Institut de Mathématiques de Toulouse
Institut National des Sciences Appliquées de Toulouse

THÈSE

Présentée pour obtenir le grade de

DOCTEUR EN SCIENCES DE L'INSA DE TOULOUSE

par

Matthieu LERASLE

"Rééchantillonnage et sélection de modèles optimale pour l'estimation de la densité de variables indépendantes ou mélangeantes"

Soutenue publiquement le **25 juin 2009** devant la commission d'examen.

M. Sylvain	**ARLOT**	CNRS	Examinateur
M. Fabrice	**GAMBOA**	Université Toulouse	Examinateur
Mme. Béatrice	**LAURENT**	INSA Toulouse	Directrice
M. Pascal	**MASSART**	Université Paris XI	Président
Mme. Clémentine	**PRIEUR**	Université de Grenoble	Directrice

Rapporteurs :

Mme Florence	**MERLEVÈDE**	Université Marne-la-Vallée
M. Aad	**VAN DER VAART**	Vrije Universiteit Amsterdam

Remerciements

Je tiens d'abord à remercier chaleureusement Béatrice et Clémentine. Cette thèse doit énormément à leurs précieux conseils et leurs nombreuses idées, je dois plus encore à la confiance et au soutien qu'elles m'ont apportés dans les moments plus difficiles de la première année.

Je suis très honoré que Aad van der Vaart et Florence Merlevède aient accepté de rapporter mon travail. Je remercie également Pascal Massart, Sylvain Arlot et Fabrice Gamboa qui ont bien voulu les accompagner dans ce jury.

Je tiens à remercier toute l'équipe du GMM de l'INSA de Toulouse, les enseignants chercheurs et le personnel administratif. Je fais un clin d'oeil particulier a mes co-bureau 135, Jérémie, Cynthia, Elie et Erwan, à Aldéric et son “pot belge”, à mes coach de tutorat Sandrine et Olivier, à Philippe P., le héros du café et des chocolats et à Jean Louis et ses innombrables anecdotes. Merci aussi aux 111, Michel, Nolwen et re-Erwan.

Je ne me serais probablement pas intéressé aux statistiques sans les cours passionnés de Philippe Berthet et les Td de son coéquipier à l'université de Rennes Florent Malrieu. Je suis également reconnaissant aux professeurs de l'université d'Orsay, particulièrement Elisabeth Gassiat et Pascal Massart, la richesse de leurs cours et la pertinence de leurs conseils m'ont lancé dans les meilleures conditions dans ces trois années.

Je voudrais remercier Adrien et Sylvain pour leurs idées enthousiasmantes sur les statistiques en général et la sélection de modèle en particulier, merci encore d'avoir relu la thèse et d'avoir permis de tant l'améliorer.

Merci à tous mes potes, surtout les meilleurs, les depuis Rennes, Charles, Jeff, Pacome, Eric, Julie et Juju, ceux d'avant, Yogi, David, Nico G., Jenni, Virginie. Merci à ceux qui sont un peu encore plus là Pierre, Nico C, Guigui et Adrien.

Merci à toute ma famille, qui est venue en nombre aujourd'hui, et surtout à mon papa, à ma môman, à mon frère Maxime et à mes deux soeurs Caroline et Marie, que j'aime très fort.

Enfin, j'embrasse ma princesse, celle avec qui ma vie est belle, celle que j'suis avec ho-ho-ho!!

Résumé

Le principal objectif de cette thèse est d'étudier deux méthodes de calibration automatique de la pénalité pour la sélection de modèle. L'avantage de ces méthodes est double, d'une part, elles sont toujours implémentables, elles ont même souvent été utilisées dans des problèmes pratiques avec succès, d'autre part, elles sont optimales puisqu'elles permettent de sélectionner asymptotiquement le meilleur modèle.

Il existe d'autres méthodes de pénalisation calculables en pratique, quand les données sont indépendantes. Néanmoins, en dehors des collections de modèles très réguliers, ces pénalités sont très pessimistes, voire dépendent de constantes inconnues comme la norme sup de la densité. De plus, quand on veut utiliser les preuves classiques pour des données mélangeantes, les pénalités que l'on obtient dépendent toujours de constantes inconnues de l'utilisateur (voir le chapitre 3).

