligned}$$

Proof. Firstly we notice that

- $i^{-2s_0} - i^{-2s} \leq 0$, $\forall i > 0$
- there exists an integer $I = [2^{\frac{1}{2(s_0-s)}}]$ such that $\forall i > I$ we have

$$i^{-2s_0} - i^{-2s} \leq -\frac{1}{2}i^{-2s}.$$

Moreover, we notice that the dominant term in $R_{i,n} = (n\lambda_n^2)^{-2} + (n\lambda_n^2)^{-1}(i^{-2s_0}+i^{-2s}) + i^{-2(s+s_0)} + n^{-1}(i^{-2s_0}-i^{-2s})$ is $i^{-2(s_0+s)}$ as soon as $i \leq r(-s_0)^{\frac{1}{s+s_0}}$. So putting together these results we see that there exists N_1 so that for every $n \geq N_1$

$$S_1 \leq -\frac{1}{12}(r(-s_0))^{\frac{1}{s_0+s}}.$$

The second term is such that:

$$\begin{aligned}
S_2 &\leq \sum_{i=1}^{n} \frac{i^{-2s} - i^{-2s_0} + r(s) - r(s_0)}{R_{i,n}} \theta_i^2 \\
&\leq \sum_{i=1}^{\infty} i^{2s_0} \theta_i^2 + \sum_{i=1}^{n} \frac{r(s)}{R_{i,n}} \theta_i^2 \\
&\leq M + r(s) \sum_{i=1}^{n} \frac{1}{i^{-2s_0} r(s)} \theta_i^2 \\
&\leq M + M \\
&< \infty.
\end{aligned}$$

So there exists N_2 such that for every $n \geq N_2$ we have

$$2M \leq \frac{1}{24} r(-s_0)^{\frac{1}{s_0+s}}$$

and as a result

$$\forall n \geq N_2,\ S_2 \leq \frac{1}{24}(n^{\frac{2s_0}{2s_0+1}\frac{1}{s_0+s}}).$$

So if we choose $N \geq \sup(N_1, N_2)$, the lemma is proved. □

So with the result of the lemma we have that there exists a finite non zero constant c_2 such that, for every $s < s_0$:

$$\mathbf{P}(\hat{s}_n = s) \leq \exp\left(-c_2^2 n^{\frac{2s_0}{2s_0+1}\frac{1}{s_0+s}}\right) \tag{4.2.23}$$

As a direct consequence

$$\begin{aligned}
\mathbf{P}(\hat{s_n} < s_0) &= \sum_{m<0} \mathbf{P}(\hat{s}_n = s_m) \\
&\leq |\mathcal{S}_-| \sup_{s \in \mathcal{S}_-} \left[\exp\left(-c^2 n^{\frac{2s_0}{2s_0+1}\frac{1}{s_0+s}}\right)\right] \\
&\leq |\mathcal{S}_-| \exp\left(-c_2^2 n^{\frac{1}{2s_0+1}}\right) \\
&\overset{n\to\infty}{\longrightarrow} 0,
\end{aligned}$$

as soon as the set $\mathcal{S}_- = \{s_m,\ m < 0\}$ is a finite set, possibly depending polyniommaly on n. So the M-estimator of the smoothness coefficient almost surely overestimates the true value s_0. □

Proof of Theorem 4.2.4

Proof. $\hat{\theta}(\hat{s}_n)$ is the estimator of the minimization problem. We have:

$$\begin{aligned} E\|\hat{\theta}(\hat{s}_n)-\theta_0\|_n^2 &= E(\|\hat{\theta}(\hat{s}_n)-\theta_0\|_n^2 1_{\hat{s}_n<s_0}) \\ &+ E(\|\hat{\theta}(\hat{s}_n)-\theta_0\|_n^2 1_{\hat{s}_n=s_0}) + E(\|\hat{\theta}(\hat{s}_n)-\theta_0\|_n^2 1_{\hat{s}_n>s_0}) \\ &= O(n^{-\frac{2s_0}{2s_0+1}}) + E(\|\hat{\theta}(\hat{s}_n)-\theta_0\|_n^2 1_{\hat{s}_n>s_0}) \end{aligned}$$

The last term describes the behavior of a MAP estimator $\tilde{\theta}_n$ where the models are taken in $\mathcal{S}_+ = \{s_m,\ m>0\}$. After some calculations we get:

$$\begin{aligned} &E(\|\hat{\theta}(\hat{s}_n)-\theta_0\|_2^2 1_{\hat{s}_n>s_0}) \\ &\le E\left(\frac{1}{n}\left(\frac{1}{nn^{-\frac{1}{2\hat{s}_n+1}}}\right)^{-\frac{1}{2\hat{s}_n}} 1_{\hat{s}_n>s_0} + \sum_{i=1}^n \frac{i^{4\hat{s}_n}}{(i^{2\hat{s}_n}+n^{\frac{2\hat{s}_n}{2\hat{s}_n+1}})^2}\theta_i^2 1_{\hat{s}_n>s_0}\right) \\ &\le E\left(n^{-\frac{2\hat{s}_n}{2\hat{s}_n+1}} 1_{\hat{s}_n>s_0} + \sum_{i=1}^n \frac{i^{4\hat{s}_n}}{(i^{2\hat{s}_n}+n^{\frac{2\hat{s}_n}{2\hat{s}_n+1}})^2}\theta_i^2 1_{\hat{s}_n>s_0}\right) \\ &\le n^{-\frac{2s_0}{2s_0+1}} + E\left(\sum_{i=1}^n \frac{i^{4\hat{s}_n}}{(i^{2\hat{s}_n}+n^{\frac{2\hat{s}_n}{2\hat{s}_n+1}})^2}\theta_i^2 1_{\hat{s}_n>s_0}\right) \end{aligned}$$

So it remains to be proved that the residual term is such that

$$E\left(\sum_{i=1}^n \frac{i^{4\hat{s}_n}}{(i^{2\hat{s}_n}+n^{\frac{2\hat{s}_n}{2\hat{s}_n+1}})^2}\theta_i^2 1_{\hat{s}_n>s_0}\right) = O(n^{-\frac{2s_0}{2s_0+1}}).$$

By the definition of $\hat{s}_n$ we have:

$$\hat{s}_n = \arg\min_{s\in\mathcal{S}}\left(\sum_{i=1}^n \frac{X_i^2}{1+n^{\frac{2s}{2s+1}}i^{-2s}} + I(s)\right)$$

so we can write expanding $X_i = \theta_i + W_i$:

$$\sum_{i=1}^n \frac{X_i^2}{1+n^{-\frac{2\hat{s}_n}{2\hat{s}_n+1}}i^{-2\hat{s}_n}} + I(\hat{s}_n) \le \sum_{i=1}^n \frac{X_i^2}{1+n^{-\frac{2s_0}{2s_0+1}}i^{-2s_0}} + I(s_0)$$

$$\begin{aligned} &\sum_{i=1}^n \frac{\theta_i^2}{1+n^{-\frac{2\hat{s}_n}{2\hat{s}_n+1}}i^{-2\hat{s}_n}} + \sum_{i=1}^n \frac{2\theta_i W_i}{1+n^{-\frac{2\hat{s}_n}{2\hat{s}_n+1}}i^{-2\hat{s}_n}} + \sum_{i=1}^n \frac{W_i^2}{1+n^{-\frac{2\hat{s}_n}{2\hat{s}_n+1}}i^{-2\hat{s}_n}} + I(\hat{s}_n) \le \\ &\sum_{i=1}^n \frac{\theta_i^2}{1+n^{-\frac{2s_0}{2s_0+1}}i^{-2s_0}} + \sum_{i=1}^n \frac{2\theta_i W_i}{1+n^{-\frac{2s_0}{2s_0+1}}i^{-2s_0}} + \sum_{i=1}^n \frac{W_i^2}{1+n^{-\frac{2s_0}{2s_0+1}}i^{-2s_0}} + I(s_0) \end{aligned}$$

We will study each term of this inequality separately. Let $s \ge s_0$, we use here the property that we have restricted our study to the cases where $\hat{s}_n \ge s_0$:

1.

$$
\begin{aligned}
(I) &= \sum_{i=1}^{n} \frac{2\theta_i W_i}{1 + n^{-\frac{2s}{2s+1}} i^{-2s}} \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i \frac{\theta_i}{1 + n^{\frac{2s}{2s+1}} i^{-2s}}
\end{aligned}
$$

where the random variables V_i are Gaussian centered with variance 1, identically independent variables. We know that we have

$$(I) = O_{\mathbf{P}}(\sigma_n) \tag{4.2.24}$$

where the asymptotic variance σ_n^2 is defined by

$$
\begin{aligned}
\sigma_n^2 &= \frac{1}{n} \sum_{i=1}^{n} \frac{i^{4s}\theta_i^2}{(i^{2s} + n^{\frac{2s}{2s+1}})^2} \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \frac{i^{2s}}{i^{2s} + n^{\frac{2s}{2s+1}}} \theta_i^2 \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \left(\frac{i^{2s}}{i^{2s} + n^{\frac{2s}{2s+1}}} \right)^{\frac{s_0}{s}} \theta_i^2 \\
&\leq \frac{1}{n} n^{-\frac{2s_0}{2s_0+1}} \sum_{i=1}^{n} i^{2s_0}\theta_i^2 \\
&\leq \frac{M}{n} n^{-\frac{2s_0}{2s_0+1}}
\end{aligned}
$$

As a result the previous majoration gives an upper bound for σ_n:

$$\sigma_n = O_{\mathbf{P}}(n^{-\frac{s_0}{2s_0+1}-\frac{1}{2}}) = O_{\mathbf{P}}(n^{-\frac{2s_0}{2s_0+1}})$$

and

$$(I) = O_{\mathbf{P}}(n^{-\frac{2s_0}{2s_0+1}})$$

2.

$$
\begin{aligned}
(II) &= E\sum_{i=1}^{n} \frac{W_i^2}{1+n^{-\frac{2\hat{s}_n}{2\hat{s}_n+1}} i^{-2\hat{s}_n}} \\
&= \sum_{s>s_0}\sum_{i=1}^{n} \frac{W_i^2}{1+n^{-\frac{2s_m}{2s_m+1}} i^{-2s_m}} q(\hat{s}_n = s) \\
&= \sum_{s>s_0}\left(\sum_{i>n^{\frac{1}{1+2s}}} \frac{W_i^2}{1+n^{-\frac{2s_m}{2s_m+1}} i^{-2s_m}} + \sum_{i\leq n^{\frac{1}{1+2s}}} \frac{W_i^2}{1+n^{-\frac{2s_m}{2s_m+1}} i^{-2s_m}} \right) q(\hat{s}_n = s) \\
&\leq \sum_{s>s_0}\left(\frac{n^{\frac{1}{2s+1}}}{n} + \sum_{i>n^{\frac{1}{1+2s}}} \frac{1}{n^{\frac{2s}{2s+1}} i^{2s}} \right) q(\hat{s}_n = s) \\
&\leq \sum_{s>s_0}\left(n^{-\frac{2s}{2s+1}} + n^{-\frac{2s}{2s+1}} \int \frac{1}{t^{2s}}\, dt \right) q(\hat{s}_n = s) \\
&\leq \sum_{s>s_0}\left(n^{-\frac{2s}{2s+1}} (1+\frac{1}{2s-1}) \right) q(\hat{s}_n = s) \\
&\leq C n^{-\frac{2s_0}{2s_0+1}}
\end{aligned}
$$

for a finite positive constant C and as soon as $s > \frac{1}{2}$.

3.

$$
\begin{aligned}
(III) &= \sum_{i=1}^{n} \frac{\theta_i^2}{1+n^{\frac{2s_0}{2s_0+1}} i^{-2s_0}} \\
&\leq n^{-\frac{2s_0}{2s_0+1}} \sum_{i=1}^{n} i^{2s_0}\theta_i^2 \\
&= O(n^{-\frac{2s_0}{2s_0+1}})
\end{aligned}
$$

4. Over the set $\hat{s}_n > s_0$, $\Phi(\hat{s}_n) = o(n^{-\frac{2s_0}{2s_0+1}})$

Finally, we have:

$$
E\left(\sum_{i=1}^{n} \frac{i^{4\hat{s}_n}\theta_i^2}{(i^{2\hat{s}_n} + n^{\frac{2\hat{s}_n}{2\hat{s}_n+1}})^2} 1_{\hat{s}_n > s_0} \right) + o(n^{-\frac{2s_0}{2s_0+1}}) = O(n^{-\frac{2s_0}{2s_0+1}})
$$

which shows that

$$
E\left(\sum_{i=1}^{n} \frac{i^{4\hat{s}_n}\theta_i^2}{(i^{2\hat{s}_n} + n^{\frac{2\hat{s}_n}{2\hat{s}_n+1}})^2} 1_{\hat{s}_n > s_0} \right) = O(n^{-\frac{2s_0}{2s_0+1}})
$$

concluding the proof. □

4.3 Bayesian Estimation with Wavelets

Assume that we observe a function f, corrupted by a white noise, at fixed points $t_i = \frac{i}{n}$ for $i = 1, \dots, n$. As a result, we observe an n-dimensionnal vector $Y = (Y_1, \dots, Y_n)$ such that:

$$Y_i = f(t_i) + W_i,\ i = 1, \dots, n \tag{4.3.1}$$

where W_i are independent random variables following $W_i \sim \mathcal{N}(0, \sigma^2)$. Given observed discrete data, we may find the vector of its sample discrete wavelet coefficients by performing the Discrete Wavelet Transform. This algorithm, developed by Mallat in [Mal89] provides an orthonormal transform. As a result, the equivalent model is given by:

$$\begin{cases} X_{jk} = \theta_{jk} + \epsilon_{jk} \\ j_1 \geq j \geq 0,\ k = 0, \dots, 2^j - 1,\ 2^{j_1} = n \end{cases} \tag{4.3.2}$$

where ϵ_{jk} are independent random variables following a Gaussian $\mathcal{N}(0, \frac{\sigma^2}{n})$. Estimating the function is performed by inverting the wavelet transform and write:

$$f(x) = \sum_{j=0}^{j_1} \sum_{k=0}^{2^j-1} \theta_{jk} \psi_{jk}(x).$$

For the wavelet model (4.3.2), we make the assumption that there exist a smoothness coefficient s_0, and a finite real M such that

$$(\theta_{jk})_{jk} \in \Theta(s_0) = \{(\theta_{jk}),\ \sum_j 2^{2js_0} \sum_k \theta_{jk}^2 \leq M\}.$$

Consider the function $f(x) = \sum_{(j,k)} \theta_{jk} \psi_{jk}(x)$ where $(\psi_{jk})_{jk}$ is a wavelet basis with regularity $r > s_0$. Then such a condition over the wavelet coefficients implies that the function f belongs to a Besov space $B_{22}^{s_0}$ due to the characterization of Besov spaces in terms of wavelet coefficients that can be found in [DJKP95]. For a fixed resolution level j, if the smoothness parameter s_0 is a known quantity, we consider, for a given sequence λ_n^2 the following prior on the coefficients:

$$\begin{cases} \theta_{jk} \sim \mathcal{N}(0, \lambda_n^2 2^{-2js_0}) \\ j \geq 0,\ k = 0, \dots, 2^j - 1 \end{cases}$$

and the coefficients are independent random variables, with the same prior distribution at each level j, which are also chosen independently from the noise ϵ_{jk}, $\forall (j,k)$. Such prior can be interprated as a modelization attempt of the sparsity of the coefficients. Indeed, as the resolution level of the wavelet decomposition increases, the number of non zero wavelet coefficients decreases, since the Gaussian distribution is more and more concentrated around zero. Using the same ideas of what has been already done, we find analogous laws cf. (4.2.7) by setting $\tau_j^2 = \lambda_n^2 2^{-2js_0}$. An estimator of each wavelet coefficient is the solution of the quadractic minimization problem. The solution of this optimization program has an explicit form:

$$\forall (j,k),\ \hat{\theta}_{jk} = \frac{\tau_j^2}{\frac{\sigma^2}{n} + \tau_j^2} X_{jk} \qquad (4.3.3)$$

We recognize here a smoothed estimator and the asymptotics of the quadratic error are well known. Moreover, since wavelet basis provides unconditional basis for a wide range of spaces, we can pay attention to convergence for other loss functions. The following theorem proves the consistency in B^{σ}_{22}-loss when $\sigma < s_0$, when only a finite number of coefficients is used. Indeed we consider wavelet coefficients between the two resolution level j_0 and j_1 and try to estimate the true coefficients using a posterior mode estimation while the higher levels coefficients are set to zero and the others are kept unchanged. This estimation procedure can be easily implemented. We recalled here a theorem proved in Chapter III:

Theorem 4.3.1. *The Posterior mode for a prior over wavelet coefficients $\theta_{jk} \sim \mathcal{N}(0, \lambda_n^2 2^{-2js_0})$ for $j = j_0, \dots, j_1$, provides a consistent estimator with the following rate of convergence of a function $f = \sum_{(j,k)} \theta_{jk}\psi_{jk} \in B^{s_0}_{22}$. If the smoothing parameter λ_n^2 is chosen such that $\lambda_n^2 = n^{-\frac{2s_0}{1+2s_0}}$ and the resolution level $j_0 = O(1)$ and $j_1 = n^{\frac{1}{2s_0+1}}$, then there exists a finite constant C so that, for every $0 \leq \sigma < s_0$:*

$$E||\hat{\theta} - \theta||^2_{B^{\sigma}_{22}} \leq Cn^{-\frac{2(s_0-\sigma)}{2s_0+1}}.$$

However, all our study relies on the fact that the true value of the smoothness parameter s_0 is known. Since we do not want to impose such a condition over the model, we try to turn this method into an adaptive version. As it has been first studied by Efromovich and Pinsker in [EP80] or in the wavelet case by Donoho, Johnstone, Kerkyacharyan and Picard in [DJKP95] or in [DJ94], we will try to find an estimator achieving optimal asymptotics without prior knowledge of its regularity. The main idea is considering the smoothness parameter as an hyperparameter of the model and put a prior on it. Then we will let the observations choose a posteriori the real model between a collection of various models. The prior law acts as a penalty in model selection. Here we will not use a penalty over the dimension of the model, techniques developed by L. Birgé and P. Massart in [BM97], but the prior we will choose penalizes the regularity of the model in a similar way to complexity regularization developed by G. Lugosi and Nobel in [LN99]. Like in the preceding section, we start from Belitser and Ghosal's result in [BG00]. Indeed, in an allocation model $X_i = \theta_i + W_i$, $\sum_i i^{2s_0}\theta_i^2 \leq M$ described in the previous section, they show that the posterior distribution converges at the optimal rate of convergence. For this, they prove an auxiliary lemma, the selection lemma.

Lemma 4.3.2. *There exist integer N and a constant $c > 0$ such that for any $m < 0$ and $n > N$,*

$$E\,\mathbf{P}(s = s_m|X) \leq \frac{q_m}{q_0} \exp\left(-cn^{\frac{1}{s_0+s_m+1}}\right) \qquad (4.3.4)$$

As a consequence we have

$$\mathbf{P}(s < s_0|X) \to 0. \qquad (4.3.5)$$

in probability.

Moreover they prove the following theorem, which states the convergence of the posterior probability at the minimax rate of convergence:

Theorem 4.3.3. *For any sequence $M_n \to \infty$ and for a rate of convergence $r_n(s_0) = n^{\frac{s_0}{2s_0+1}}$ the posterior probability*

$$\mathbf{P}(\theta : \, r_n(s_0)||\theta - \theta_0|| > M_n \,|X) \to 0 \tag{4.3.6}$$

in probability as n goes to infinity.

For sake of completeness we recall briefly the sketch of the proof in the appendix.

Such results are central in our work since we will use the same ideas in all this section to extend them to the wavelet model (4.3.2).
Set a positive constant M. Consider a collection of smoothness indexes $\mathcal{S}$, and to every $s \in \mathcal{S}$ we associate the model $\Theta(s) = \{\theta = (\theta_{jk})_{j,k}, \sum_j 2^{2js} \sum_k \theta_{jk}^2 \leq M\}$. Using the definition of s_0, this smoothness index can be characterized among other candidates in a discrete set $\mathcal{S}$ by:

$$s_0 = \arg\max_{(s \in \mathcal{S})} \left(\sum_j 2^{2js} \sum_k \theta_{jk}^2 \leq M \right).$$

We choose for prior on $\Theta(s)$ the same prior we had chosen on $\Theta(s_0)$ but replacing s_0 by s. On $\mathcal{S}$ consider a prior q such that $q(s) > 0$, $\forall s \in \mathcal{S}$. We prove the following lemma, which is equivalent to the selection lemma in the previous paper [BG00].

Lemma 4.3.4. *There exist integer N and a constant $c > 0$ such that, for any $m < 0$ and $n > N$*

$$E_{\theta_0}\mathbf{P}(s = s_m|X) \leq \frac{q_m}{q} \exp\left(-cn^{\frac{2}{s_0+s_m}}\right) \tag{4.3.7}$$

and therefore

$$\mathbf{P}(s < s_0|X) \to 0 \tag{4.3.8}$$

in $\mathbf{P}_{\theta_0}$−probability as $n \to \infty$.

This lemma shows that, also in the wavelet case, the posterior probability of mischoosing the model in favor of a less regular one tends to zero.
Given this selection lemma, the following theorem proves the convergence of the posterior distribution at the rate of convergence $n^{-\frac{2s_0}{2s_0+1}}$.

Theorem 4.3.5. *For any sequence $M_n \to \infty$, the posterior probability*

$$\mathbf{P}\left((\theta_{jk})_{jk}, \, n^{\frac{s_0}{2s_0+1}}||\theta - \theta_0|| > M_n|X\right) \to 0 \tag{4.3.9}$$

in $\mathbf{P}_{\theta_0}$-probability as n goes to infinity.

We have proven that the posterior mass is concentrated around the ball centered in θ_0 at the minimax rate of convergence, but the Bayesian point of view does not provide an estimator that can be easily constructed. E. Belitser and S. Ghosal have proven that there exists an estimator based on the posterior distribution which is adaptive but they do not construct it on a practical point of view. Yet A. van der Vaart in [GvdV00] tackles that problem by choosing a maximizer of the posterior probability of a ball with a well chosen radius but the estimator found does not correspond to the ones usually used for such estimation problems. That is the reason why we have investigated in Section 4.2 the choice of the posterior Bayes estimator and given its adaptive properties.

Such results could be extended to this model but, recently, some work has been conducted over Bayesian estimation with wavelet coefficients and some authors have looked at prior which describe, more precisely, the sparsity properties of wavelet bases. Write $\theta = \sum_{jk} w_{jk}\psi_{jk}$. Few coefficients can represent a large amount of information and the number of non zero coefficients tend towards zero when the resolution level increases. This prior information over the model can be incorporated into the coefficients's prior law to model such behavior. Abramovich, Sapatinas and Silverman in [ASS98] consider the following model

$$\begin{aligned} w_{jk} &\sim (1-\pi_j)\delta_0 + \pi_j \mathcal{N}(0,\tau_j^2) \\ \tau_j^2 &= c_1 2^{-\alpha j} \\ \pi_j &= \min(1, c_2 2^{-\beta j}) \end{aligned}$$

This model takes into account that with a probability $1-\pi_j \to 1$ when $j \to \infty$ large coefficients are less and less numerous. The constants are taken positive and can be chosen to maximize the log-likelihood. The distribution function of the posterior law is determined by

$$F(w_{jk}|d_{jk}) = \frac{1}{1+\eta_{jk}}\Phi\left(\frac{w_{jk} - d_{jk}\frac{\tau_j^2}{\sigma^2+\tau_j^2}}{\sigma\tau_j/\sqrt{\sigma^2+\tau_j^2}}\right) + \frac{\eta_{jk}}{1+\eta_{jk}}1_{w_{jk}\geq 0}$$

with

$$\eta_{jk} = \frac{1-\pi_j}{\pi_j}\frac{\sqrt{\sigma^2+\tau_j^2}}{\sigma}\exp\left(-\frac{\tau^2 d_{jk}^2}{2\sigma^2(\sigma^2+\tau_j^2)}\right),$$

where Φ is the standard Gaussian distribution function. Due to discontinuity in 0, solving $F(w_{jk}|d_{jk}) = \frac{1}{2}$, leads to posterior median:

- for $\eta_{jk} \geq 1$, the median is equal to 0.
- the median still equals 0 under the following conditions:

$$\begin{aligned} &\eta_{jk} < 1 \\ \frac{1}{2}(1-\eta_{jk}) \leq &\Phi\left(-\frac{d_{jk}\tau_j}{\sigma\sqrt{\sigma^2+\tau_j^2}}\right) \leq \frac{1}{2}(1+\eta_{jk}). \end{aligned}$$

So we have

$$\mathrm{Med}(w_{jk}|d_{jk}) = \mathrm{sign}(d_{jk}) \max(0, \xi_{jk})$$

for

$$\xi_{jk} = \frac{\tau_j^2}{\sigma^2 + \tau_j^2}|d_{jk}| - \frac{\tau_j \sigma}{\sqrt{\sigma^2 + \tau_j^2}} \Phi^{-1} \left(\frac{1 + \min(\eta_{jk}, 1)}{2} \right).$$

Such estimator is a Bayes thresholded estimator. Indeed ξ_{jk} is negative for $d_{jk} \in [-\lambda_j, \lambda_j]$ so the posterior median is set to zero when the observed coefficient is below a fixed level.
This model is the limit model of the one proposed by Chipman in [CW99] or Ruggeri and Vidakovic in [RV99]. As a matter of fact, these authors consider the following prior:

$$\begin{aligned} w_{jk}|\gamma_{jk} &\sim \gamma_{jk} N(0, c_j^2 \tau_j^2) + (1 - \gamma_{jk}) N(0, \tau_j^2) \\ \gamma_{jk} &\sim \mathrm{Bern}(p_j) \\ c_j^2 &>> 1. \end{aligned}$$

It is composed of a mixture of two Gaussian variables, one concentrated around zero stands for small coefficients while the other is wide spread and represents large coefficients which are vanishing with probability p_j. The estimator studied is the posterior mean:

$$w_{jk}|d_{jk}, \gamma_{jk} = 1 \sim N(\frac{c_j^2 \tau_j^2}{\sigma^2 + c_j^2 \tau_j^2} d_{jk}, \frac{\sigma^2 \tau_j^2 c_j^2}{\sigma^2 + c_j^2 \tau_j^2})$$

$$w_{jk}|d_{jk}, \gamma_{jk} = 0 \sim N(\frac{\tau_j^2}{\sigma^2 + \tau_j^2} d_{jk}, \frac{\sigma^2 \tau_j^2}{\sigma^2 + \tau_j^2}).$$

So it implies that

$$\begin{aligned} E(w_{jk}|d_{jk}) &= E_{\gamma_{jk}|\hat{w}_{jk}} E(w_{jk}|d_{jk}, \gamma_{jk}) \\ &= \mathbf{P}(\gamma_{jk} = 1|d_{jk}) \frac{c_j^2 \tau_j^2}{\sigma^2 + c_j^2 \tau_j^2} d_{jk} \\ &+ \mathbf{P}(\gamma_{jk} = 0|d_{jk}) \frac{\tau_j^2}{\sigma^2 + \tau_j^2} d_{jk}. \end{aligned}$$

So the estimator is a smoothed estimator. The constants of the law can also be chosen to maximize the log-likelihood. But, due to the structure of the prior, EM algorithm, described by Dempster, Laird an Rubin in [DLR77] can be used. Clyde and Geoges in [CPV98] or Neal and Hinton in [KCGN98] have used such algorithm. MCMC algorithm can also be used, as it is suggested by Müller and Vidakovic in [MV99]. Such methods lead to good numerical results. The asymptotic rate of convergence has only been proved very recently by I. Johnstone and B. Silverman in [JS01]. As a consequence, their result takes into account the properties of wavelets bases when representing functions with Besov-type regularity.

To extend the scope of this method, we should focus on either very regular functions, see Section 4.4, or very irregular functions such as the multifractal functions. We give some hints in the last chapter, Chapter V but most of the work is still in progress.

4.4 Super smooth functions

We consider in this part a class of functions, analytic functions, which has been originally studied by B. Levit in [GL96] and we will try to adapt previous result to this estimation problem. Consider a parameter $\gamma > 0$, we make the assumption that there exists a basis such that the functions $\mathbf{f}$ can be written $\mathbf{f} = \sum_j \theta_j \psi_j$ and such that the coefficients satisfy $\theta = (\theta_i)_{i \geq 1}$ belongs to the space

$$\Theta = \{(\theta_i)_i, \sum_{i=1}^{\infty} e^{\gamma i} \theta_i^2 \leq M\} \tag{4.4.1}$$

So define the true γ as the largest real in a collection of indexes Γ such that the preceding sum is convergent.

$$\gamma = \arg\max_{t \in \Gamma} \{t, \sum_{i=1}^{\infty} e^{ti} \theta_i^2 \leq M\}$$

This model appears in mixture problem or in exponential family problems. Our aim is still to estimate the infinite dimensional parameter of interest $\theta = (\theta_i)_{i \geq 1}$ using the Maximum a posteriori estimator.
Define the prior $\theta_i \sim \mathcal{N}(0, \tau_i^2)$ where we have set $\tau_i^2 = \lambda_n^2 e^{-\gamma i}$. Following the guidelines of the preceding work, we find

$$\forall i \geq 1,\ \hat{\theta}_i = \frac{1}{1 + \frac{\sigma^2}{\lambda_n^2 n} e^{\gamma i}} X_i \tag{4.4.2}$$

This estimator is convergent and the following theorem gives its asymptotic behavior.

Theorem 4.4.1. *In the exponential model hypothesis* (4.4.1)*, the estimator corresponding to the maximum of the posterior law is convergent at the following rate of convergence: there exist a constant $C < \infty$ so that $\forall \epsilon > 0$ we have*

$$E||\hat{\theta} - \theta||_2^2 \leq C \frac{\log(\gamma n)}{n}.$$

We recall that in that model we have: to every $\gamma \in \Gamma$, we associate a corresponding model

$$\Theta(\gamma) = \{\theta \in l^2, \sum_{i=1}^{n} e^{\gamma i} \theta_i^2 \leq M\}.$$

γ is considered as an hyperparameter of the model. In the non adaptive case where γ_0, the true value of the regularity coefficient is known, the maximum a posteriori estimator is defined by

$$\forall i,\ \hat{\theta}_i = \frac{1}{\frac{1}{n} + \lambda_n^2 e^{\gamma i}} X_i.$$

It converges at a rate of convergence as close as possible to the parametric rate of convergence. Now consider $q(\gamma)$, $\gamma \in \Gamma$ a prior distribution over $\Gamma = \{\gamma_m, m \in \mathbb{Z}\}$. This time, the selection lemma does not hold for this model and a similar method does not apply. E. Belitser and B. Levitt in [BB01] have solved this issue by using an empirical Bayes method based on a marginal of the joint distribution.

4.5 Appendix

Proof of Theorem 4.3.3

Proof.

$$\begin{aligned}\mathbf{P}(r_n(s_0)||\hat{\theta}(s)-\theta|| > M_n|X) = & \mathbf{P}(r_n(s_0)||\hat{\theta}(s)-\theta|| > M_n, s > s_0|X) \quad (I)\\ & + \mathbf{P}(r_n(s_0)||\hat{\theta}(s)-\theta|| > M_n, s = s_0|X) \quad (II)\\ & + \mathbf{P}(r_n(s_0)||\hat{\theta}(s)-\theta|| > M_n, s < s_0|X) \quad (III)\end{aligned}$$

for $r_n(s_0) = n^{\frac{s_0}{2s_0+1}}$ the optimal rate of convergence and M_n a sequence, growing to infinity. The second term (II) can be easily handled using the results of the preceding subsection, proving the convergence of the estimator at the minimax rate of convergence $r_n(s_0)$, and a Chebychev's inequality:

$$\begin{aligned}E\mathbf{P}(r_n||\hat{\theta}(s)-\theta|| > M_n, s = s_0|X) &\leq \frac{E(||\hat{\theta}(s_0)-\theta||^2)}{r_n^2(s_0)M_n^2}\\ &\leq r_n^2(s_0)\frac{1}{M_n^2 r_n^2(s_0)}\\ &\leq \frac{1}{M_n^2}\\ &\leq \epsilon\end{aligned}$$

for any ϵ as soon as M_n is large enough.
The last term (III) goes to zero due to a lemma in [BG00] called the selection lemma where they prove that the smoothness of the model is never underestimated, i.e

$$\mathbf{P}(s < s_0|X) \to 0.$$

The first term (I) is proved to go to zero using a theorem proved by van der Vaart in [GvdV00] and recalled below, depending on the choice of the prior on $\mathcal{S}$ which defines the space Θ.

Theorem 4.5.1. *Suppose we have i.i.d observations $Y_1, Y_2, \ldots$ from a distribution $\mathbf{P}$ having a density $p \in \mathbf{P}$.Let H the Hellinger distance and Π a prior on $\mathcal{P}$ and $\bar{\mathcal{P}} \subset \mathcal{P}$. Let $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$. Let $D(\epsilon, \mathcal{S}, d)$ the largest m such that there exist $s_1, \ldots, s_m \in \mathcal{S}$ with $d(s_j, s_k) > \epsilon$ for $j \neq k$. Then if we make the assumptions that*

$$\log D(\epsilon_n, \bar{\mathcal{P}}, H) \leq c_1 n\epsilon_n^2$$

and also that

$$\Pi\left(P, -\int \log\frac{p}{p_0}dP_0 \leq \epsilon_n^2, \int(\log\frac{p}{p_0})^2 dP_0 \leq \epsilon_n^2\right) \geq c_2 e^{-c_3 n\epsilon_n^2}$$

we have for a large M

$$\Pi\left(P \in \bar{\mathcal{P}}, H(P_0, P) \geq M\epsilon_n|Y_1, \ldots, Y_n\right) \to 0$$

in P_0 probability.

The first condition is an entropy condition over the class $\bar{\mathcal{P}}$ since if we consider $N(\epsilon, \mathcal{S}, d)$ the minimal number of balls for an ϵ-covering of the set $\mathcal{S}$ we have:

$$N(\epsilon, \mathcal{S}, d) \leq D(\epsilon, \mathcal{S}, d) \leq N(\epsilon/2, \mathcal{S}, d).$$

The second term means that the prior distribution puts enough mass near $\mathbf{P}_0$ where "enough" is measured in terms of Küllback-Leibner distance and quadratic norm of the log-likelihood. Belitser and Ghosal use this theorem with $\pi = \sum_{m>0} q_m \pi_m$, $\mathcal{P} = \{\theta \in l^2\}$, and $\bar{\mathcal{P}} = \{\theta \in l^2, \sum_{i=1}^n i^{2s_0}\theta_i^2 < \infty\}$, and conclude the proof by the equivalence of Hellinger distance and l^2 norm for allocation models, proved by Ghosal and Belitser in [BG00]. □

Proof of Lemma 4.3.4

Proof. The proof follows the guidelines of the proof by Belitser and Ghosal in [BG00]. By the martingale convergence theorem in [Wil91] we have

$$\mathbf{P}(s = s_m|X) = \lim_{i\to\infty} \mathbf{P}(s = s_m|X_1, \dots, X_i) \text{ a.s } \mathbf{P}_\theta.$$

So, if we assume that $\mathbf{P}_\theta$ and $\mathbf{P}_{\theta_0}$ are mutually absolutely continuous, we can calculate the posterior probability. Such an assumption can be proven by probabilist arguments as in [BG00]. By Bayes theorem we have:

$$\begin{aligned}
&\mathbf{P}(s = s_m|X_1, \dots, X_i) \\
&= \frac{q_m \prod_{j=1}^i \prod_{k=0}^{2^j-1} (\frac{\sigma^2}{n} + 2^{-2js_m})^{-1/2} \exp(-\frac{1}{2}(\frac{\sigma^2}{n} + 2^{-2js_m})^{-1} X_{jk}^2)}{\sum_{m=-\infty}^{\infty} q_m \prod_{j=1}^i \prod_{k=0}^{2^j-1} (\frac{\sigma^2}{n} + 2^{-2js_m})^{-1/2} \exp(-\frac{1}{2}(\frac{\sigma^2}{n} + 2^{-2js_m})^{-1} X_{jk}^2)} \\
&\leq \frac{q_m}{q_0} \frac{\prod_{j=1}^i (\frac{\sigma^2}{n} + 2^{-2js_m})^{-2^{j-1}} \exp(-\frac{1}{2}(\frac{\sigma^2}{n} + 2^{-2js_m})^{-1} \sum_{k=0}^{2^j-1} X_{jk}^2)}{\prod_{i=1}^j (\frac{\sigma^2}{n} + 2^{-2js_0})^{-2^{j-1}} \exp(-\frac{1}{2}(\frac{\sigma^2}{n} + 2^{-2js_0})^{-1} \sum_{k=0}^{2^j-1} X_{jk}^2)}
\end{aligned}$$

For reasons of simplicity, we consider $\sigma^2 = 1$. Set

$$a_j = (\frac{\sigma^2}{n} + 2^{-2js_m})^{-1} - (\frac{\sigma^2}{n} + 2^{-2js_0})^{-1}.$$

Since the random variables X_{jk} are independent, and using the following majoration: if X is a random variable distributed as $\mathcal{N}(\mu, \sigma^2)$, then for every $\alpha > -\sigma^2$

$$E\left(\exp(-\frac{\alpha}{2}X^2)\right) = \frac{1}{\sqrt{1 + \alpha\sigma^2}} \exp\left(-\frac{\mu^2\alpha}{2(1 + \alpha\sigma^2)}\right)$$

we have the following inequality

$$\begin{aligned}
&E_{\theta_0}\mathbf{P}(s = s_m|X) \\
&= \lim_{i\to\infty} E_{\theta_0}\mathbf{P}(s = s_m|X_1,\ldots,X_i) \\
&\leq \frac{q_m}{q_0}\prod_{j=1}^{\infty}\sqrt{\left(\frac{\frac{\sigma^2}{n}+2^{-2js_0}}{\frac{\sigma^2}{n}+2^{-2js_m}}\right)^{-2^j}}(1+\frac{a_j}{n})^{-2^{j-1}}\exp\left(-\frac{a_i\sum_{k=0}^{2^j-1}\theta_{jk}^2}{2(1+\frac{a_j}{n})}\right) \\
&\leq \frac{q_m}{q_0}\prod_{j=1}^{\infty}\left(\left[1+\frac{a_j 2^{-2js_0}}{1+\frac{a_j}{n}}\right]^{2^{j-1}}\exp\left(-\frac{a_j\sum_{k=0}^{2^j-1}\theta_{jk}^2}{2(1+\frac{a_j}{n})}\right)\right) \\
&\leq \frac{q_m}{q_0}\exp\left(\frac{1}{2}\sum_{j=1}^{\infty}\frac{a_j}{1+\frac{a_j}{n}}\sum_{k=0}^{2^j-1}(2^{-2js_0}-\theta_{jk}^2)\right) \\
&\leq \frac{q_m}{q_0}\exp\left(\frac{1}{2}\sum_{j=1}^{\infty}\frac{2^{-2js_0}-2^{-2js_m}}{2^{-j(s_0+s_m)}+2n^{-1}2^{-2js_0}+n^{-2}}\sum_{k=0}^{2^j-1}(2^{-2js_0}-\theta_{jk}^2)\right)
\end{aligned}$$

We have used that $1+x \leq \exp(x)$. We notice that the same inequality holds for a generic l.

$$E_{\theta_0}\mathbf{P}(s = s_m|X) \leq \frac{q_m}{q_l}\exp\left(\frac{1}{2}\sum_{j=1}^{\infty}\frac{2^{-2js_l}-2^{-2js_m}}{2^{-2j(s_l+s_m)}+2n^{-1}2^{-2js_l}+n^{-2}}\sum_{k=0}^{2^j-1}(2^{-2js_l}-\theta_{jk}^2)\right)$$

Now divide this sum into two terms and define

$$S_1 = \sum_{j=1}^{\infty}\frac{2^{-2js_0}-2^{-2js_m}}{2^{-2j(s_0+s_m)}+2n^{-1}2^{-2js_0}+n^{-2}}\sum_{k=0}^{2^j-1}2^{-2js_0}, \tag{4.5.1}$$

$$S_2 = -\sum_{j=1}^{\infty}\frac{2^{-2js_0}-2^{-2js_m}}{2^{-2j(s_0+s_m)}+2n^{-1}2^{-2js_0}+n^{-2}}\sum_{k=0}^{2^j-1}\theta_{jk}^2. \tag{4.5.2}$$

Remark that:

- $2^{-2js_0} - 2^{-2js_m} < 0$
- as soon as $j \geq \frac{1}{s_0 - s_{-1}}$ we have

$$2^{-2js_0} - 2^{-2js_m} \leq \frac{-1}{2}2^{-2js_m}$$

- $\forall j \leq n^{\frac{2}{s_0+s_m}}$ the first term in the denominator of the common fraction is dominant, so

$$\begin{aligned}
&2^{-2j(s_0+s_m)} \geq 2n^{-1}2^{-2js_0} \\
&2^{-2j(s_0+s_m)} \geq n^{-2}.
\end{aligned}$$