Le chapitre 2 étudie l'heuristique de pente et les pénalités par rééchantillonnage dans le cas de données indépendantes. On donne une condition suffisante pour que l'heuristique de la pente soit optimale, en utilisant l'inégalité de concentration de Talagrand pour le supremum du processus empirique. On étudie aussi l'approximation du processus empirique par sa version rééchantillonnée et on en déduit que la même condition suffit à garantir l'optimalité des méthodes par rééchantillonnage.

Le chapitre 3 est consacré à l'étude de pénalités classiques quand les observations sont mélangeantes. On montre des inégalités oracles et l'adaptativité de l'estimateur sélectionné à la régularité de la densité. La pénalité dépend des coefficients de mélange qui peuvent parfois être évalués.

Le chapitre 4 étend les résultats du chapitre 2 au cas de données mélangeantes. On montre ainsi que les méthodes de la pente et bootstrap sont également optimales dans ce cas, sous le même type de conditions. Ces nouvelles pénalités sont toujours calculables en pratique et le modèle sélectionné est asymptotiquement un oracle, ce qui améliore beaucoup les résultats du chapitre 3.

Le chapitre 5 traite du problème des régions de confiance adaptatives. Contrairement au cas de l'estimation, cette adaptation n'est que très rarement possible. Quand elle l'est, nous construisons des régions adaptatives. En particulier, on améliore quelques résultats de concentration du chapitre 2 lorsque les données sont à valeurs réelles, notamment ceux des U-statistiques.

Contents

Chapter 1

Introduction

Le hasard est le plus grand romancier du monde; pour être fécond, il n'y a qu'à l'étudier.
Honoré de Balzac

Cette thèse présente un ensemble de contributions au problème de sélection de modèles optimale ainsi qu'aux méthodes de sélection de modèles quand les observations ne sont pas indépendantes mais sont seulement supposées mélangeantes. Le chapitre 2 concerne le problème de sélection optimale dans le modèle de densité, nous commençons notre étude de la sélection de modèles en milieu mélangeant au chapitre 3, le chapitre 4 améliore les résultats du chapitre 3 en apportant les idées de sélection optimale pour des données mélangeantes, il constitue l'apport le plus conséquent de la thèse tant du point de vue théorique que pratique. Le chapitre 5 porte sur le problème des régions de confiance adaptatives pour la densité. Les chapitres 2, 4 et 5 correspondent chacun à un article soumis pour publication. Le contenu du chapitre 3 est un article accepté pour publication dans la revue *Mathematical Methods of Statistics.*

Les statistiques non paramétriques étudient des quantités s sur lesquelles on ne dispose que de très peu d'informations a priori, par exemple des fonctions. Une des principales difficultés est de définir un modèle S_m dans lequel choisir, à partir des observations, un estimateur $\tilde{s}$ de s. S_m doit être suffisamment riche pour contenir au moins un "bon" estimateur de s, mais en même temps pas trop complexe pour limiter le risque de sélectionner un mauvais estimateur. Un bon modèle optimise ces deux contraintes et est inconnu en pratique.

Pour remédier à cette difficulté, les méthodes de sélection de modèles proposent de partir d'une collection d'ensembles $(S_m)_{m\in\mathcal{M}_n}$ et de choisir à partir des observations le mieux adapté au paramètre inconnu s. Comme nous allons le voir dans cette thèse, les procédures de sélection commencent à être bien comprises théoriquement et il existe des moyens pratiques d'obtenir de bonnes estimées des procédures idéales. Dans le problème de l'estimation de densité, nous avons d'abord compris comment sélectionner un modèle de façon optimale lorsque les observations sont indépendantes. Nous avons ensuite relâché cette hypothèse et étendu au cadre de données mélangeantes les résultats de sélection classiques, ayant pour corollaire l'adaptativité des estimateurs. Nous avons également obtenu des résultats de sélection optimale pour des données mélangeantes. Le dernier chapitre est consacré au problème des régions de confiance adaptatives.