Using these three inequalities, we obtain that:

$$S_1 \leq -\frac{1}{12} n^{\frac{2}{s_0+s_m}}.$$

$$\begin{aligned} S_2 &= \sum_{j\geq 1} \frac{2^{-2js_m}}{2^{-2j(s_0+s_m)} + 2n^{-1}2^{-2js_0} + n^{-2}} \sum_k \theta_{jk}^2 \\ &\leq \sum_{j\leq j_1} 2^{2js_0} \sum_k \theta_{jk}^2 + \sum_{j>j_1} n^2 2^{-2js_m} \sum_k \theta_{jk}^2 \\ &\leq j_1 M + n^2 \sum_{j>j_1} 2^{-2j(s_0+s_m)} 2^{js_0} \sum_k \theta_{jk}^2 \\ &\leq j_1 M + n^2 \sum_{j>j_1} 2^{-2(k+1)(s_m+s_0)} 2^{2js_0} \sum_k \theta_{jk}^2 \\ &\leq \frac{1}{24} n^{\frac{2}{s_0+s_m}} \end{aligned}$$

for n large enough since the last sum is convergent. So with these two majorations the selection lemma is proved in the wavelet case. □

proof of theorem 4.3.5

Proof. The only thing that remains to be seen is the adaptation of van der Vaart's theorem (4.5.1) that has been recalled previously.
First of all, the entropy condition is satisfied since the entropy of a Besov body $B_{2\infty}^s$ has been calculated by Birgé and Massart in [BM00] and one has:

Lemma 4.5.2. *For a constant A depending on p and $s > 1/p - 1/2$,*

$$H(\delta, \mathcal{B}_{p,\infty}^s) \leq A\delta^{-\frac{1}{s}}, \quad \delta > 0. \tag{4.5.3}$$

Second, since the set

$$\{\theta,\ -E_{\theta_0}\left(\log \frac{d\mathbf{P}_\theta}{d\mathbf{P}_{\theta_0}}(Y)\right) \leq \epsilon^2, E_{\theta_0}\left(\log \frac{d\mathbf{P}_\theta}{d\mathbf{P}_{\theta_0}}(Y)\right)^2 \leq \epsilon^2\}$$

contains $\{||\theta - \theta_0||_2 \leq \epsilon\}$, we have to evaluate $\mathbf{P}_{\theta_0}(||\theta - \theta_0||^2 \leq \epsilon^2)$. For simplicity reasons, write $\mathbf{P}_0 = \mathbf{P}_{\theta_0}$.

$$\begin{aligned} &\mathbf{P}_0(||\theta - \theta_0||^2 \leq \epsilon^2) \\ \geq &\mathbf{P}_0(\sum_{j\leq J} \sum_k |\theta_{jk} - \theta_{jk}^0||^2)\, \mathbf{P}_0(\sum_{j>J} \sum_k |\theta_{jk} - \theta_{jk}^0||^2) \end{aligned}$$

On the one hand, under $\mathbf{P}_{\theta_0}$, the variables $\theta_{jk} - \theta^0_{jk} = W_{jk}$ are independent Gaussian variables following a law $\mathcal{N}\left(0, 2^{-2js}\right)$. We notice that

$$\begin{aligned}
&\mathbf{P}_0(\sum_{j\le J}\sum_k W^2_{jk} \le \epsilon^2)\\
&= \int_{\sum_{j\le J}\sum_k W^2_{jk}\le\epsilon^2} \prod_{j\le J,\, k}\left((2\pi)^{-1/2} 2^{2js_0/2}\exp(-\frac{1}{2}2^{2js}(w_{jk}+\theta^0_{jk})^2)dw_{jk}\right)\\
&\ge \prod_{j\le J,\, k}\left((2\pi)^{-1/2}2^{2js/2}\right)\int_{\sum_{j\le J}\sum_k W^2_{jk}\le\epsilon^2}\prod_{j\le J,\, k}\exp(-2^{2js}(w^2_{jk}+(\theta^0_{jk})^2))dw_{jk}\\
&\ge \prod_{j\le J,\, k}\left((2\pi)^{-1/2}2^{2js/2}\right)\int_{\sum_{j\le J}\sum_k W^2_{jk}\le\epsilon^2}\exp(-\sum_{j\le J}2^{2js}\sum_k(w^2_{jk}+(\theta^0_{jk})^2))dw_{jk}\\
&\ge \prod_{j\le J,\, k}\left((2\pi)^{-1/2}2^{2js/2}\right)\exp(-\sum_{j\le J}2^{2js}\sum_k(\theta^0_{jk})^2)\int_{\sum_{j\le J}\sum_k W^2_{jk}\le\epsilon^2}\exp(-\sum_{j\le J}2^{2js}\sum_k w^2_{jk})dw_{jk}\\
&\ge \prod_{j\le J,\, k}\left((\pi)^{-1/2}\right)\exp(-\sum_{j\le J}2^{2js}\sum_k(\theta^0_{jk})^2)\int_{\sum_{jk}2^{-2js-1}V^2_{jk}\le\epsilon^2}\exp(-\frac{1}{2}\sum_{j\le J,k}v^2_{jk})dv_{jk}\\
&\ge \pi^{-\frac{1}{2}\sum_{j=0}^J 2^j}\exp(-\sum_{j\le J}2^{2js}\sum_k(\theta^0_{jk})^2)\mathbf{P}(\sum_{j\le J,k}2^{-2js-1}V^2_{jk}\le\epsilon^2).
\end{aligned}$$

where V_{jk} are standard Gaussian random variables. Such result is similar to the majoration given by Shen and Wasserman in [XS98].
On the other hand we have,

$$\begin{aligned}
&\mathbf{P}_0(\sum_{j>J}\sum_k(\theta-\theta^0_{jk})^2\le\epsilon^2)\\
&\ge \mathbf{P}_0(\sum_{j>J}\sum_k 2(\theta^2_{jk}+(\theta^0_{jk})^2)\le\epsilon^2)\\
&\ge \mathbf{P}_0(\sum_{j>J}2\frac{2^{2js_0}}{2^{2Js_0}}\sum_k(\theta^0_{jk})^2+2\sum_{j>J}\sum_k\theta^2_{jk}\le\epsilon^2)\\
&\ge \mathbf{P}_0(2\sum_k\theta^2_{jk}+22^{-2Js_0}M\le\epsilon^2)\\
&\ge \mathbf{P}_0(2\sum_k\theta^2_{jk}\le\epsilon^2/2)
\end{aligned}$$

where J is chosen $J \ge J_1$ in order to have $2^{-2Js_0}M \le \epsilon^2/4$. Using Chebychev's inequality, there exists J_2 so that $\sum_{j>J_2,\, k}|\theta_{jk}|^2 \le \epsilon^2/4$ with probability

$$1 - 4/\epsilon^2 \sum_{j>2_1,\, k} E|\theta_{jk}|^2 \ge \frac{1}{2}.$$

So for J large enough, the two conditions of theorem 2.1 in [GvdV00] are fulfilled, so the theorem applies in the wavelet case which proves the convergence of the posterior probability in $\mathbf{P}_{\theta_0}$-probability. □

Proof of Theorem 4.4.1

Proof. We recall the form of the estimator: $\hat{\theta}_i = \frac{1}{1+\frac{\sigma^2}{\lambda_n^2 n}e^{\gamma i}} X_i$.

$$\begin{aligned} E\|\hat{\theta} - \theta\|_2^2 &= \sum_{i\geq 1} \left(\frac{\sigma_n^2}{\sigma_n^2 + \tau_i^2} \theta_i \right)^2 + \sum_{i \geq 1} \left(\frac{\tau_i^2}{\tau_i^2 + \sigma_n^2} \right)^2 \frac{\sigma^2}{n} \\ &= T_1 + T_2. \end{aligned}$$

As precedingly, we have the following majorations:

$$T_1 \leq \frac{M}{n\lambda_n^2}$$

where we have used the definition of the set Θ.
The second term can be compared with an integral. Using that, for all $\epsilon > 0$, there exist a fixed constant $A(\epsilon) = A$ and $M_1 < \infty$ such that

$$\forall x \geq A, \ \log(1+x) \leq Ax^\epsilon$$

we have:

$$\begin{aligned} T_2 &= \sum_{i \geq 1} \frac{\sigma^2}{n} \left(\frac{1}{1 + \sigma_n^2 \tau_i^{-2}} \right)^2 \\ &= \sum_{i\geq 1} \frac{\sigma^2/n}{(1 + \frac{\sigma^2 e^{\gamma i}}{n\lambda_n^2})^2} \\ &\leq \frac{\sigma^2}{\gamma} \frac{1}{n} \left(\frac{n\lambda_n^2}{\sigma^2} \right)^\epsilon \\ &\leq C_1 \frac{\lambda_n^{2\epsilon}}{n^{1-\epsilon}} \end{aligned}$$

for a constant C_1. We find the following majoration: there exists a finite constant C such that

$$E\|\hat{\theta} - \theta\|_2^2 \leq C(\frac{1}{n\lambda_n^2} + \frac{\lambda_n^{2\epsilon}}{n^{1-\epsilon}}).$$

If the smoothing parameter λ_n is choosen such that $\lambda_n^2 = n^{-\frac{\epsilon}{1+\epsilon}}$ we have the rate of convergence

$$E\|\hat{\theta} - \theta\|_2^2 \leq Cn^{-\frac{\epsilon}{1+\epsilon}}.$$

If we go into the details of the previous calculations and keep the logarithmic term, we can find the optimal trade-off parameter which gives the optimal rate of convergence concluding the proof.:

$$\begin{aligned}
E||\hat{\theta}-\theta||_2^2 &\leq T_1+T_2 \\
&\leq \frac{M}{n\lambda_n^2}+\frac{\sigma^2}{n\gamma}\left(\log(1+\frac{n\lambda_n^2}{\sigma^2})-\frac{1}{1+\frac{\sigma^2}{n\lambda_n^2}}\right) \\
&\leq \frac{M}{n\lambda_n^2}+\frac{\sigma^2}{n\gamma}\log(1+\frac{n\lambda_n^2}{\sigma^2}) \\
&\leq \frac{M}{n\lambda_n^2}+c\frac{\sigma^2}{n\gamma}\log(\frac{n\lambda_n^2}{\sigma^2}) \\
&\leq A\frac{\sigma^2}{\gamma}\frac{\log(\frac{n\gamma/c}{\sigma^2})}{n}
\end{aligned}$$

for $n \geq N$ and a finite positive constant A. □

Chapter 5

Multifractal Estimation and Bayes M-estimation

5.1 Wavelet estimation of a multifractal function

Abstract We prove that multifractal functions, characterized by their wavelet representation can be estimated in the white noise model by a Bayesian estimation method. We give rates of convergence for two different models. Finally, we estimate the parameters of the model using EM-algorithm.

AMS 1991 subject classifications. Primary: 60G17, Secondary:62G07
Key words and phrases. Multifractal analysis, Bayesian statistics, Wavelet Bases

5.1.1 Introduction

In the last decade many emphasis have been placed on non parametric estimation by wavelet methods. The reasons of the success of wavelets in non parametric statistics are mainly twofold. First, wavelet basis are unconditional basis of at most all usual function space [Mey87]. Further, estimates built on wavelets are easy to compute [Mal89] and are asymptotically optimal [DJKP95], [HKPT98], [DJ96a].

In this paper, we will focus on wavelet estimates of very irregular functions namely multifractal functions. Roughly speaking, a multifractal function is a function whose Hölder local regularity index does not range in a singleton. That means that the function may be very regular in some areas while it is very irregular in others. Such function with rapid changes of regularity have been first introduced to modelize physical phenomena as turbulence [BAF$^+$91], or net events as the road or data traffic [RCRB99]. A way to study these functions is the multifractal analysis first introduced in [FP85]. This analysis is concerned with the repartition of points having a given regularity.

Jaffard et al [ABJM99],[Jaf00b],[Jaf00a], [AJ01] or Roueff in [Rou01] have recently shown that lacunary random series built on wavelets have multifractal properties. In others words, using wavelets, they built a random process having trajectories in a multifractal set of functions. That is a probability measure $\mathcal{P}$ on this set. We will consider here an unknown function f^* lying in the support of $\mathcal{P}$. In the frame of the Gaussian white noise model built on f^*, we perform and study Bayesian procedures with prior $\mathcal{P}$. We show that the Bayesian estimate converges in L^2 in mean and give the rate of convergence using a model which differs from the one recently studied by Sapatinas et al [ASS98] and Johnstone et al [JS01].

The paper is organized as follows. In the next Section we first recall some topics on multifractal analysis. Then, following Jaffard et al we present some lacunary random wavelet series having multifractal properties. Section 5.1.3 is devoted to the study of Bayesian estimates built with the priors of Section 5.1.2. Section 4 is devoted to numerical simulations. In Section 5.1.5, we study the statistical multifractal models in a parametric frame.

5.1.2 Multifractal wavelet models

Multifractal formalism

In this paper, we will always work with function on $[0,1]$. To begin with, let first introduce some useful definitions around multifractal functions.

Definition 5.1.1. *Let f be a function on $[0,1]$.*

1) Let $x_0 \in [0,1]$ and $h \geq 0$, the set $C^h(x_0)$ is the set of all functions f on $[0,1]$ such that there exist a polynomial P of degree less than h and a neighborhood V of x_0 satisfying

$$|f(x) - P(x)| = O\left(x - x_0\right)^h \quad (x \in V).$$

2) Let $h_f(x_0) = \sup\{h \geq 0, f \in C^h(x_0)\}$ and

$$E_h = \{x \in \mathbb{R},\ h_f(x) = h\} \quad (h \geq 0).$$

The spectrum of singularity d_f of f is the function on $\mathbb{R}^+$ which associates to each $h \geq 0$ the Haussdorf dimension of the set E_h.

Multifractal analysis of a function was first introduced in a physical frame in [FP85]. Given a function f, one of the main goal of this analysis is the computation of the spectrum of singularities d_f. When d_f does not vanish in at least two points we say that f is multifractal. The spectrum of singularities of a function f is a relevant quantity to describe the smoothness variation of f. Multifractal functions can be constructed using their decomposition onto an appropriate wavelet basis as it is done by Arnéodo, Bacry and Jaffard in [ABJM99] or in [Jaf00a] and [Jaf00b]. Since we pay attention on function on $[0,1]$, we will consider periodized wavelets described by Daubechies in [Dau92]. Following Daubechies we define $\tilde{\psi}$ a wavelet in the Schwartz class with enough vanishing moments, and construct the periodized wavelet

$$\psi(x) = \sum_{l \in \mathbb{Z}} \tilde{\psi}(x - l).$$

The functions $\psi_{jk} = \psi(2^j x - l), \quad \forall j \in \mathbb{N},\ k \in [0, 2^j - 1]$ are obtained from the first wavelet by dilatation and translation. Then $(2^{j/2}\psi_{jk})_{(j,k)}$ provides an orthonormal basis of the Hilbert space $L^2([0,1])$ (observe the presence of a normalizing factor $2^{j/2}$). Let $f \in L^2([0,1])$ on one hand, its wavelet coefficients may be computed as

$$w_{jk} = 2^j \int_0^1 f(t)\psi_{jk}(t)dt \quad (j \in \mathbb{N},\ k \in [0, 2^j - 1]).$$

On the other hand, f may be reconstructed using its wavelet coefficients

$$f = \sum_j \sum_{k=0}^{2^j - 1} w_{jk}\psi_{jk} \tag{5.1.1}$$

Using this wavelet representation, we now turn on the construction of random functions exhibiting multifractal properties. This will be done considering sparse random wavelet series. First let define some functions quantifying sparsity that will be useful in our study. Let $(w_{j,k})_{j\in\mathbb{N},\, k\in[0,2^j-1]}$ be any sequence of real numbers, we set for a positive real number α:

$$N_j(\alpha) = \#\{k,\ |w_{jk}| \geq 2^{-\alpha j}\} \tag{5.1.2}$$

$$\rho(\alpha) = \inf_{\epsilon>0} \limsup_{j\to\infty} \frac{\log_2(N_j(\alpha+\epsilon) - N_j(\alpha-\epsilon))}{j}, \tag{5.1.3}$$

where $\log_2$ is the base 2 logarithm (hereafter, log will denote the natural logarithm). Roughly speaking, for large j there are about $2^{\rho(\alpha)j}$ coefficients $(w_{j,k})_{j\in\mathbb{N}}$ of size of order $2^{-\alpha j}$.
The random functions used in this paper are obtained as follows. Let $\rho_j,\ j\in\mathbb{N}$ be a repartition function on $\mathbb{R}$. Further, let $(X_{jk})_{k=1,\ldots,2^j}$ be 2^j independent random variables having common distribution ρ_j. Now, build a random function F using the reconstruction formula (5.1.1) with for any $j\in\mathbb{N}, k=1,\ldots,2^j$ $|w_{jk}| = 2^{-jX_{jk}}$. To study the multifractal properties of the random function F, Aubry and Jaffard [AJ01], [Jaf00b] introduced the following functions:

$$\begin{aligned}\tilde{\rho}(\alpha,\epsilon) &= \limsup_{j\to\infty} \frac{\log_2(2^j\rho_j[\alpha-\epsilon,\alpha+\epsilon])}{j}\\ &= 1 + \limsup_{j\to\infty} \frac{\log \mathbf{P}(X_{jk}\in[\alpha-\epsilon,\alpha+\epsilon])}{j}\\ \tilde{\rho}(\alpha) &= \inf_{\epsilon>0} \tilde{\rho}(\alpha,\epsilon)\end{aligned}$$

One of the main assumption on $(\rho_j)_{j\in\mathbb{N}}$ in [Jaf00b], [AJ01] is that the support of the wavelet coefficient distribution is compact. Under some additional hypothesis, Jaffard et al prove that the two quantities ρ and $\tilde{\rho}$ are equal and that the spectrum of singularity of F can be calculated. Indeed we have, for all $h>0$

$$d_F(h) = h \sup_{\alpha\in(0,h]} \frac{\rho(\alpha)}{\alpha} \text{ (a.s.)}. \tag{5.1.4}$$

More recently, this result has been extended for Gaussian coefficients by Aubry and Jaffard in [AJ01].

Multifractal Model

Let $\mathcal{P}$ be the probability distribution on the Borel measurable space $L^2([0,1])$ induced by the previous random serie. In this paper, we will made Bayesian inference with $\mathcal{P}$ on various simple functional statistical models. These models will satisfy assumptions warranting that (5.1.4) holds. These simple multifractal models will be characterized by two parameters η and α in $[0,1]$. On one hand η will describe the lacunarity of the wavelet series (that is its sparsity). On the other hand the coefficient α will be inversely proportional to the intensity of the value of the wavelet coefficients. These parameters will completely characterize the spectrum of singularity of the random functions involved. We now introduce and discuss these models.

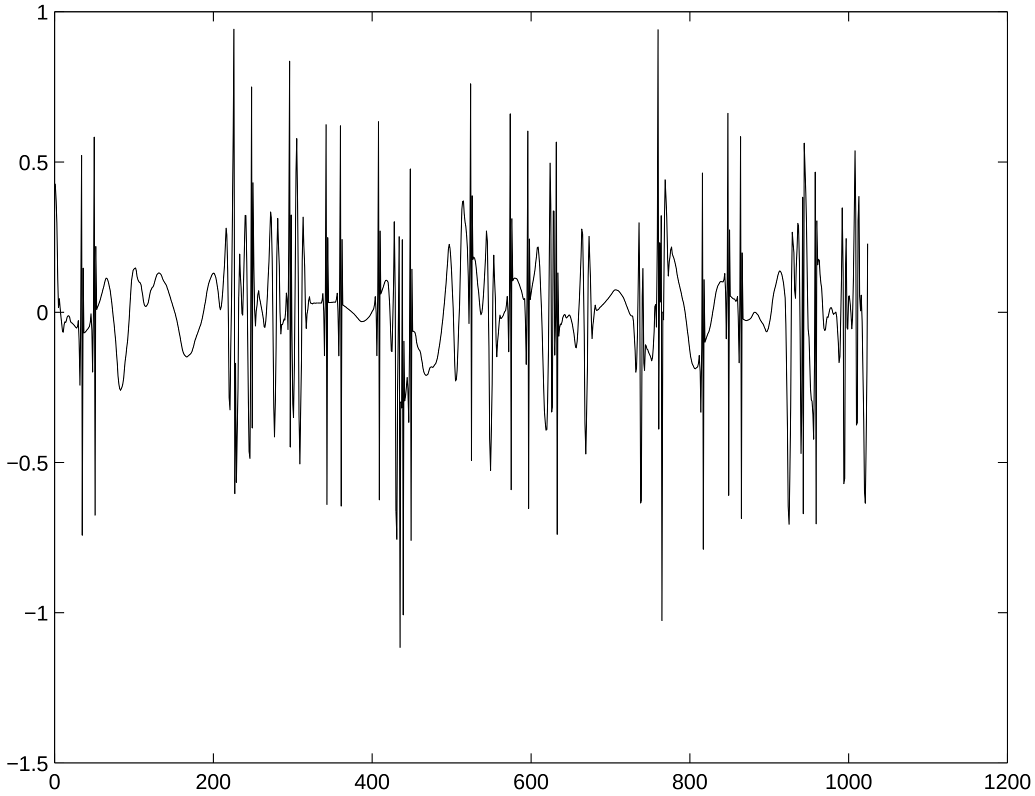


Figure 5.1: Multifractal process

Bernoulli constrained model The first simplest model is an exact representation of the structure of the multifractal processes described in term of wavelet series by S. Jaffard in [Jaf00b]. At each resolution level j, pick at random $[2^{\eta j}]$ locations among the 2^j wavelet coefficients, and put these coefficients to the value $2^{-\alpha j}$ while the $2^j - [2^{\eta j}]$ are setted to zero. This choice of coefficients is made independently between each level. Generating a function with this method may seem too restrictive. However, such processes appear naturally when studying multifractal processes and their spectrum of singularity can be described using parameters α and η. As a matter of fact, using (5.1.4), the assumptions over the wavelet coefficients lead to the following spectrum of singularities.

$$\forall h \in [\alpha, \frac{\alpha}{1-\eta}],\ d_f(h) = \frac{1-\eta}{\alpha} h. \tag{5.1.5}$$

The Bernoulli constrained model enables us to modelize functions with a linear spectrum of singularity.

In figure 5.1 we plot a realization of a multifractal function of the Bernoulli constrained model. The lacunarity coefficient is $\eta = 0.4$ while $\alpha = 0.3$.

Gaussian extension to Bernoulli constrained model The second model we consider is an extension of the previous one. It allows more flexibility in the choice of the wavelet coefficients: in the first description, they could only take two values: either $2^{-\alpha j}$ or 0. Here, we allow non zero coefficients to take values different from $2^{-\alpha j}$ but still close to that value. Hence, we consider that these coefficients are distributed following a Gaussian random variable centered in $2^{-\alpha j}$ with variance $\Delta_j^2 > 0$ $\left(\mathcal{N}\left(2^{-\alpha j}, \Delta_j^2\right)\right)$. The other coefficients are still equal to zero. Such model is a generalization of the first rough model. It is described in Aubry and Jaffard [AJ01].

Towards a mixture model Further we may consider the more general frame where at each level j the wavelet coefficients form an i.i.d. sample drawn from of simple mixture distribution. Where this mixture depends only on j and on two parameters: the lacunarity parameter η and the intensity parameter α. Assuming that these parameters are unknown, their estimation is an interesting statistical problem. This will be considered in the forthcoming paper [GL01]. One simple case is the Gaussian model considered in the previous subsection where at level j the mixture distribution is

$$2^{(\eta-1)j}\mathcal{N}\left(2^{-\alpha j}, \sigma_n^2\right) + (1 - 2^{(\eta-1)j})\mathcal{N}\left(0, \sigma_n^2\right).$$

In Section 5.1.5 we will describe a block likelihood method for estimating the parameters α and η and the corresponding E.M.-algorithm.
The two last models provide functions whose spectrum of singularity has the same expression as in the Bernoulli constrained model in (5.1.5).

Remark 5.1.2. *The lacunar random wavelet serie $f^* = \sum_{jk} w_{jk}^* \psi_{jk}$, where the wavelet coefficients are drawn according to the previous statistical model, is such that, for a positive p, there exists a finite positive constant C:*

$$\sum_{k=0}^{2^j-1} \mathbf{E}|w_{jk}^*|^p \leq C 2^{(-\alpha p+\eta)j}.$$

This emplies that the function f^* belongs a.s to the sparse Besov spaces $B_{p\infty}^s$ for $s \leq \alpha + \frac{1-\eta}{p}$. General references about Besov spaces are Besov, Ill'in and Nikol'skii [BIN78], Triebel [ET92] or DeVore and Lorentz [DL93].

5.1.3 Bayesian estimation

Assuming that a multifractal function f^* is drawn from the Bernoulli constained model (or its extension), our aim is to estimate this function when it is observed in the white noise model. This will be performed using the maximum a posteriori estimator. In the white noise model, we observe the wavelet coefficients w_{jk}^* of the function f^* together with a Gaussian white noise ϵ_{jk} with variance $\frac{\sigma^2}{n}$ where n is the number of observations. We assume that the observations are dyadic and $n = 2^{j_1}$, $(j_1 > 0)$. Recall that the wavelet coefficients are obtained from discrete regression model $Y_i = f^*(i/2^{j_1}) + W_i$, $i = 1, \ldots, n$ by performing the

Discrete Wavelet Transform (DWT). Such transform is performed by Mallat's fast algorithm [Mal89] that requires only $O(n)$ operations. Hence, the observations are

$$d^*_{jk} = w^*_{jk} + \epsilon_{jk},\ j = 0, \dots, j_1,\ k = 0, \dots, 2^j - 1$$

The prior distribution is defined on the space of wavelet coefficients. A posteriori mode is the Bayesian estimator. that maximizes the posterior likelihood (the law of the coefficients given the observations). We focuss our study on the properties of the a posteriori mode. We first consider the Bernoulli constrained model.

Bernoulli constrained model

Given $\alpha > 0$ and $\eta > 0$, at each fixed level $j > 0$, we set randomly $[2^{\eta j}]$ coefficients w_{jk} to the value $2^{-\alpha j}$ and the other coefficients to zero. So that at level j, the wavelet coefficients of the unknown function f^* lies in the set

$$\Omega_j = \{\omega = (\omega_k)_{k=0,\dots,2^j-1} \in \{0, 2^{-\alpha j}\}, \sum_{k=0}^{2^j-1} \omega_k = 2^{(\eta-\alpha)j}\}.$$

We take as prior probability on this set the uniform one. Hence if $w_j = (w_{j1}, \dots, w_{j\,2^j-1})'$, then

$$\forall \omega \in \Omega_j,\ \mathbf{P}(w_j = \omega) = \frac{1}{C_{2^j}^{[2^{\eta j}]}}$$

so at each level, the coefficients follow the uniform prior

$$w_j \sim \frac{1}{C_{2^j}^{[2^{\eta j}]}} \sum_{\omega \in \Omega_j} \delta_\omega.$$

For $\omega \in \Omega_j$, the canonical distribution of $d_j = (d_{j1}, \dots, d_{j2^j-1})'$ given $\{\omega_j = \omega\}$ is the Gaussian distribution $N(\omega, \sigma^2 Id_{2^j})$. Given $d_j = d^*_j$, the posterior distribution puts the weight:

$$\frac{\exp(-\frac{1}{2\sigma^2}||d^* - w||^2)}{C_{2^j}^{[2^{\eta j}]} \sum_{w_j \in \Omega_j} \exp(-\frac{1}{2\sigma^2}||d^* - w||^2)}$$

on the configuration $\omega \in \Omega_j$. So the posterior mode $\hat{w_j}$ satisifies

$$\begin{aligned} \hat{w_j} &= \arg\max_{w \in \Omega_j} p((w_j)|d^*_j) \\ &= \arg\min_{w \in \Omega_j} -\log p((w_j)|d^*_j) \\ &= \arg\min_{w_j \in \Omega_j} \frac{1}{2\sigma^2} \sum_{k=0}^{2^j} |d^*_{jk} - w_{jk}|^2, \end{aligned} \tag{5.1.6}$$

where $p(.|d_j^*)$ is the posterior density with respect to the uniform measure on Ω_j. With the particular form of the optimization problem (5.1.6), we recognize a constrained least squares estimator. To solve this minimization problem, first observe that: $|x| < |x - 2^{-\alpha j}|$ if and only if $x < 2^{-\alpha j - 1}$. Hence, a good candidate to be solution of (5.1.6) could be the thresholded estimator

$$\hat{w}_{jk} = 2^{-\alpha j} 1_{d_{jk}^* > 2^{-\alpha j - 1}}$$

But this estimator does not necessarly satisfies the constraint that the number of its non zero coefficients is equal to $[2^{\eta j}]$.

To take into account this constraint, sort, for each given j, the d_{jk}^*'s in a decreasing order:

$$d_{(0)}^* \geq \cdots \geq d_{([2^{\eta j}])}^* \geq \ldots d_{(2^j - 1)}^*$$

and estimate the $[2^{\eta j}]$ first coefficients by $2^{-\alpha j}$ and the others by zero. This estimation is the solution to the minimization criterium. We may think that the obtained estimator is not very efficient, since the estimated coefficients are very far from the observations and large coefficients are not represented. But we have to keep in mind that we estimate a multifractal function with representing coefficients numerous but close to zero. Such an estimator can be viewed as a hard-thresholding wavelet estimator similar to the ones studied by Donoho, Johnstone, Kerkyacharian and Picard in [DJKP97] or in [DJKP95] and in [DJKP96b]. But the threshold level is random and depends on the data at each level. Indeed the nonparametric estimator of f^* obtained by this method is:

$$\hat{f}_n = \sum_{j=0}^{j_1} \sum_{k=0}^{2^j} 2^{-\alpha j} 1_{|d_{jk}^*| \geq d_{(j[2^{\eta j}])}^*}$$

where $2^{j_1} = n$.

For simplicity reasons we fix a resolution level $j > 0$ and forget indexes in j for a while. Write $n = 2^j$, $p = 2^{\eta j}$ and, for $x = (x_1, \ldots, x_n)' \in \mathbb{R}^n$ define $\mathbf{k} \in \{1, \ldots, n\}^n$ a vector of indices such that, $x_{k_i} \neq x_{k_j}$ for $i \neq j$, we have

$$x_{k_1} \geq x_{j_2} \geq \cdots \geq x_{k_n}.$$

Indeed k_1 is the position of the greatest value of $(x_i)_{i=1,\ldots,n}$, k_2 the position of the second largest coefficient and so on. Set $\mathbf{k}^* = (\mathbf{k_1^*}, \ldots, \mathbf{k_n^*})$ the order indices of the true configuration at a fixed level j, which gives the position of the $w_{j.}$ in the decreasing order. Set also $\hat{\mathbf{k}} = (\hat{\mathbf{k}}_\mathbf{1}, \ldots, \hat{\mathbf{k}}_\mathbf{n})$ the configuration of the observed data, so we now have

$$d_{\hat{k}_1}^* \geq d_{\hat{k}_2}^* \geq \cdots \geq d_{\hat{k}_n}^*.$$

According to our former calculations, the estimator maximizing the posterior likelihood is defined by the estimated coefficients $(\hat{w}_{jk})$, where for at each level the estimators $(\hat{w}_{jk})_{k=0,\ldots,2^j-1}$ are given by:

$$\begin{cases} \hat{w}_{jk} = 2^{-\alpha j} & , \text{if } k \in \{\hat{k}_0, \ldots, \hat{k}_p\} \\ \hat{w}_{jk} = 0 & , \text{if } k \in \{\hat{k}_{p+1}, \ldots, \hat{k}_n\}. \end{cases} \tag{5.1.7}$$

So the quality of our approximation will depend on the quality of the estimation of the true position of the maximal wavelet coefficients.

The following theorem describes the behavior of our nonparametric estimator.

Theorem 5.1.3. *Assume that the multifractal function* $f^* = \sum_{j=0}^{j_1} \sum_{k=0}^{2^j-1} w_{jk}^* \psi_{jk}$ *has been drawn according to the Bernoulli constrained model. Assume also that* $\alpha < \frac{1}{2}$. *Define* Π_{j_1} *the projection operator onto the space* V_{j_1} *with* $2^{j_1} = n$. *Then there exist positive finite constant* c *and* c_1 *such that:*

$$\mathbf{E}\|\Pi_{j_1} f^* - \hat{f}_n\|_2^2 \le c_1 \exp(-c^2 n^{1-2\alpha}) n^{\eta+1-2\alpha} \tag{5.1.8}$$

Remark 5.1.4. *The condition* $\alpha < \frac{1}{2}$ *implies that the wavelet coefficients can not be too small. Otherwise, the function* f^* *can not be differentiated from the noise which prevents any estimation.*

Corollary 5.1.5. *Assume that* $\alpha < \frac{1}{2}$. *The rate of convergence of the estimator* $\hat{f}_n$ *is given by the remainder term* $\sum_{j>j_1} 2^{-j}(w_{jk}^*)^2$. *Indeed, there exists a positive finite constant* c_2 *such that:*

$$\begin{aligned}\mathbf{E}\|f^* - \hat{f}_n\|_2^2 &= O_P\left(\sum_{j>j_1} 2^{-j}(w_{jk}^*)^2\right) \\ &= O_P\left(c_2 n^{-(1-\eta+2\alpha)}\right)\end{aligned}$$

which goes to zero when n *goes to infinity.*

Corollary 5.1.6. *Assume that* $\alpha < \frac{1}{2}$. *A straightforward application of Borel-Cantelli's lemma gives that* $\hat{f}_n \to f^*$ *a.s .*

The proof follows from linking this estimation method and cluster analysis of a Gaussian mixture. As a matter of fact, due to orthonormality of wavelet basis we have the following decomposition:

$$\begin{aligned}
\mathbf{E}||\hat{f} - \Pi_{j_1} f^*||_2^2 &= \mathbf{E} \sum_{(j,k)} 2^{-j} |\hat{w}_{jk} - w^*_{jk}|^2 \\
&= \mathbf{E} \sum_{j} 2^{-j} \sum_{l=0}^{2^j-1} |\hat{w}_{jk_l} - w^*_{jk^*{}_l}|^2 \\
&= \mathbf{E} \sum_{j} 2^{-j} \left(\sum_{l=0}^{p_j} |\hat{w}_{jk^*_l} - 2^{-\alpha j}|^2 + \sum_{l=p_j+1}^{2^j-1} |\hat{w}_{jk^*_l}|^2 \right) \\
&= \sum_{j} 2^{-j} 2^{-2\alpha j} \mathbf{E} \left(\sum_{l=0}^{p_j} 1_{k^*_l \notin \{\hat{k}_0,\ldots,\hat{k}_{p_j}\}} + \sum_{l=p_j+1}^{2^j-1} 1_{k^*_l \in \{\hat{k}_0,\ldots,\hat{k}_{p_j}\}} \right) \\
&= \sum_{j} 2^{-j} 2^{-2\alpha j} \left(\sum_{l=0}^{p_j} \mathbf{P}(k^*_l \notin \{\hat{k}_0, \ldots, \hat{k}_{p_j}\}) + \sum_{l=p_j+1}^{2^j-1} \mathbf{P}(k^*_l \in \{\hat{k}_0, \ldots, \hat{k}_{p_j}\}) \right) \\
&\leq \sum_{j} 2^{-j} 2^{-2\alpha j} \Big([2^{-\eta j}] \mathbf{P}(d_{k^*_0} < d_{\hat{k}_{p_j}}) + (2^j - [2^{\eta j}])(1 - \mathbf{P}(d_{k^*_{2^j-1}} < d_{\hat{k}_{p_j}})) \Big) \\
&\leq T_1 + T_2.
\end{aligned}$$

where we have set $p_j = [2^{\eta j}] - 1$, and use the independence of the random variables between each group. Such majorations give rise to another statistical problem that can be described as follows and is similar to the one described by Mc Lachlan in [McL82]:
Consider n random variables, X_i, $i = 1, \ldots, n$ belonging to two different populations such that

$$\underbrace{X_1, \ldots, X_p}_{(I)}, \underbrace{X_{p+1}, \ldots, X_n}_{(II)} \tag{5.1.9}$$

where the population (I) consists of independent Gaussian variables $\mathcal{N}(a, \sigma^2)$ and the population (II) consists of independent Gaussian variables $\mathcal{N}(0, \sigma^2)$. Moreover, the two groups are assumed to be independent. We consider a decreasing reordering of the variables

$$X_{(1)} \geq \cdots \geq X_{(p)} \geq \cdots \geq X_{(n)} \tag{5.1.10}$$

This model is a mixture model, as defined by Mc Leish and Small in [MS86], where we know precisely the different proportions and the values of the different means. If no assumptions were made, we could use well-known technics developed by Basford and Mc Lachlan in [BM85] to run a EM-algorithm [DLR77] in order to estimate both a and σ^2.
The link between the two problems is the following: at each fixed level j the wavelet coefficients can take two different values $a = a_j = 2^{-\alpha j}$ or 0 whether they are recognized as part of the $p_j + 1 = [2^{\eta j}]$ greatest coefficients. So if we rescale the coefficients by multiplying them by the same parameter $\frac{\sqrt{n}}{\sigma}$, the estimation problem turns to be a classification problem of X_k

random variables following Gaussian laws $\mathcal{N}(0,1)$ or $\mathcal{N}(\frac{\sqrt{n}}{\sigma}a,1)$. Our aim here is to bound the error of misclassifying a variable. Hence, we want to bound the following quantities

$$\mathbf{P}(d_{k_0^*} < d_{\hat{k}_{p_j}}), \text{ and } \mathbf{P}(d_{k_{2^j-1}^*} < d_{\hat{k}_{p_j}}) \tag{5.1.11}$$

which can be rewritten using the cluster analysis formalism:

$$\mathbf{P}(X_1 < X_{(p)}), \text{ and } \mathbf{P}(X_n < X_{(p)}). \tag{5.1.12}$$

If we define the rank statistics R_i, $i = 0, \ldots, n-1$ the two probabilities can be rewritten as $\mathbf{P}(R_1 < p)$ and $\mathbf{P}(R_n < p)$. Such problem has been studied very early in statistics (see Gumbel in [Gum58] for example). The following lemma gives a first rough majoration of the errors that will be sufficient in our work. The proof follows from straightforward combinatory calculations:

Lemma 5.1.7.
$$\begin{cases}\mathbf{P}(X_1 < X_{(p)}) & \leq (n-p)\mathbf{P}(X_1 < X_{p+1}) \\ \mathbf{P}(X_n < X_{(p)}) & \geq \mathbf{P}(\max_{i>p} X_i < \min_{i\leq p} X_i).\end{cases}$$

We have made the assumption that the function to be estimated is significant, in the sense that it can be distinguished from the white noise. Such assumptions imply that the wavelet coefficient can not be too small so we impose the condition

$$1 - 2\alpha > 0.$$

With this hypothesis, the two groups of Gaussian variables can be differianted since the mean $m_n = 2^{j_1/2-\alpha j}$ goes far from zero quickly enough.
The following lemma whose proof can be found in the appendix describes the asymptotic behavior of the two previous probabilitties.

Lemma 5.1.8. *There exist two finite positive constants c_1 and c_2 such that*

$$\mathbf{P}(X_1 < X_{(p+1)}) \leq c_1 \mathrm{exp}\left(-\frac{m_n^2}{4}\right)$$
$$\mathbf{P}(X_n > X_{(p)}) \leq c_2 \exp\left(-\frac{m_n^2}{4}\right).$$

Putting together all the results we obtain:

$$\begin{aligned} T_1 &\leq \sum_{j\leq j_1} 2^{-j} 2^{\eta j - 2\alpha j} \mathbf{P}(X_1 < X_{(p)}) \\ &\leq \sum_{j\leq j_1} 2^{(\eta-2\alpha)j} 2^{-\frac{j_1}{2}} 2^{\alpha j} \exp(-\frac{2^{j_1-2\alpha j}}{4}) \\ &\leq \exp(-\frac{2^{j_1(1-2\alpha)}}{4}) 2^{(\eta-\alpha-\frac{1}{2})j_1} \end{aligned}$$

But since $1-2\alpha > 0$ we can conclude that T_1 goes to zero with exponential rate of convergence whatever the value of η may be. For the second term we have the following bound:

$$\begin{aligned} T_2 &\leq \sum_{j\leq j_1} 2^{-2\alpha j}\mathbf{P}(X_n > X_{(p)}) \\ &\leq \sum_{j\leq j_1} 2^{j(1+\eta-2\alpha)}\exp(-c^2 2^{j_1-2\alpha j}) \\ &\leq \exp(-c^2 2^{(1-2\alpha)j_1})2^{(1+\eta-2\alpha)j_1} \end{aligned}$$

With the same assumption over α, the last inequality proves that $T_1 \to 0$ at an exponential rate.

Gaussian Model

Up to now, we have tried to recover functions whose wavelet coefficients can only take two values: 0 and $2^{-\alpha j}$. From now on, we extend our results to the cases where we allow non zero coefficients to take values different from $2^{-\alpha j}$ as it stated in Section 5.1.2. As a result we may rewrite the model as follows. Let $F_j = (f_{jk})_{k=0,\ldots,2^j-1}$, $j = 0,\ldots,j_1$ a random variable with values on the set $\{0, 2^{-\alpha j}\}$ following an uniform law over the set Ω_j described in Section 3.1. Let (z_{jk}), $j = 0,\ldots,j_1$, $k = 0,\ldots,2^j-1$ independent Gaussian variables $\mathcal{N}\left(0,\Delta_j^2\right)$, taken also independent from the noise. The variance $\Delta_j > 0$ are such that $\sum_j 2^{-j}\Delta_j^2 < \infty$. The coefficients of the observed random function $f^* = \sum_{j=0}^{j_1}\sum_k w^*_{jk}\psi_{jk}$ are:

$$w^*_{jk} = f^*_{jk} + z^*_{jk}1_{f_{jk}\neq 0},\ j = 0,\ldots,j_1,\ k = 0,\ldots,2^j-1 \tag{5.1.13}$$

We observe this function with a Gaussian additive noise:

$$\left\{ \begin{array}{l} d^*_{jk} = w^*_{jk} + \epsilon_{jk} \\ j = 0,\ldots,j_1,\ k = 0,\ldots,2^j-1 \end{array} \right. \tag{5.1.14}$$

We propose to use an estimator close to the maximum a posteriori estimator used previously with a slight change: we try to determine the highest coefficients that will be non zero and then smooth them in order to give a good estimation. The ideas of the smoothing effect come from the following obvious lemma.

Lemma 5.1.9. *Consider two independent Gaussian variables*

$$X \sim \mathcal{N}\left(m_1,\alpha^2\right), \quad Y \sim \mathcal{N}\left(m_2,\beta^2\right)$$

We have

$$\begin{aligned} \mathbf{E}(X|X+Y) &= m_1 + \frac{\alpha^2}{\alpha^2+\beta^2}(X+Y-(m_1+m_2)) \\ \mathrm{Var}(X-\mathbf{E}(X|X+Y)) &= \frac{\alpha^2\beta^4+\alpha^4\beta^2}{(\alpha^2+\beta^2)^2} \end{aligned}$$

To estimate f^*, we use the following estimator:

$$\hat{f}_n = \sum_{j=0}^{j_1} \sum_{k=0}^{2^j-1} \hat{w}_{jk}\psi_{jk}$$

where:

$$\begin{aligned} \hat{w}_{jk} = &\quad 2^{-\alpha j} + \frac{\Delta_j^2}{\Delta_j^2+\frac{\sigma^2}{n}}(d_{jk}^* - 2^{-\alpha j}) &\quad , \forall k \in \{\hat{k}_0, \ldots, \hat{k}_p\} \\ = &\quad 0 &\quad , \forall k \notin \{\hat{k}_0, \ldots, \hat{k}_p\} \end{aligned}$$

where $\hat{\mathbf{k}} = (\hat{\mathbf{k}}_\mathbf{0}, \ldots, \hat{\mathbf{k}}_\mathbf{p})$ corresponds to the position for a fixed level j of the p highest observed coefficients which must correspond to the true non-zero coefficients of the function. We can see that there are slight changes with the first model. As a matter of fact, an additional estimation issue is added to the original classification problem: the quadratic loss is divided into three terms corresponding to the misschosing the position of the greatest coefficients and an extra term corresponding to the estimation error. The following theorem describes the behavior of our estimator.

Theorem 5.1.10. *Assume that f^* has been drawn according to the Gaussian extension of the Bernoulli constained model. There exists a finite positive constant c_3 such that for $f^* = \sum_{j=0}^{j_1}\sum_k w_{jk}^*\psi_{jk}$.*

$$\mathbf{E}||\Pi_1 f^* - \hat{f}_n||_2^2 \leq c_3 n^{-(2-\eta)} \tag{5.1.15}$$

Corollary 5.1.11. *The estimator $\hat{f}_n$ converges for $1-\eta+2\alpha \geq 0$. Indeed there exist positive finite constants c_3 and c_2 such that we get the following upper bound.*

$$\begin{aligned} \mathbf{E}||f^* - \hat{f}_n||_2^2 &\leq c_3 n^{-(2-\eta)} + \sum_{j>j_1} 2^{-j}\mathbf{E}|d_{jk}^*|^2 \\ &\leq c_3 n^{-(2-\eta)} + \sum_{j>j_1} 2^{-j}\Delta_j^2 + c_2 n^{-(1-\eta+2\alpha)} \\ &= O_P(n^{-(1-\eta+2\alpha)} \wedge n^{-(2-\eta)}) \end{aligned}$$

as soon as $\sum_{j>j_1} 2^{-j}\Delta_j^2 \leq \frac{1}{n}$.

Reasoning as previously we have:

$$\begin{aligned} \mathbf{E}||\hat{f} - \Pi_1 f^*||_2^2 &= \mathbf{E}\sum_{(j,k)} 2^{-j}|\hat{w}_{jk} - w_{jk}^*|^2 \\ &= \sum_j 2^{-j}\mathbf{E}\left(\sum_{l=0}^{p_j}(w_{j,k_l^*}^*)^2 1_{k_l^*\notin\{\hat{k}_0,\ldots,\hat{k}_{p_j}\}}\right) \quad (I) \\ &+ \sum_j 2^{-j}\mathbf{E}\left(\sum_{l=0}^{p_j}(\hat{w}_{j,k_l^*} - w_{j,k_l^*})^2 1_{k_l^*\in\{\hat{k}_0,\ldots,\hat{k}_{p_j}\}}\right) \quad (II) \\ &+ \sum_j 2^{-j}\mathbf{E}\left(\sum_{l=p_j+1}^{2^j-1} \hat{w}_{j,k_l^*}^2 1_{k_l^*\in\{\hat{k}_0,\ldots,\hat{k}_{p_j}\}}\right) \quad (III) \end{aligned}$$

The three quantities can be majorated as shown in the following lemma:

Lemma 5.1.12.

$$(I) \leq \sum_j 2^{(\eta-1)j} \mathbf{P}^{\frac{1}{2}}(k_l^* \notin \{\hat{k}_0, \ldots, \hat{k}_{p_j}) A_j^{1/2} \tag{5.1.16}$$

$$(II) \leq \sum_j 2^{-j} \sum_{l>p_j} \mathbf{P}^{\frac{1}{2}}(k_l^* \in \{\hat{k}_0, \ldots, \hat{k}_{p_j}\}) \mathbf{E}^{\frac{1}{2}} \left(\frac{2^{-\alpha j} \frac{\sigma^2}{n}}{\Delta_j^2 + \frac{\sigma^2}{n}} + \frac{\Delta_j^2}{\Delta_j^2 + \frac{\sigma^2}{n}} \epsilon_{jk} \right)^4 \tag{5.1.17}$$

$$(III) \leq \sum_j 2^{(\eta-1)j} (2^j - 2^{\eta j}) \frac{\mathbf{P}^{\frac{1}{2}}(k_n^* \in \{\hat{k}_0, \ldots, \hat{k}_{p_j}\})}{(\Delta_j^2 + 2^{-j_1}\sigma^2)^2} B_j^{1/2} \tag{5.1.18}$$

where, for $j = 1, \ldots, 2^{j_1}$, A_j and B_j are defined by if we have set $\sigma_j^2 = \Delta_j^2 + 2^{-j_1}\sigma^2$:

$$\begin{aligned} A_j &= 2\sigma_j^4 + 6\, 2^{-2\alpha j}\sigma_j^2 + 2^{-4\alpha j} \\ B_j &= 32^{-2j_1}\sigma^4\Delta_j^8 + 2^{-4\alpha j}\sigma^8 2^{-4j_1} + 6\sigma^6\Delta_j^5 2^{-2\alpha j} 2^{-3j_1} \end{aligned}$$

Proof. The proof of this result is rather technical and imply calculations. It is postponed to the appendix. □

We point out that A_j and B_j both tend towards zero as n increases so we the convergence of the first and second terms of the quadratic loss will be ensured by the good classification properties of the model. As a matter of fact the only modifications with the first model are the change of the variance of the errors but they still have the same asymptotic behavior. So due to Lemma 3.6 which can be proved in that specific case, we know that the probability of misclassifying the coefficients tends towards zero at an exponential rate of convergence. As a consequence, the quadratic rate of convergence will only depends on the central term. Indeed, we get the following bounds for positive constants c_1, c_2 and c:

$$\begin{aligned} (I) &\leq \sum_{j=0}^{j_1} 2^{(\eta-1)j} \\ &< \exp(-\frac{m_n^2}{8})[3(\Delta_j^2 + \frac{\sigma^2}{n})^2 \\ &+ 2\, 2^{-2\alpha j}(\Delta_j^2 + \frac{\sigma^2}{n}) + 2^{-4\alpha j}] \\ &\leq c_1 \exp(-c^2 n^{1/2-\alpha}) \sup_{j \leq j_1}(\Delta_j^2) 2^{(\eta-1)j_1} \end{aligned}$$

The last expression goes to zero at an exponential rate of convergence as soon as the variance term Δ_j^2 does not go to infinity at the same rate.

$$\begin{aligned}
(II) &\leq \sum_{j=0}^{j_1} 2^{(\eta-1)j} \frac{\Delta_j^2 \frac{\sigma^4}{n^2} + \Delta_j^4 \frac{\sigma^2}{n}}{(\Delta_j^2 + \frac{\sigma^2}{n})^2} \\
&\leq \sum_{j=0}^{j_1} 2^{(\eta-1)j} \Delta_j^2 \frac{\sigma^2}{n} \\
&\leq \frac{\Delta_j^2 + \frac{\sigma^2}{n}}{(\Delta_j^2 + \frac{\sigma^2}{n})^2} \\
&\leq \sum_{j=0}^{j_1} 2^{(\eta-1)j} \sum_{j=0}^{j_1} 2^{(\eta-1)j} \frac{\sigma^2}{n} \\
&\leq c_2 n^{\eta-2}
\end{aligned}$$

which goes to zero as well.

To conclude the last term can be written as follows: let $F_j = \dfrac{2^{\eta j} B_j^{\frac{1}{2}}}{(\Delta_j^2 + 2^{-j_1}\sigma^2)^2}$, we get

$$(III) \leq \sum_{j=0}^{j_1} \exp(-c^2 m_n^2) F_j$$

where F_j does not go to infinity at an exponential rate which enables us to conclude that the last term goes to zero at an exponential rate of convergence which concludes the proof.

5.1.4 Simulation results

The following results have been obtained using Matlab software in the Bernoulli constraint model. In Figure 2, we present the Bayesian reconstruction of multifractal function generated with a choice of $\eta = 0.4$ and $\alpha = 0.1$ observed with a Gaussian noise with variance 1. In the Figure 3, the coefficients of the multifractal function are drawn with a choice of $\eta = 0.6$ and $\alpha = 0.25$, while the function is observed with a Gaussian noise with variance 4. Each figure is divided into four part: in the first subfigure, we plot the multifractal function. The second subplot shows the observed data while the third subplot shows the estimator of the function. Finally, in the last subplot, the estimator together with the true function are plotted. Even if some peaks are badly allocated, the Bayesian reconstruction provide good visual performances and preserves the energy of the signal.

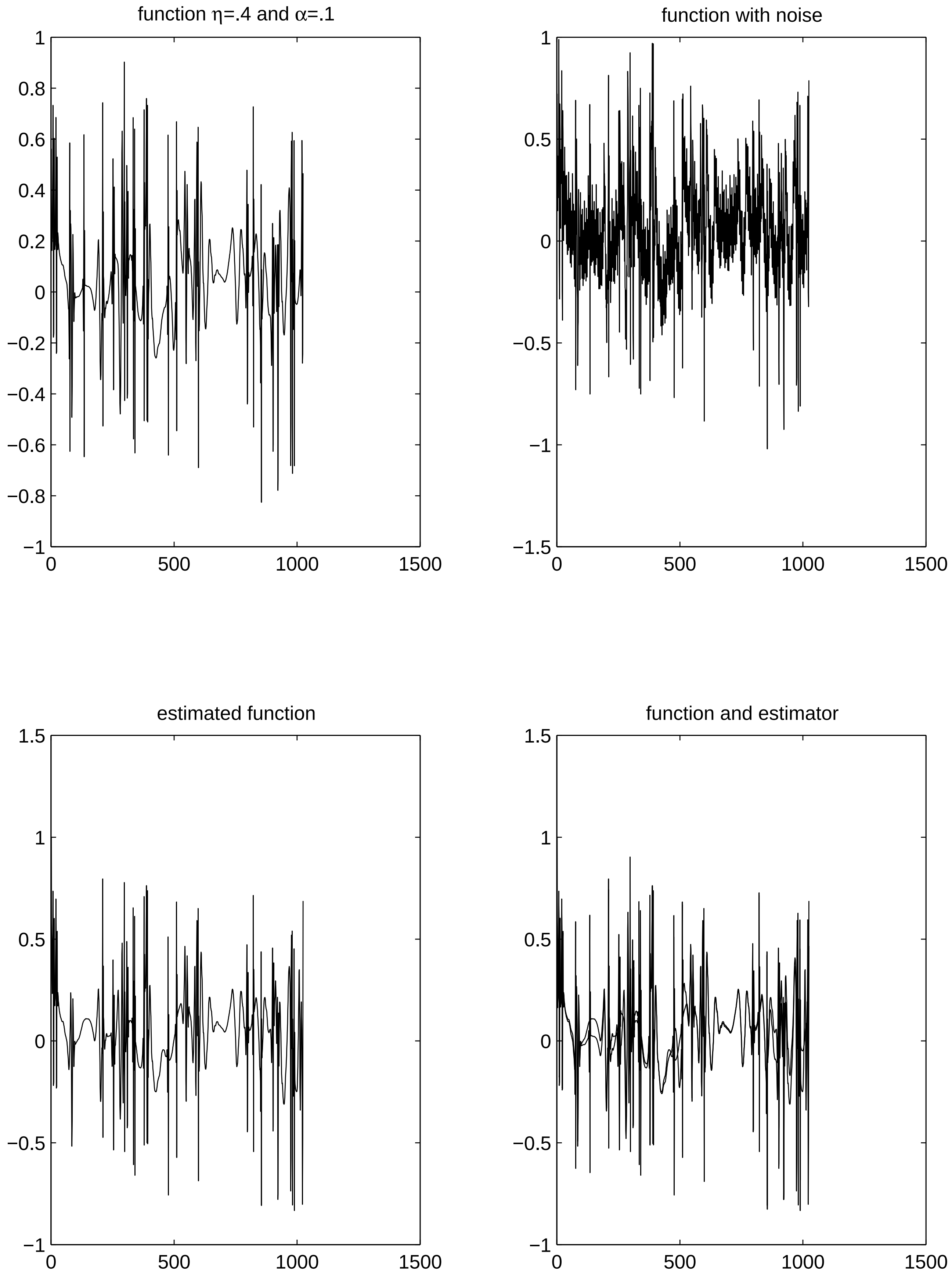


Figure 5.2: Bayesian reconstruction of a multifractal process

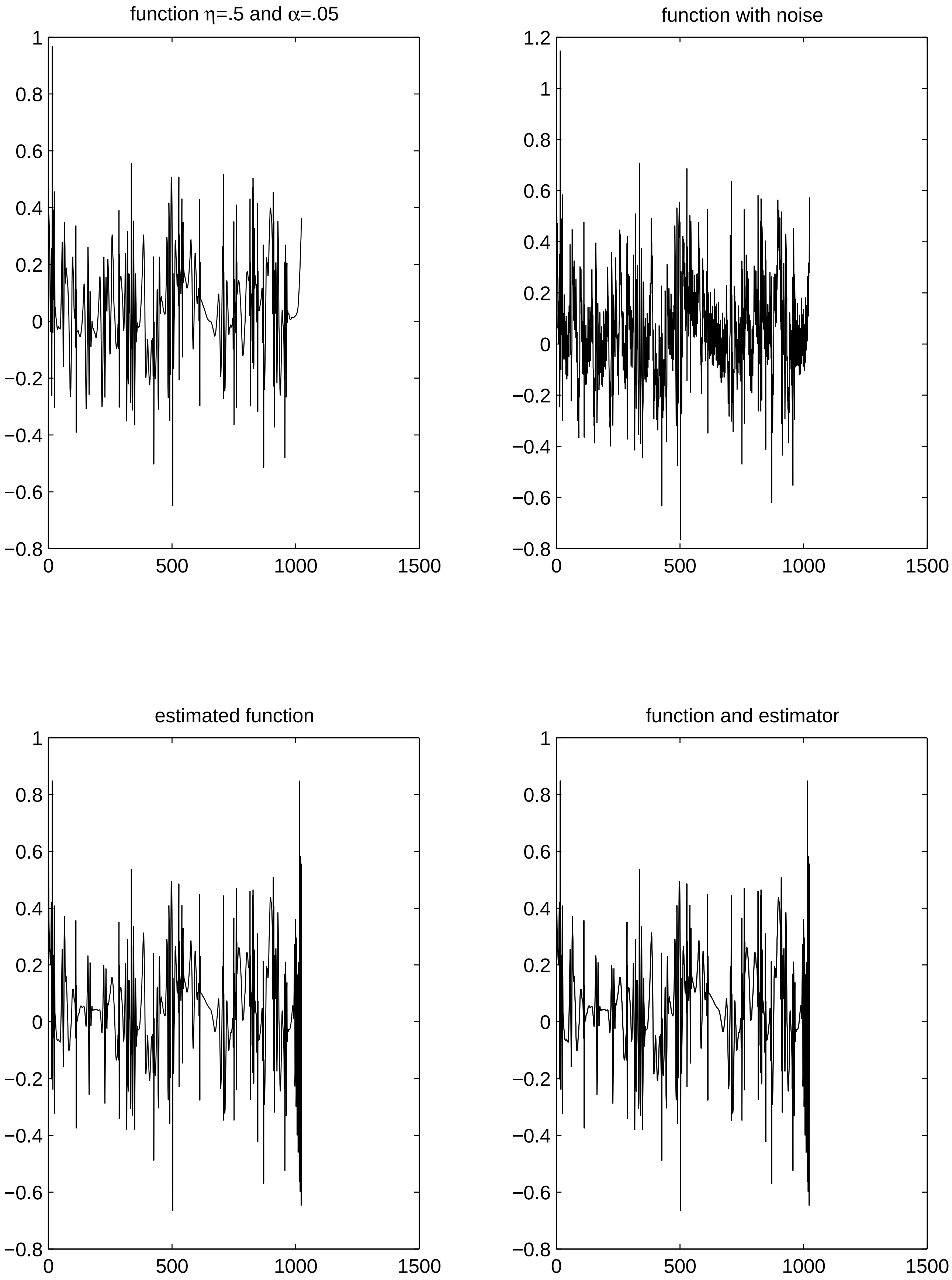


Figure 5.3: Bayesian reconstruction of a multifractal process

5.1.5 Estimation of model parameters

EM-algorithm

The general framework of the EM algorithm is the following: consider random variables $X = (Y, Z)$, where only Y can be observed. Z is a missing unobservable data while X is the complete data. Assume that X has been drawn according to $\mathbf{P}_{YZ}$ and that only y a realization of Y is available. The problem is to estimate θ^* and to predict Z using y.

Let μ, a Borelian positive measure σ-finite over $\mathbb{R}^n$. Consider the parametric framework : let $\theta \in \Theta \subset \mathbb{R}^m$ and write $\mathbf{P}_{YZ}(., \theta)$ the distribution under θ of X

$$d\mathbf{P}_{YZ}(y, z, \theta) = p_{YZ}(y, z, \theta) d\mu(y, z)$$

and $\mathbf{P}_Y(., \theta)$ the distribution of Y

$$d\mathbf{P}_Y(\omega, \theta) = p_Y(y, \theta) d\tilde{\mu}(\omega),$$

where $\tilde{\mu} = \int \mu(., dz)$. A classical estimation with a maximization of the log-likelihood can lead to a solution of the problem but the difficulty lies in the fact, except in very simple cases, the distribution of Y is difficult to write whereas the joint distibution function (Y, Z) is easier. The EM algorithm enables to solve the maximization problem of $p_Y(y, \theta)$ by only optimizing $p_{YZ}(y, z, \theta)$. So a direct application of such algorithm is the classification problem in mixture settings, see for instance Mc Leish in [MS86].
Thus we illustrate the previous frame by a single mixture model. This contains two Gaussian variables. Let $Y_1, \ldots, Y_n$ be an i.i.d sample of a random vector Y with density:

$$f\left(y, \Psi\right) = \sum_{i=1}^{2} \pi_i \phi\left(y; \mu_i, \sigma\right),$$

where $\pi_1 + \pi_2 = 1$, and $\phi\left(y; \mu_i, \sigma\right)$ is the Gaussian distribution function with mean μ_i and variance σ^2. Define the parameter of interest, $\Psi = (\pi_1, \theta^T)^T$, with θ a vector with components μ_i and σ, $(i = 1, 2)$. The log-likelihood is:

$$L\left(\Psi\right) = \sum_{j=1}^{n} \log\left(\pi_1 \phi\left(y_j; \mu_1, \sigma\right) + (1 - \pi_1)\phi\left(y_j; \mu_2, \sigma\right)\right).$$

An estimator of Ψ can be obtained my maximizing the log-likelihood so is a solution of :

$$\frac{\partial L\left(\Psi\right)}{\partial \Psi} = 0. \tag{5.1.19}$$

To apply EM-algorithm, we transform this model into a missing observation model. We define z_{ij}, a random variable equal to 1 if y_j comes from the i component , (with law $\mathcal{N}(\mu_i, \sigma)$,) and 0 otherwise (for $i = 1, 2; j = 1, ..., n,$) and set $z_j = (z_{1j}, z_{2j})^T$. Moreover set $y_=(y_1, ..., y_n)^T$, and consider that y_i comes from the mixture random vector, indeed there exists only one 1 in every z_j. The complete data is then:

$$x_c = (x_1^T, ..., x_n^T),$$

with $X_1 = (Y_1, Z_1^T)^T, ..., X_n = (Y_n, Z_n^T)^T$. Suppose that $X_1, ..., X_n$ are i.i.d with $z_1, ..., z_n$, n realizations of a Bernoulli with parameter π. In our new model the log-likelihood can be rewritten in the following way:

$$L_c(\Psi) = \sum_{i=1}^{2} \sum_{j=1}^{n} z_{ij} \log\left[\pi_i \phi\left(y_j; \mu_i, \sigma\right)\right].$$

First, the theory of EM algorithm tells us that maximizing the log-likelihood is equivalent to maximizing the following quantity:

$$\begin{aligned} Q\left(\Psi, \Psi^{(k)}\right) &= \mathbf{E}\left(L_c(\Psi)/y_{obs}; \Psi^{(k)}\right) \\ &= \sum_{i=1}^{g} \sum_{j=1}^{n} \mathbf{E}\left(Z_{ij}/y_{obs}; \Psi^{(k)}\right) \log\left[\pi_i \phi\left(y_j; \mu_i, \sigma\right)\right]. \end{aligned}$$

Then, the E stage is done by replacing z_{ij} by

$$\begin{aligned} \tau_i\left(y_j; \Psi^{(k)}\right) &= \mathbf{E}\left(Z_{ij}/y_j; \Psi^{(k)}\right) \\ &= \mathbf{P}\left[Z_{ij} = 1/y_j; \Psi^{(k)}\right] \\ &= \frac{\pi_i^{(k)} \phi\left(y_j; \mu_i^{(k)}, \sigma^{(k)}\right)}{\sum_{h=1}^{2} \pi_h^{(k)} \phi\left(y_j; \mu_h^{(k)}, \sigma^{(k)}\right)} \end{aligned}$$

for $i = 1$, and $j = 1, ..., n$.
The M-step of the $(k+1)^{th}$ iteration, consists in choosing Ψ $\left(\text{writed } \Psi^{(k+1)}\right)$ in order to maximize $Q\left(\Psi, \Psi^{(k)}\right)$. The estimators after the $(k+1)^{th}$ step are defined by:

$$\begin{aligned} \pi_i^{(k+1)} &= \frac{1}{n} \sum_{j=1}^{n} \tau_i^{(k)}\left(y_j\right) \\ \mu_i^{(k+1)} &= \frac{\sum_{j=1}^{n} \tau_i^{(k)}\left(y_j\right) y_j}{\sum_{j=1}^{n} \tau_i^{(k)}\left(y_j\right)} \\ \sigma^{(k+1)} &= \frac{\sum_{j=1}^{n} \tau_i^{(k)}\left(y_j\right)\left(y_j - \mu_i^{(k+1)}\right)^2}{\sum_{j=1}^{n} \tau_i^{(k)}\left(y_j\right)} \end{aligned}$$

for $i = 1, 2$ and$\tau_i^{(k)}\left(y_j\right) = \tau_i\left(y_j; \Psi^{(k)}\right)$ We now can apply this general algorithm to our concrete case with the assumption that variance σ^2 is known: write $m = 2^{-\alpha j}$ and $\pi = 2^{(\eta_1)j}$. At a fixed level j, the augmented likelihood is

$$\begin{aligned}
L(d^*_{jk}, m, \pi) &= \sum_k \log \pi^{z_{jk}} \exp(-\frac{n}{2\sigma^2}(d^*_{jk} - m)^2 z_{jk})(1-\pi)^{1-z_{jk}} \exp(-\frac{n}{2\sigma^2} {d^*_{jk}}^2 (1 - z_{jk})) \\
&= \sum_k z_{jk} \log \frac{\pi}{1-\pi} - \frac{n}{2\sigma^2} \sum_k (z_{jk}(d^*_{jk} - m)^2 - (1 - z_{jk}) {d^*_{jk}}^2) + \sum_k \log(1-\pi) \\
&= \log\left(\frac{\pi}{1-\pi}\right) \sum_k z_{jk} - \frac{n}{2\sigma^2} m^2 \sum_k z_{jk} + \frac{nm}{\sigma^2} \sum_k d^*_{jk} z_{jk} + 2^j \log(1-\pi) \\
&= (\log \frac{\pi}{1-\pi}; m^2; m)(\sum_k z_{jk}; -\frac{n}{2\sigma^2} \sum_k z_{jk}; \frac{n}{\sigma^2} \sum_k d^*_{jk} z_{jk})' + 2^j \log(1-\pi) \\
&= a(\theta)' b(X) + c(\theta) + d(X)
\end{aligned}$$

We recognize an exponential family. The EM algorithm can then be written at the $i+1$-step:

- E step:
$$\mathbf{E}(b(X)|d^*, \theta^i) = (\sum_k \hat{z}^{(i)}_{jk}; -\frac{n}{2\sigma^2} \sum_k \hat{z}^{(i)}_{jk}; \frac{n}{\sigma^2} \sum_k d^*_{jk} \hat{z}^{(i)}_{jk})$$
where $\hat{z}^{(i)}_{jk} = \mathbf{P}(z_{jk} = 1 | d^*, \theta^{(i)})$.

- M step: in order to maximize the functions:
$$\begin{cases} f(\pi) & = \log\left(\frac{\pi}{1-\pi}\right) \sum_k z_{jk} + 2^j \log(1-\pi) \\ g(m) & = -\frac{n}{2\sigma^2} m^2 \sum_k z_{jk} + \frac{nm}{\sigma^2} \sum_k d^*_{jk} z_{jk} \end{cases}$$
write the first order condition and this gives raise to the two estimated parameters:
$$\hat{m}^{(i+1)} = \frac{\sum_k d^*_{jk} \hat{z}^{(i)}_{jk}}{\sum_k \hat{z}^{(i)}_{jk}} \tag{5.1.20}$$
$$\hat{\pi}^{(i+1)} = \frac{1}{2^j} \sum_k \hat{z}^{(i)}_{jk} \tag{5.1.21}$$

Perspectives

We have used the EM algorithm on data sets d^*_{jk}, $k = 0, \ldots, 2^j - 1$ at fixed j. So in practice, we get two possibilities: either we use an iteration of the algorithm as j increases, using the result of each iteration as a starting point of the next step, or we use the EM algorithm on the last scale j_1 where we have the most information. If we use log-likelihood estimation over the whole data set, the behavior of the likelihood process depends heavily on the range of the parameter η and α and its maximization does not give always consistent estimators. So using a weighted version of the process or empirical mean estimators of the parameters seem to be more relevant.

5.1.6 Appendix

Proof of Lemma 5.1.8:

Proof. First of all, we point out that, the probabilities remain unchanged if we multiply the random variables by the same constant, so now the random variables follow either $\mathcal{N}(0,1)$ or $\mathcal{N}(m_n,1)$ where $m_n = a\frac{\sqrt{n}}{\sigma}$. Such hypothesis is satisfied for small choices of α that leads to significant wavelet coefficients. Otherwise, the coefficients of the signal are too small to be differentiated from the Gaussian white noise and the estimation problem is made impossible. Under this assumption, when n goes to infinity $m_n \to \infty$, so the two components of the Gaussian mixture are well divided, and the classification issue leads to efficient results.

- The first probability can be majorated as follows:

$$\begin{aligned}
&\mathbf{P}(X_1 < X_{p+1})\\
&= \mathbf{P}\left(\mathcal{N}(m_n,1) < \mathcal{N}(0,1)\right)\\
&= \int\int_{x<y} \frac{1}{2\pi}\exp(-\frac{(x-m_n)^2}{2})\exp(-\frac{y^2}{2})dxdy\\
&= \frac{1}{\sqrt{2\pi}}\int \exp(-\frac{y^2}{2})\int_{-\infty}^{y}\frac{1}{\sqrt{2\pi}}\exp(-\frac{(x-m_n)^2}{2})dxdy\\
&= \frac{1}{\sqrt{2\pi}}\int \exp(-\frac{y^2}{2})(1-\Phi(m_n-y))dy\\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{m_n-1} \exp(-\frac{y^2}{2})(1-\Phi(m_n-y))dy + \frac{1}{\sqrt{2\pi}}\int_{m_n-1}^{\infty}\exp(-\frac{y^2}{2})(1-\Phi(m_n-y))dy
\end{aligned}$$

where Φ is the repartition function of a standard normalized Gaussian variable. Using an asymptotic equivalence, we know that

$$1-\Phi(y-m_n) \le \frac{1}{\sqrt{2\pi}}\frac{\exp(-(m_n-y)^2/2)}{m_n-y}.$$

With this inequality we can write:

$$\begin{aligned}
\mathbf{P}(X_1 < X_{p+1}) &\le \frac{1}{\sqrt{2\pi}}\int_{m_n-1}^{\infty}\exp(-\frac{y^2}{2})(1-\Phi(m_n-y))dy\\
&+\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{m_n-1}\exp(-\frac{y^2}{2})(1-\Phi(m_n-y))dy\\
&\le \int_{y\ge m_n-1}\exp(-\frac{y^2}{2})dy + \frac{1}{\sqrt{2\pi}}\int_{y<m_n-1}\exp(-\frac{y^2}{2})\frac{\exp(-\frac{(m_n-1-y)^2}{2})}{m_n-1-y}dy\\
&\le \exp(-\frac{(m_n-1)^2}{2}) + c_1\exp(-\frac{m_n^2}{4})\\
&\le c_2\exp(-\frac{m_n^2}{4}) + c_1\exp(-\frac{m_n^2}{4})\\
&\le c\exp(-\frac{m_n^2}{4}).
\end{aligned}$$

where c_1, c_2 and c are positive finite constants. So we can conclude that there exists a finite constant $c_1 \geq c$ such that:

$$\mathbf{P}(X_1 < X_{p+1}) \leq c_1 \mathrm{exp}(-\frac{m_n^2}{4}) \tag{5.1.22}$$

- The second inequality can be bounded as follows:

For the second probability, we use the law of order statistics since, between each group the random variables are independently equi-distributed. The next lemma gives the law of the variables between each group.

Lemma 5.1.13. *The density of* $\min_{i=1,\dots,p} Y_i$ *is*

$$\frac{n-p}{\sqrt{2\pi}} \exp(-\frac{(x-m_n)^2}{2}) \left(\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-\frac{(t-m_n)^2}{2}) dt\right)^{n-p-1}$$

The density of $\max_{i=p+1,\dots,n} Y_i$ *is*

$$\frac{p}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) \left(\int_{x}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2}) dt\right)^{p-1}$$

The proof is obvious using definitions of order statistics.

$$\begin{aligned}
&1 - \mathbf{P}(\max_{i>p} X_i < \min_{i \leq p} X_i) \\
&= \int\int_{x>y} \frac{(n-p)p}{\sqrt{2\pi}} \exp(-x^2/2) \exp(-(y-m_n)^2/2) \Phi(x)^{n-p-1} (1 - \Phi(y-m_n))^{p-1} dx dy \\
&\leq p(n-p) \int\int_{x>y+m_n} \exp(-\frac{x^2}{2}) \Phi(x)^{n-p-1} \exp(-\frac{y^2}{2})(1-\Phi(y))^{p-1} dx dy \\
&\leq p(n-p) \int \left(\int_{x>y+m_n} \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) dx\right) \exp(-\frac{y^2}{2})(1-\Phi(y))^{p-1} dy \\
&\leq p(n-p) \int \exp(-\frac{(y+m_n)^2}{2}) \exp(-\frac{y^2}{2}) \exp(-(p-1)\frac{y^2}{2}) dy \\
&\leq p(n-p) \exp\left(-\frac{pm_n^2}{2(p+1)}\right)
\end{aligned}$$

As a result, we have proved that, there exists a positive constant c such that

$$\mathbf{P}(X_n \geq X_{(p)}) \leq cnp \exp\left(-\frac{m_n^2}{4}\right) \tag{5.1.23}$$

So we have found two majorations for the errors which concludes the proof of the lemma. □

Proof of Lemma 5.1.12:

Proof. • (I):

$$\begin{aligned}
\text{(I)} =&\mathbf{E}\sum_j 2^{-j}\sum_{l=0}^{p_j} w^2_{j,k_l^*}1_{k_l^*\notin\{\hat{k}_0,\ldots,\hat{k}_{p_j}\}}\\
\leq&\sum_j 2^{-j}(p_j+1)(\mathbf{E}w^4_{j,k_0^*})^{\frac{1}{2}}\mathbf{P}^{\frac{1}{2}}(k_l^*\notin\{\hat{k}_0,\ldots,\hat{k}_{p_j})\\
\leq&\sum_j 2^{(\eta-1)j}(\mathbf{E}w^4_{j,k_0^*})^{\frac{1}{2}}\mathbf{P}^{\frac{1}{2}}(k_0^*\notin\{\hat{k}_0,\ldots,\hat{k}_{p_j})
\end{aligned}$$

using Cauchy-Schwarz inequality. But if X is Gaussian variables with mean m and variance σ^2, we have

$$\mathbf{E}X^4 = 3\sigma^4 + 6m^2\sigma^2 + m^4.$$

So since $w_{j,k_l^*} \sim \mathcal{N}\left(2^{-\alpha j}, \sigma^2 2^{-j_1} + \Delta_j^2\right)$ we have if we set $\sigma_j^2 = \Delta_j^2 + 2^{-j_1}\sigma^2$:

$$\text{(I)} \leq \sum_j 2^{(\eta-1)j}\mathbf{P}^{\frac{1}{2}}(k_l^*\notin\{\hat{k}_0,\ldots,\hat{k}_{p_j})(2\sigma_j^4 + 6\;2^{-2\alpha j}\sigma_j^2 + 2^{-4\alpha j}).$$

- (II):

$$\begin{aligned}
\text{(II)} =&\sum_j 2^{-j}\mathbf{E}\left(\sum_{l=0}^{p_j}(\hat{w}_{j,k_l^*} - w_{j,k_l^*})^2 1_{k_l^*\in\{\hat{k}_0,\ldots,\hat{k}_{p_j}\}}\right)\\
\leq& 2^{-j}(p_j+1)\mathbf{E}(\hat{w}_{j,k_0^*} - w_{j,k_0^*})^2\\
\leq& 2^{(\eta-1)j}\frac{\Delta_j^2\frac{\sigma^4}{n} + \Delta_j^4\frac{\sigma^2}{n}}{(\frac{\sigma^2}{n} + \Delta_j^2)^2}
\end{aligned}$$

using again Cauchy-Scwharz inequality.

- (III):

$$\begin{aligned}
\text{(III)} =& \sum_j 2^{-j}\mathbf{E}\left(\sum_{l=p_j+1}^{2^j-1}\hat{w}^2_{j,k_l^*}1_{k_l^*\in\{\hat{k}_0,\ldots,\hat{k}_{p_j}\}}\right)\\
\leq& \sum_j 2^{-j}\sum_{l>p_j}\mathbf{E}(2^{-\alpha j} + \frac{\Delta_j^2}{\Delta_j^2+\frac{\sigma^2}{n}}(d_{jk} - 2^{-\alpha j})^2 1_{k_l^*\in\{\hat{k}_0,\ldots,\hat{k}_{p_j}\}})\\
\leq& \sum_j 2^{-j}\sum_{l>p_j}\mathbf{P}^{\frac{1}{2}}(k_l^*\in\{\hat{k}_0,\ldots,\hat{k}_{p_j}\})\mathbf{E}^{\frac{1}{2}}\left(\frac{2^{-\alpha j}\frac{\sigma^2}{n}}{\Delta_j^2+\frac{\sigma^2}{n}} + \frac{\Delta_j^2}{\Delta_j^2+\frac{\sigma^2}{n}}\epsilon_{jk}\right)^4
\end{aligned}$$

where we have set $\epsilon_{jk} = d_{jk} - 2^{-\alpha j}$. So

$$\text{(III)} \leq \sum_j 2^{(\eta-1)j}(2^j - 2^{\eta j})\frac{\mathbf{P}^{\frac{1}{2}}(k_n^*\in\{\hat{k}_0,\ldots,\hat{k}_{p_j}\})}{(\Delta_j^2 + 2^{-j_1}\sigma^2)^2}R_j$$

with

$$R_j = (3.2^{-2j_1}\sigma^4\Delta_j^8 + 2^{-4\alpha j}\sigma^8 2^{-4j_1} + 6\sigma^6\Delta_j^5 2^{-2\alpha j}2^{-3j_1})^{1/2}$$

□

Acknowledgements

Many thanks are due to Antoine Ayache from who we were awared of Jaffard's works on sparse wavelet series.

5.2 Estimation of lacunarity parameter

The aim of this part is to provide estimators of the parameters (η, α) of a multifractal function represented as a lacunary wavelet series. We provide empirical estimators and give their asymptotic behavior and distribution.

5.2.1 Rates of convergence for estimators

We recall that in the general mixing model describe in the previous Section, the observed coefficients d_{jk}, $j = 1, \dots, J$, $k = 0, \dots, 2^{j-1}$ are such that

$$d_{jk} \sim z_{jk}\mathcal{N}(2^{-\alpha j}, \frac{\sigma^2}{n}) + (1 - z_{jk})\mathcal{N}(0, \frac{\sigma^2}{n}) \tag{5.2.1}$$

where z_{jk} is an independent Bernoulli random variable such that $P(z_{jk} = 1) = 2^{(\eta-1)j}$. We assume that α is known and give an estimator of η.

An estimator of η is given by considering the mean of a coefficient: since we have

$$\sum_{k=0}^{2^j-1} Ed_{jk} = 2^{(\eta-\alpha)j}$$

let us consider the estimator

$$\hat{\eta}_n = \alpha + \frac{1}{J\log 2} \log \sum_{j=1}^{J} \sum_{k=0}^{2^j-1} d_{jk} \tag{5.2.2}$$

The following theorem describes the asymptotic behavior of this estimator.

Theorem 5.2.1. *Consider a multifractal function characterized in terms of lacunarity wavelet series by a known intensity parameter α and an unknown lacunarity parameter η_0 such that $\eta_0 - \alpha > 0$. For*

$$\kappa = \frac{2^{\eta_0-1-2\alpha} - 2^{2(\eta_0-1\alpha)}}{(1 - 2^{\eta_0-1-2\alpha})(1 - 2^{2(\eta_0-1-\alpha)})} > 0$$

and for $C_J = 2^{(\alpha-\eta_0)J}$ we have

$$\log(2^J)C_J^{-1}\left(\hat{\eta}_n - \eta_0\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \kappa + \sigma^2\right) \tag{5.2.3}$$

With the notations $2^J = n$ the following Central Limit theorem holds:

$$\log(n)n^{\eta_0-\alpha}\left(\hat{\eta}_n - \eta_0\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \kappa + \sigma^2\right) \tag{5.2.4}$$

For the remaining subcase $\eta_0 - \alpha_0 < 0$, the mean of d_{jk} goes to zero. So an estimator based on the mean is not consistent.
The second parameter to be estimated is α, the intensity coefficient. We consider the quadratic mean of the coefficients:

$$Ed_{jk}^2 = \frac{\sigma^2}{n} + 2^{(\eta-1-2\alpha)j}.$$

So define the following estimator:

$$\hat{\alpha}_n = \frac{1}{J\log 2}\left(\log\left[\frac{\sum_{jk} d_{jk}}{\sum_{jk} d_{jk}^2 - \sigma^2}\right]\right) \tag{5.2.5}$$

The following theorem describes the behavior of the estimator:

Theorem 5.2.2. *The estimator $\hat{\alpha}_n$ is a consistent estimator of the intensity coefficient in the case $\eta_0 - 2\alpha_0 > 0$.*

$$\hat{\alpha}_n = \frac{1}{J\log 2}\left(\log\left[\frac{\sum_{jk} d_{jk}}{\sum_{jk} d_{jk}^2 - \sigma^2}\right]\right) \to \alpha_0 \tag{5.2.6}$$

Set $d_J = 2^{(2\alpha_0-\eta_0)J} = n^{2\alpha_0-\eta_0} \to 0$, the asymptotic distribution is given by

$$\sqrt{n^{\eta_0}}\left[d_J\left(\sum_{jk} d_{jk}^2 - \sigma^2\right) - 1\right] \to \mathcal{N}(0,1) \tag{5.2.7}$$

So we have the following Central Limit theorem

$$\log(n)\sqrt{n^{\eta_0}}(\hat{\alpha}_n - \alpha_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \tag{5.2.8}$$

Remark 5.2.3. *The two conditions for estimating the parameters by empirical mean estimators are $\eta_0 > \alpha_0$ and $\eta_0 > 2\alpha_0$. They mean that the true signal must convey enough information in the sense that the number of non zero coefficients has too be large enough for the signal and noise to be differentiated.*

Remark 5.2.4. *The two parameters α and η are deeply linked. The first estimator estimates in fact $\eta - \alpha$ while the second approximates the quantity $\eta - 2\alpha$. Only a computationnal trick enables us to estimate them separately, but the relations between these parameters are worth a closer attention. Perhaps a change in the representation of the model should be appropriate. This problem must be linked with the important issue of the estimation of the index of a multifractionnal process. Such issue is tackled by Benassi, Cohen and Istas in [BBCI99] and [BCI98].*

5.2.2 Proofs

Proof of Theorem 5.2.1:

Proof. For $\eta \in (0,1)$, define $L_n(\eta) = C_J \sum_{j=1}^{J} \sum_k d_{jk}$. Consider the sub-case $\eta_0 - \alpha > 0$ and set $C_J = 2^{(\alpha-\eta_0)J}$. By easy calculations we have

$$\begin{aligned} EL_n(\eta) &= C_J \sum_{j=1}^{J} 2^{(\eta_0-\alpha)J} \\ &\longrightarrow 1 \end{aligned} \tag{5.2.9}$$

Taking into account the independence of the coefficients, the variance of $L_n(\eta)$ is:

$$\mathrm{Var}(L_n(\eta)) = C_J^2\sigma^2 + C_J^2 \sum_{j=1}^{J} 2^{(\eta_0-2\alpha)j}(1 - 2^{(\eta_0-1)j}) \tag{5.2.10}$$

Chebychev's inequality, as well as the calculations (5.2.9) and (5.2.10) leads to the following result

$$L_n(\eta) \xrightarrow{\mathbf{P}} 1 \tag{5.2.11}$$

So consistency of the empirical mean in probability is proven.
Using the Delta method, see for instance van der Vaart and Wellner in [vdVW96], the asymptotic distribution will give the rate of convergence. By straightforward calculations we obtain:

$$\begin{aligned} E\exp(itL_n(\eta)) &= E\exp(itC_J \sum_{j=1}^{J}\sum_k d_{jk}) \\ &= \prod_{j=1}^{J} \left[\exp(-\frac{t^2C_J^2\sigma^2}{2n})(1 + 2^{(\eta_0-1)j}(\exp(itC_J 2^{-\alpha j}) - 1))\right]^{2^j} \\ &= \exp(-\frac{t^2\sigma^2C_J^2}{2}) \prod_{j=1}^{J} \left(1 + 2^{(\eta_0-1)j}(\exp(itC_J 2^{-\alpha j}) - 1)\right)^{2^j} \end{aligned}$$

A Taylor's expansion of order 2 in t gives rise to the following result

$$E\exp(itL_n(\eta)) = \exp(it)\exp(-C_J^2\frac{t^2}{2}(\sigma^2+\kappa)) + O(t^2C_J^2) \tag{5.2.12}$$

where κ is such that

$$\begin{aligned} \kappa &= -\sum_{j=1}^{J} 2^{2(\eta_0-1-\alpha)j} - 2^{(\eta_0-1-2\alpha)j} \\ &= \frac{2^{\eta_0-1-2\alpha} - 2^{2(\eta_0-1-\alpha)}}{(1-2^{2(\eta_0-1-\alpha)})(1-2^{(\eta_0-1-2\alpha)})} \end{aligned}$$

Define the quantity T_n as

$$\begin{aligned} T_n &= \frac{1}{J}\log_2 C_J \sum_{j=1}^{J}\sum_{k=0}^{2^j-1} d_{jk} \\ &= \frac{1}{J}\log_2 C_J + \alpha - \hat{\eta}_n \\ &= \eta_0 - \hat{\eta}_n \end{aligned}$$

Since (5.2.12) shows that

$$C_J^{-1}\left(C_J \sum_{j=1}^{J}\sum_{k=0}^{2^j-1} d_{jk} - 1\right) \xrightarrow{\mathbf{P}} \mathcal{N}\left(0, \sigma^2 + \kappa\right) \tag{5.2.13}$$

the Delta method together with (5.2.13) provides the asymptotic behavior:

$$JC_J^{-1}T_n \xrightarrow{\mathbf{P}} \mathcal{N}\left(0, \sigma^2 + \kappa\right) \tag{5.2.14}$$

Indeed if Φ is Hadamard differentiable at a point θ, provided there exists a sequence r_n such that

$$r_n(X_n - \theta) \xrightarrow{\mathcal{L}} X$$

the following convergence holds

$$r_n(\Phi(X_n) - \Phi(\theta)) \xrightarrow{\mathcal{L}} \Phi^{'}(\theta)X.$$

Applying the last result with $\theta = 0$ and $\Phi(x) = \log(1 + x)$ proves (5.2.14). Finally, the form of the estimator together with the result (5.2.14) proves the statement of theorem 5.2.1. We must keep in mind that the data are dyadic with the correspondence $2^J = n$. This gives a rate of convergence in $\log(n)n^{\alpha-\eta_0}$. □

Proof of Theorem 5.2.2:

Proof. We only consider in this proof the true parameters of the function, so we will write for simplicity reasons $\alpha = \alpha_0$ and $\eta = \eta_0$. Set a scaling coefficient $d_J = n^{2\alpha-\eta} \to 0$ under the assumption that $\eta > 2\alpha$. Our aim is to prove a Central Limit theorem for

$$d_J[\sum_{jk} d_{jk}^2 - \sigma^2] - 1$$

with a proper scaling coefficients.
By some calculations and using the law of the coefficients d_{jk} we obtain the characteristic

function of $\sum_{jk} d_{jk}^2$.

$$\begin{aligned}
&\mathbf{E}\exp(it\sum_{jk} d_{jk}^2) \\
&= \prod_{jk}\left(2^{(\eta-1)j}\left(1-2it\frac{\sigma^2}{n}\right)^{-\frac{1}{2}} + (1-2^{(\eta-1)j})\exp\left(\frac{i2^{-2\alpha j}t}{1-2it\frac{\sigma^2}{n}}\right)\left(1-2it\frac{\sigma^2}{n}\right)^{-\frac{1}{2}}\right) \\
&= \left(1-2it\frac{\sigma^2}{n}\right)^{-\frac{n}{2}}\prod_j\left[1+2^{(\eta-1)j}\left(\exp\left(\frac{i2^{-2\alpha j}t}{1-2it\frac{\sigma^2}{n}}\right)-1\right)\right]^{2^j} \qquad (5.2.15)
\end{aligned}$$

Using a Taylor's expansion up to the second order in t we find the following developments:

$$\begin{aligned}
\left(1-2it\frac{\sigma^2}{n}\right)^{-\frac{n}{2}} &= \exp(it\sigma^2-\frac{t^2\sigma^4}{n})+o(t^2/n) \\
\exp\left(\frac{i2^{-2\alpha j}t}{1-2it\frac{\sigma^2}{n}}\right) &= 1+i2^{-2\alpha j}t-(22^{-2\alpha j}\frac{\sigma^2}{n}+\frac{2^{-4\alpha j}}{2})t^2 \\
\left[1+2^{(\eta-1)j}\left(\exp\left(\frac{i2^{-2\alpha j}t}{1-2it\frac{\sigma^2}{n}}\right)-1\right)\right]^{2^j} &= \exp(2^{(\eta-2\alpha)j}t-t^2\frac{n^{(\eta-4\alpha)j}}{2}).
\end{aligned}$$

As a result we have the following development of the characterisitc function up to the second order

$$\mathbf{E}\exp(itd_J\sum_{jk} d_{jk}^2) = \exp(it(1+d_J\sigma^2)-\frac{t^2}{2n^\eta})$$

As a result, the following expansion up to second order holds:

$$\mathbf{E}\exp(it[d_J(\sum_{jk} d_{jk}^2-1)-1]) = \exp(-t^2\frac{1}{2n^\eta}) \qquad (5.2.16)$$

So with a scaling factor of order $n^{\frac{\eta}{2}}$, (5.2.16) proves that

$$\mathbf{E}\exp(it\sqrt{n^\eta}[d_J(\sum_{jk} d_{jk}^2-1)-1]) = \exp(-\frac{t^2}{2}) \qquad (5.2.17)$$

As a conclusion, (5.2.17) together with Levy's theorem enables us to conclude that

$$\sqrt{n^\eta}[d_J(\sum_{jk} d_{jk}^2-\sigma^2)-1] \to \mathcal{N}(0,1) \qquad (5.2.18)$$

As in the previous proof, the Delta method, with $\Phi(x)=\log(1+x)$ gives the asymptotic distribution of the estimator.

$$\sqrt{n_0^\eta}\log[d_J(\sum_{jk} d_{jk}^2-\sigma^2)] \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \qquad (5.2.19)$$

As a result, we decompose the estimator of α_0 as shown:

$$\begin{aligned}\hat{\alpha}_n &= \frac{1}{J\log 2}\left(\log\left[\frac{\sum_{jk} d_{jk}}{\sum_{jk} d_{jk}^2 - \sigma^2}\right]\right)\\ &= \frac{1}{J\log 2}\left(\log(\sum_{jk} d_{jk}) + \log(\sum_{jk} d_{jk}^2 - \sigma^2)\right)\\ &= \frac{1}{J\log 2}\left(\hat{\eta}_n - \eta_0 + \log(d_J(\sum_{jk} d_{jk}^2 - \sigma^2))\right) + \alpha_0\end{aligned}$$

It implies that

$$\log(n)\sqrt{n_0^\eta}\hat{\alpha}_n - \alpha_0 = \frac{\sqrt{n^{\eta_0}}}{n^{\eta_0-\alpha_0}} n^{\eta_0-\alpha_0}(\hat{\eta}_n - \eta_0) + \sqrt{n^{\eta_0}}\log(d_J(\sum_{jk} d_{jk}^2 - \sigma^2))$$

But $\eta_0 > 2\alpha_0$. As a result the first term in the previous sum goes to zero since it implies that

$$\frac{\sqrt{n^{\eta_0}}}{n^{\eta_0-\alpha_0}} \to 0.$$

So the asymptotic distribution is given by the second term, which proves the statement (5.2.8) of the theorem. □

5.3 Simulations

The simulations have been done using MatLab software. We estimate a multifractal function generated by its wavelet coefficients. We display the results in two graphics. Each is divided into four cases where is shown the true function, the observations, the reconstructed function and, in the last box, the function and its estimator. We point out that the knowledge of the inner structure of the signal (i.e the parameters η and α) provides an efficient reconstruction of the signal. In bad cases, if some bumps are badly allocated, most of the information conveyed by the signal is still preserved. However, for the moment, estimating the parameters of the signal with the EM algorithm does not give significant results. As a matter of fact, if the convergence of the algorithm is fast, the convergence of the estimators to the true values may be slow. One of the key could be the fact that it does not use the whole data - all the wavelet coefficients d_{jk}, $j = 0, \dots, J$, $k = 0, \dots, 2^j - 1$ -, but more or less, only the last line in the array of data - d_{Jk}, $k = 0, \dots, n_1$ -. Moreover we do not have any control about the approximation error done when replacing η_0 and α_0 by their estimators. Such lack of efficiency justifies the research of Section 5.2 and the attempts to find others more efficient estimators.

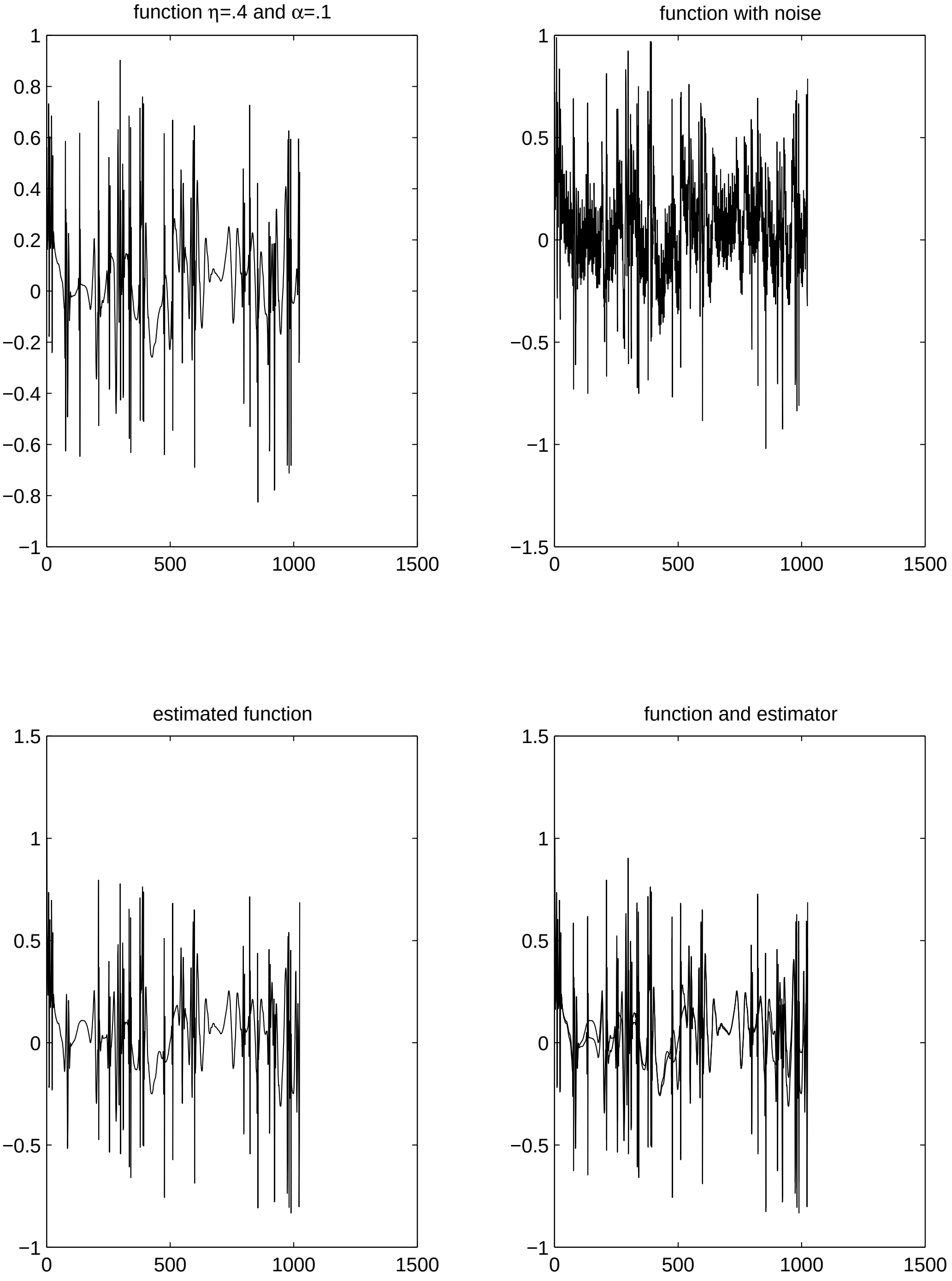

function η=.4 and α=.1
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
0
500
1000
1500
function with noise
1
0.5
0
−0.5
−1
−1.5
0
500
1000
1500
estimated function
1.5
1
0.5
0
−0.5
−1
0
500
1000
1500
function and estimator
1.5
1
0.5
0
−0.5
−1
0
500
1000
1500

Figure 5.4:

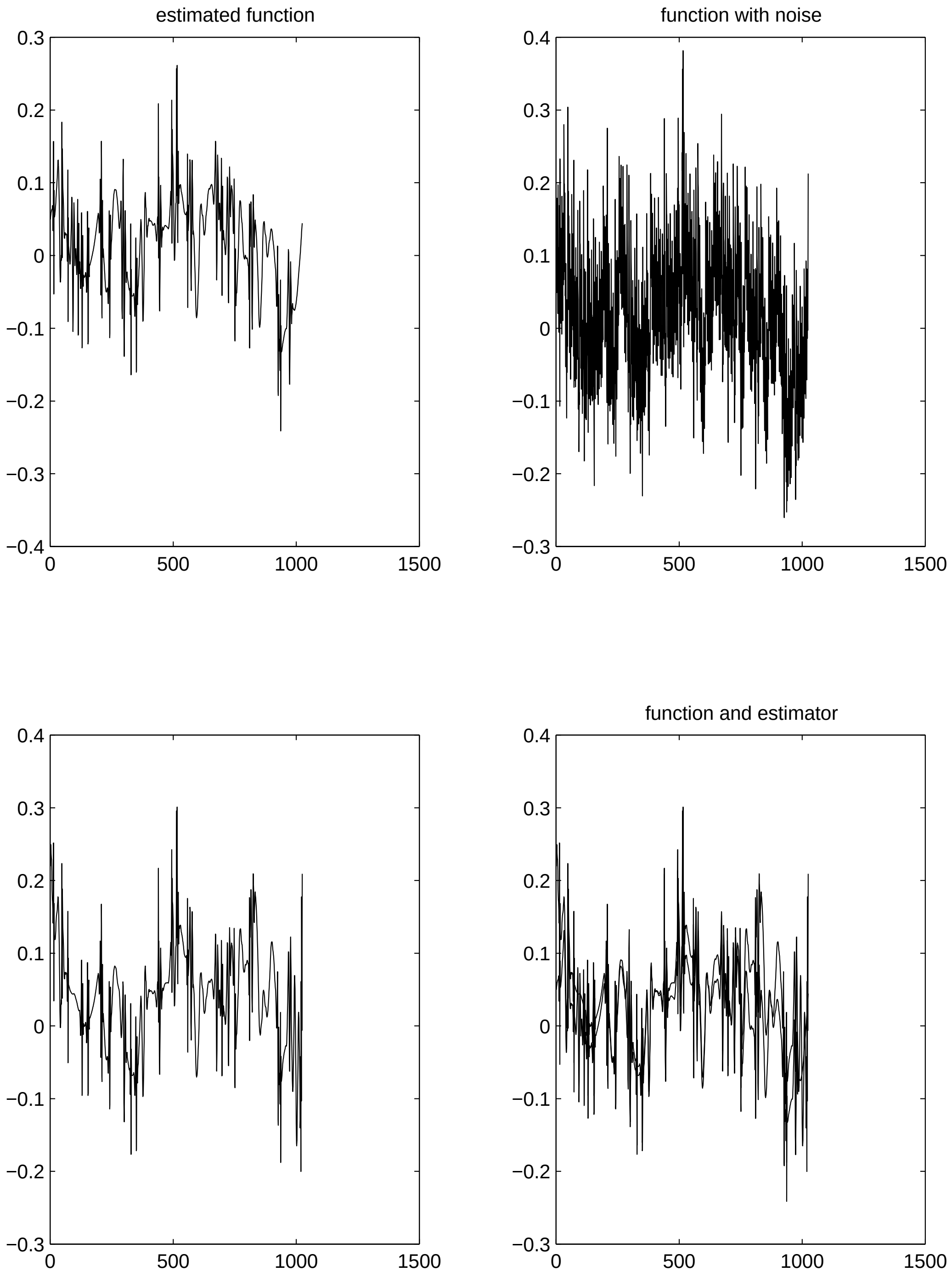

estimated function
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
0
500
1000
1500
function with noise
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
0
500
1000
1500
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
0
500
1000
1500
function and estimator
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
0
500
1000
1500

Figure 5.5:

Bibliography

[ABJM99] A. Arneodo, E. Bacry, S. Jaffard, and J. F. Muzy. Oscillating singularities and fractal functions. In *Spline functions and the theory of wavelets (Montreal, PQ, 1996)*, pages 315–329. Amer. Math. Soc., Providence, RI, 1999.

[ABM99] A. Arneodo, E. Bacry, and J. F. Muzy. The thermodynamics of fractals revisited with wavelets. In *Wavelets in physics*, pages 339–390. Cambridge Univ. Press, Cambridge, 1999.

[ADCL00] C. Angelini, D. De Canditis, and F. Leblanc. Wavelet regression estimation in nonparametric mixed effect models. *preprint*, 2000.

[AGG97] A. Antoniadis, I. Gijbels, and G. Grégoire. Model selection using wavelet decomposition and applications. *Biometrika*, 84(4):751–763, 1997.

[AJ01] J. M. Aubry and S. Jaffard. Random wavelet series. *preprint*, 2001.

[Ale85] K. Alexander. Recent results on central limit theorem for empirical processes, with applications. In *Proceedings of the 45th session of the International Statistical Institute, Vol. 4 (Amsterdam, 1985)*, volume 51, pages No. 25.3, 13, 1985.

[ASS98] F. Abramovich, T. Sapatinas, and B. W. Silverman. Wavelet thresholding via a Bayesian approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(4):725–749, 1998.

[BAF$^+$91] E. Bacry, Al. Arneodo, U. Frisch, Y. Gagne, and E. Hopfinger. Wavelet analysis of fully developed turbulence data and measurement of scaling exponents. In *Turbulence and coherent structures (Grenoble, 1989)*, pages 203–215. Kluwer Acad. Publ., Dordrecht, 1991.

[Bar00] Y. Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4):467–493, 2000.

[BB01] E. Belitser and Levit B. Empirical bayes estimation in gaussian model. *preprint*, 2001.

[BBCI99] A. Benassi, P. Bertrand, S. Cohen, and J. Istas. Identification d'un processus gaussien multifractionnaire avec des ruptures sur la fonction d'échelle. *C. R. Acad. Sci. Paris Sér. I Math.*, 329(5):435–440, 1999.

[BBM99] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[BCI98] A. Benassi, S. Cohen, and J. Istas. Identifying the multifractional function of a Gaussian process. *Statist. Probab. Lett.*, 39(4):337–345, 1998.

[BD93] D. Birkes and Y. Dodge. *Alternative methods of regression.* John Wiley & Sons Inc., New York, 1993. A Wiley-Interscience Publication.

[BDH$^+$85] A. Beck, R. Dudley, M. Hahn, J. Kuelbs, and M. Marcus, editors. *Probability in Banach spaces. V*, Berlin, 1985. Springer-Verlag.

[BG00] E. Belitser and S. Ghosal. Adaptive bayesian inference on the mean of an infinite dimensional normal distribution. *Technical Report, University of Leiden 2000-06*, 2000.

[BIN78] O. V. Besov, V. P. Ilin, and S. M. Nikolskiĭ. *Integral representations of functions and imbedding theorems. Vol. I.* V. H. Winston & Sons, Washington, D.C., 1978. Translated from the Russian, Scripta Series in Mathematics, Edited by Mitchell H. Taibleson.

[Bir83] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete*, 65(2):181–237, 1983.

[BL96] L. Brown and M. Low. Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, 24(6):2384–2398, 1996.

[BLV00] A. Berlinet, F. Liese, and I. Vajda. Necessary and sufficient conditions for consistency of M-estimates in regression models with general errors. *J. Statist. Plann. Inference*, 89(1-2):243–267, 2000.

[BM85] K. E. Basford and G. J. McLachlan. Estimation of allocation rates in a cluster analysis context. *J. Amer. Statist. Assoc.*, 80(390):286–293, 1985.

[BM97] L. Birgé and P. Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[BM98] L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.

[BM00] L. Birgé and P. Massart. Compression algorithm in besov spaces. *Constructive Approximation*, 16:1–36, 2000.

[BM01] L. Birgé and P. Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 2001.

[BMP92] G. Brown, G. Michon, and J. Peyrière. On the multifractal analysis of measures. *J. Stat. Phys.*, 41:909–996, 1992.

[BP90] K. Ball and A. Pajor. The entropy of convex bodies with “few” extreme points. In *Geometry of Banach spaces (Strobl, 1989)*, pages 25–32. Cambridge Univ. Press, Cambridge, 1990.

[BR73] I. Barrodale and F. D. K. Roberts. An improved algorithm for discrete l_1 linear approximation. *SIAM J. Numer. Anal.*, 10:839–848, 1973.

[BS67] M. Š. Birman and M. Z. Solomjak. Piecewise polynomial approximations of functions of classes ${W_p}^{\alpha}$. *Mat. Sb. (N.S.)*, 73 (115):331–355, 1967.

[BS91] AR. Barron and C. Sheu. Approximation of density functions by sequences of exponential families. *Ann. Statist.*, 19(3):1347–1369, 1991.

[Cas00] G. Castellan. Sélection d’histogrammes à l’aide d’un critère de type Akaike. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(8):729–732, 2000.

[CDS99] S. Chen, D. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61 (electronic), 1999.

[Cia82] P. Ciarlet. *Introduction à l’analyse numérique matricielle et à l’optimisation.* Masson, Paris, 1982.

[CK89] D. Cox and E. Koh. A smoothing spline based test of model adequacy in polynomial regression. *Ann. Inst. Statist. Math.*, 41(2):383–400, 1989.

[CKWY88] D. Cox, E. Koh, G. Wahba, and B. Yandell. Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *Ann. Statist.*, 16(1):113–119, 1988.

[CPV98] M. Clyde, G. Parmigiani, and B. Vidakovic. Multiple shrinkage and subset selection in wavelets. *Biometrika*, 85(2):391–401, 1998.

[CW99] H. Chipman and L. Wolfson. Prior elicitation in the wavelet domain. In *Bayesian inference in wavelet-based models*, pages 83–94. Springer, New York, 1999.

[Dau92] I. Daubechies. *Ten lectures on wavelets.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992.

[DeV98] Ronald A. DeVore. Nonlinear approximation. In *Acta numerica, 1998*, pages 51–150. Cambridge Univ. Press, Cambridge, 1998.

[DF97] P. Diaconis and D. A. Freedman. Consistency of Bayes estimates for nonparametric regression: a review. In *Festschrift for Lucien Le Cam*, pages 157–165. Springer, New York, 1997.

[DFR01] S. Darolles, J-P. Florens, and E. Renault. Nonparametric instrumental regression. *preprint*, 2001.

[DH81] R. Dutter and P. Huber. Numerical methods for the nonlinear robust regression problem. *J. Statist. Comput. Simulation*, 13(2):79–113, 1981.

[DJ94] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[DJ95] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90(432):1200–1224, 1995.

[DJ96a] B. Delyon and A. Juditsky. On minimax wavelet estimators. *Appl. Comput. Harmon. Anal.*, 3(3):215–228, 1996.

[DJ96b] D. L. Donoho and I. M. Johnstone. Neo-classical minimax problems, thresholding and adaptive function estimation. *Bernoulli*, 2(1):39–62, 1996.

[DJ98] D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *Ann. Statist.*, 26(3):879–921, 1998.

[DJ99] D. L. Donoho and I. M. Johnstone. Asymptotic minimaxity of wavelet estimators with sampled data. *Statist. Sinica*, 9(1):1–32, 1999.

[DJKP95] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B*, 57(2):301–369, 1995. With discussion and a reply by the authors.

[DJKP96a] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2):508–539, 1996.

[DJKP96b] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2):508–539, 1996.

[DJKP97] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Universal near minimaxity of wavelet shrinkage. In *Festschrift for Lucien Le Cam*, pages 183–218. Springer, New York, 1997.

[DL93] R. A. DeVore and G. G. Lorentz. *Constructive approximation.* Springer-Verlag, Berlin, 1993.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977. With discussion.

[DMA97] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Constr. Approx.*, 13(1):57–98, 1997.

[Dro96] B. Droge. Some comments on cross-validation. In *Statistical theory and computational aspects of smoothing (Semmering, 1994)*, pages 178–199. Physica, Heidelberg, 1996.

[Dro99] B. Droge. Asymptotic optimality of full cross-validation for selecting linear regression models. *Statist. Probab. Lett.*, 44(4):351–357, 1999.

[EP80] S. Yu. Efroĭmovich and M. S. Pinsker. On the problem of asymptotically sufficient statistics. In *Methods of information transmission and processing (Russian)*, pages 55–73, 116. “Nauka”, Moscow, 1980.

[ET92] D. E. Edmunds and H. Triebel. Entropy numbers and approximation numbers in function spaces. II. *Proc. London Math. Soc. (3)*, 64(1):153–169, 1992.

[FP85] U. Frisch and G. Parisi. On the singularity structure of fully developped turbulence; appendix to fully developped turbulence and intermittency. *Proc. Int. Summer School Phys. Enrico Fermi*, (1):84–88, 1985.

[Fre99] D. Freedman. On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Ann. Statist.*, 27(4):1119–1140, 1999.

[Fri95] U. Frisch. *Turbulence.* Cambridge University Press, Cambridge, 1995. The legacy of A. N. Kolmogorov.

[GCP93] Valentine Genon-Catalot and D. Picard. *Éléments de statistique asymptotique.* Springer-Verlag, Paris, 1993.

[GL96] G. K. Golubev and B. Y. Levit. Asymptotically efficient estimation for analytic distributions. *Math. Methods Statist.*, 5(3):357–368, 1996.

[GL01] F. Gamboa and J-M. Loubes. Empirical parametric estimation in multifracatl wavelet models. *preprint*, 2001.

[GLT76] R. Glowinski, J.-L. Lions, and R. Trémolières. *Analyse numérique des inéquations variationnelles. Tome 1.* Dunod, Paris, 1976. Théorie générale premiéres applications, Méthodes Mathématiques de l’Informatique, 5.

[Gro85] P. Groeneboom. Estimating a monotone density. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, pages 539–555, Belmont, CA, 1985. Wadsworth.

[GS94] P. J. Green and B. W. Silverman. *Nonparametric regression and generalized linear models.* Chapman & Hall, London, 1994. A roughness penalty approach.

[Gum58] E. J. Gumbel. *Statistics of extremes.* Columbia University Press, New York, 1958.

[GvdV00] S. Ghosal and A. van der Vaart. On bayesian adaptation. *Preprint*, 2000.

[GW91] C. Gu and G. Wahba. Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Statist. Comput.*, 12(2):383–398, 1991.

[Ham83] F. Hampel. The robustness of some nonparametric procedures. In *A Festschrift for Erich L. Lehmann*, pages 209–238. Wadsworth, Belmont, Calif., 1983.

[HI86] R. Z. Hasminskii and I. A. Ibragimov. Asymptotically efficient nonparametric estimation of functionals of a spectral density function. *Probab. Theory Related Fields*, 73(3):447–461, 1986.

[HKP98] P. Hall, G. Kerkyacharian, and D. Picard. Block threshold rules for curve estimation using kernel and wavelet methods. *Ann. Statist.*, 26(3):922–942, 1998.

[HKP99] P. Hall, G. Kerkyacharian, and D. Picard. On the minimax optimality of block thresholded wavelet estimators. *Statist. Sinica*, 9(1):33–49, 1999.

[HKPT98] W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov. *Wavelets, approximation, and statistical applications.* Springer-Verlag, New York, 1998.

[HL00] S. Huang and Henry H. Lu. Bayesian wavelet shrinkage for nonparametric mixed-effects models. *Statist. Sinica*, 10(4):1021–1040, 2000.

[HL01] S. Huang and Henry H. Lu. Extended Gauss-Markov theorem for nonparametric mixed-effects models. *J. Multivariate Anal.*, 76(2):249–266, 2001.

[HP95] P. Hall and P. Patil. On wavelet methods for estimating smooth functions. *Bernoulli*, 1(1-2):41–58, 1995.

[HP96] P. Hall and P. Patil. On the choice of smoothing parameter, threshold and truncation in nonparametric regression by non-linear wavelet methods. *J. Roy. Statist. Soc. Ser. B*, 58(2):361–377, 1996.

[Hub81] P. Huber. *Robust statistics.* John Wiley & Sons Inc., New York, 1981. Wiley Series in Probability and Mathematical Statistics.

[INK86] I. A. Ibragimov, A. S. Nemirovskiĭ, and R. Z. Khasminskiĭ. Some problems of nonparametric estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.*, 31(3):451–466, 1986.

[Jaf00a] S. Jaffard. Conjecture de Frisch et Parisi et généricité des fonctions multifractales. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(4):265–270, 2000.

[Jaf00b] S. Jaffard. On lacunary wavelet series. *Ann. Appl. Probab.*, 10(1):313–329, 2000.

[JM89] S. Jaffard and Y. Meyer. Les ondelettes. In *Harmonic analysis and partial differential equations (El Escorial, 1987)*, pages 182–192. Springer, Berlin, 1989.

[JS01] I. M. Johnstone and B. W. Silverman. Risk bounds for empirical bayes estimates of sparse sequences, with applications to wavelet smoothing. *preprint*, 2001.

[Kar92] N. Karmarkar. Interior-point methods in optimization. In *ICIAM 91 (Washington, DC, 1991)*, pages 160–181. SIAM, Philadelphia, PA, 1992.

[KCGN98] R. Kass, B. Carlin, A. Gelman, and R. Neal. Markov chain Monte Carlo in practice: a roundtable discussion. *Amer. Statist.*, 52(2):93–100, 1998.

[Koh99] M. Kohler. Nonparametric estimation of piecewise smooth regression functions. *Statist. Probab. Lett.*, 43(1):49–55, 1999.

[Koh00] M. Kohler. Inequalities for uniform deviations of averages from expectations with applications to nonparametric regression. *J. Statist. Plann. Inference*, 89(1-2):1–23, 2000.

[KP93] G. Kerkyacharian and D. Picard. Density estimation by kernel and wavelets methods: optimality of Besov spaces. *Statist. Probab. Lett.*, 18(4):327–336, 1993.

[KT59] A. N. Kolmogorov and V. M. Tihomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspehi Mat. Nauk*, 14(2 (86)):3–86, 1959.

[LC91] L. M. Le Cam. Some recent results in the asymptotic theory of statistical estimation. In *Proceedings of the International Congress of Mathematicians, Vol. I, II (Kyoto, 1990)*, pages 1083–1090, Tokyo, 1991. Math. Soc. Japan.

[Leu82] S. Leurgans. Asymptotic distributions of slope-of-greatest-convex-minorant estimators. *Ann. Statist.*, 10(1):287–296, 1982.

[LN99] G. Lugosi and A. Nobel. Adaptive model selection using empirical complexities. *Ann. Statist.*, 27(6):1830–1864, 1999.

[LT91] M. Ledoux and M. Talagrand. *Probability in Banach spaces.* Springer-Verlag, Berlin, 1991. Isoperimetry and processes.

[LvdG00] J-M. Loubes and S. van de Geer. Adaptive estimation using thresholding type penalties. *Technical Report, University of Leiden 2000-18*, 2000.

[Mal89] S. G. Mallat. Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbf{r})$. *Trans. Amer. Math. Soc.*, 315(1):69–87, 1989.

[Mal98] S. Mallat. *A wavelet tour of signal processing.* Academic Press Inc., San Diego, CA, 1998.

[Man97] B. Mandelbrot. *Fractals and scaling in finance.* Springer-Verlag, New York, 1997. Discontinuity, concentration, risk, Selecta Volume E, With a foreword by R. E. Gomory.

[McL82] G. J. McLachlan. On the bias and variance of some proportion estimators. *Comm. Statist. B—Simulation Comput.*, 11(6):715–726, 1982.

[Mey87] Y. Meyer. Les ondelettes. In *Contributions to nonlinear partial differential equations, Vol. II (Paris, 1985)*, pages 158–171. Longman Sci. Tech., Harlow, 1987.

[MS86] D. L. McLeish and C. G. Small. Likelihood methods for the discrimination problem. *Biometrika*, 73(2):397–403, 1986.

[Mum97] D. Mumford. Pattern theory: a unifying perspective. In *Fields Medallists' lectures*, pages 226–261. World Sci. Publishing, River Edge, NJ, 1997.

[MV99] P. Müller and B. Vidakovic. MCMC methods in wavelet shrinkage: non-equally spaced regression, density and spectral density estimation. In *Bayesian inference in wavelet-based models*, pages 187–202. Springer, New York, 1999.

[MvdG97] E. Mammen and S. van de Geer. Penalized quasi-likelihood estimation in partial linear models. *Ann. Statist.*, 25(3):1014–1035, 1997.

[NP86] Per Nilsson and Jaak Peetre. On the K functional between L^1 and L^2 and some other K functionals. *J. Approx. Theory*, 48(3):322–327, 1986.

[Nus96] M. Nussbaum. Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.*, 24(6):2399–2430, 1996.

[Pee70] J. Peetre. A new approach in interpolation spaces. *Studia Math.*, 34:23–42, 1970.

[Pin80] M. S. Pinsker. Optimal filtration of square-integrable signals in Gaussian noise. *Problemy Peredachi Informatsii*, 16(2):52–68, 1980.

[Pol84] D. Pollard. *Convergence of stochastic processes.* Springer-Verlag, New York, 1984.

[PT00] D. Picard and K. Tribouley. Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.*, 28(1):298–335, 2000.

[RCRB99] R. H. Riedi, M. S. Crouse, V. J. Ribeiro, and R. G. Baraniuk. A multifractal wavelet model with application to network traffic. *IEEE Trans. Inform. Theory*, 45(3):992–1018, 1999.

[Roc97] R. T. Rockafellar. *Convex analysis.* Princeton University Press, Princeton, NJ, 1997. Reprint of the 1970 original, Princeton Paperbacks.

[ROF92] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

[Rou01] F. Roueff. *Ph. D. Dissertation*, 2001.

[RV99] F. Ruggeri and B. Vidakovic. A Bayesian decision-theoretic approach to the choice of thresholding parameter. *Statist. Sinica*, 9(1):183–197, 1999.

[Sil82] B. W. Silverman. On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.*, 10(3):795–810, 1982.

[Sil85] B. W. Silverman. Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *J. Roy. Statist. Soc. Ser. B*, 47(1):1–52, 1985. With discussion.

[SM98] J-L. Starck and F. Murtagh. Automatic noise estimation from the multiresolution support. *Publications of the Astronomical Society of Pacific*, 110:193–199, 1998.

[SS95] J. Stander and B. W. Silverman. Minimax estimation of linear functionals, particularly in nonparametric regression and positron emission tomography. *Comput. Statist.*, 10(3):259–283, 1995.

[SS99] A. Bruce S. Sardy, P. Tseng. Robust wavelet denoising. *preprint*, 1999.

[Sto90] C. Stone. Large-sample inference for log-spline models. *Ann. Statist.*, 18(2):717–741, 1990.

[Vaj99] I. Vajda. On consistency of M-estimators in models with a linear substructure. I, II. *Discuss. Math. Algebra Stochastic Methods*, 19(2):355–373, 375–392, 1999. International Conference on Statistical Inference (Ł agów, 1998).

[Vap00] V. Vapnik. *The nature of statistical learning theory.* Springer-Verlag, New York, second edition, 2000.

[VC81] V. N. Vapnik and A. Ya. Chervonenkis. Necessary and sufficient conditions for the uniform convergence of empirical means to their true values. *Teor. Veroyatnost. i Primenen.*, 26(3):543–563, 1981.

[vdG90] S. van de Geer. Estimating a regression function. *Ann. Statist.*, 18(2):907–924, 1990.

[vdG00] S. van de Geer. *Applications of empirical process theory.* Cambridge University Press, Cambridge, 2000.

[vdG01] S. van de Geer. Adaptive estimation with complexity penalties. *preprint*, 2001.

[vdGW96] S. van de Geer and M. Wegkamp. Consistency for the least squares estimator in nonparametric regression. *Ann. Statist.*, 24(6):2513–2523, 1996.

[vdVW96] A. van der Vaart and J. Wellner. *Weak convergence and empirical processes.* Springer-Verlag, New York, 1996. With applications to statistics.

[Wah85] G. Wahba. A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.*, 13(4):1378–1402, 1985.

[Wah90] G. Wahba. *Spline models for observational data.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

[Weg] M. Wegkamp. Entropy methods in statistical estimation. *Ph. D Dissertation.*

[Wil91] D. Williams. *Probability with martingales.* Cambridge University Press, Cambridge, 1991.

[XS98] L. Wasserman X. Shen. Rates of convergence of posterior distributions. *preprint*, 1998.

[XW96] D. Xiang and G. Wahba. A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statist. Sinica*, 6(3):675–692, 1996.

Auteur : Jean-Michel Loubes
Titre : Estimation non paramétrique par M-estimateurs pénalisés
Date et lieu de soutenance : Le 7 décembre 2001 à l'université Paul Sabatier (Toulouse III).

Résumé

Les M-estimateurs pénalisés jouent un rôle clef en statistique non paramétrique. Définis comme réalisant le minimum d'une fonction de perte sous certaines contraintes, ils ont été étudiés d'un point de vue théorique par de nombreux auteurs et trouvent des applications en pratique dans les domaines les plus divers comme la physique ou encore l'économie. Dans ce travail, nous nous sommes intéressés à décrire le comportement asymptotique de ces estimateurs à partir de la complexité de l'espace où a lieu la minimisation, mesurée en terme d'entropie.

Cet objectif nous a conduit, dans un premier temps à une double approche. D'une part, nous avons étudié des estimateurs obtenus par projection sur des bases, notamment des bases d'ondelettes et dont les coefficients ont été soit lissés, soit seuillés. D'autre part cette méthodologie nous a permis de prouver la convergence d'estimateurs Bayésiens. Dans un second temps, nous nous sommes attachés à dégager des procédures d'estimation adaptative qui rendent possible l'estimation d'une fonction, sans faire d'hypothèses a priori sur sa régularité tout en obtenant un estimateur qui converge à la vitesse optimale au sens minimax.

Enfin, nous avons étendu ces résultats à des fonctions très irrégulières dont la régularité varie rapidement et n'a de sens que de façon locale: les fonctions multifractales. Les méthodes précédentes permettent de construire un estimateur de telles fonctions qui échappent aux méthodes d'estimation usuelles.

Mots clés : Estimation non paramétrique - Processus empiriques - Inégalités de concentration - Ondelettes - Estimation Bayésienne - Choix de modèles - Formalisme Multifractal.
Discipline : Mathématiques, Statistique.

Laboratoire de Statistique et Probabilités
UFR MIG - UMR CNRS 5883
Université Paul SABATIER, Toulouse III