1.1 Le problème de l'estimation de la densité

Nous travaillons dans le cadre de l'estimation de la densité avec perte L^2. Les observations sont des variables aléatoires $X_1, ..., X_n$ de même loi P à valeurs dans un espace mesurable $(\mathbb{X}, \mathcal{X})$. Étant donnée une mesure de référence μ sur $(\mathbb{X}, \mathcal{X})$, on s'intéresse à l'estimation de la densité s de P par rapport à μ. Nous supposons que s appartient à $L^2(\mu)$, l'espace des fonctions réelles, définies sur $\mathbb{X}$ et de carré intégrable

$$L^2(\mu) = \left\{ t : \mathbb{X} \to \mathbb{R},\ \int_{\mathbb{X}} t^2(x) d\mu(x) < \infty \right\}.$$

On note respectivement $\|.\|$ et $\langle .,. \rangle$ la norme et le produit scalaire associés à cet espace. Rappelons qu'ils sont définis pour tout t et t' de $L^2(\mu)$ par

$$\|t\| = \sqrt{\int_{\mathbb{X}} t^2(x) d\mu(x)},\ \langle t, t' \rangle = \int_{\mathbb{X}} t(x) t'(x) d\mu(x).$$

Désignons par X une copie indépendante de X_1, indépendante des observations $X_1, ..., X_n$ et, pour toute fonction t de $L^2(\mu)$, définissons

$$Pt = \mathbb{E}\left(t(X)\right) = \int_{\mathbb{X}} t(x) s(x) d\mu(x) = \langle t, s \rangle.$$

s minimise alors sur $L^2(\mu)$ le critère suivant

$$\|t - s\|^2 - \|s\|^2 = \|t\|^2 - 2\langle t, s \rangle = PQ(t)$$

où $Q : L^2(\mu) \to L^1(P)$, $t \mapsto \|t\|^2 - 2t$. L'estimation de la densité est donc un cas particulier de problèmes de M-estimation. Ces problèmes sont étudiés via la théorie du processus empirique (voir Dudley [29], Pollard [56], Ledoux & Talagrand [48], van der Vaart & Wellner [69], van der Vaart [68] ou van de Geer [67] pour une introduction à cette théorie). Disons juste qu'il s'agit de décrire la convergence uniforme sur des classes de fonctions $\mathcal{F}$ du processus empirique P_n vers P. Rappelons que P_n est défini pour toute fonction t de $L^2(\mu)$ par

$$P_n t = \frac{1}{n} \sum_{i=1}^{n} t(X_i).$$

Pour définir un M-estimateur de s, on se donne un modèle S_m, c'est-à-dire un sous-ensemble de $L^2(\mu)$ et on minimise sur S_m la version empirique du critère des moindres carrés. On obtient

$$\hat{s}_m = \arg\min_{t \in S_m} P_n Q(t) = \arg\min_{t \in S_m} \|t\|^2 - \frac{2}{n} \sum_{i=1}^{n} t(X_i).$$

Pour certains sous-ensembles S_m, ce problème de minimisation peut être très difficile en pratique et l'estimateur $\hat{s}_m$ peut même ne pas être calculable. Pour éviter ce problème, nous prendrons toujours pour S_m un sous-espace vectoriel de $L^2(\mu)$. On vérifie alors que $\hat{s}_m$ est l'estimateur par projection de s sur S_m défini sur une base orthonormée $(\psi_\lambda)_{\lambda \in m}$ de S_m par

$$\hat{s}_m = \sum_{\lambda \in m} (P_n \psi_\lambda) \psi_\lambda.$$

Il existe d'autres méthodes classiques d'estimation de la densité, comme l'estimation par maximum de vraisemblance ou les méthodes à noyaux (voir par exemple Tsybakov [66]), nous ne les aborderons pas ici.
La fonction de perte utilisée en M-estimation est la fonction excès de risque définie par

$$l(\hat{s}_m, s) = PQ(\hat{s}_m) - PQ(s).$$

Cette perte est aléatoire et vaut, avec notre critère

$$l(\hat{s}_m, s) = \|s - \hat{s}_m\|^2.$$

C'est la perte L^2 classique. Là encore, nous n'avons considéré que ce risque alors que d'autres fonctions ont été étudiées, par exemple les pertes L^q, pour $1 \leq q < \infty$ (voir par exemple Donoho *et.al* [27] et les références associées) ou L^∞ (Giné & Nickl [37]).
Le risque de $\hat{s}_m$ est décomposé grâce à l'égalité de Pythagore. On introduit la projection orthogonale s_m de s sur S_m, on a alors

$$\|s - \hat{s}_m\|^2 = \|s - s_m\|^2 + \|s_m - \hat{s}_m\|^2. \tag{1.1}$$

Le premier terme $\|s - s_m\|^2$ est une erreur de modélisation appelée généralement biais du modèle S_m, elle est incompressible, même si on a de nouvelles observations. Le second terme $\|s_m - \hat{s}_m\|^2$ est une erreur d'estimation appelée terme de variance. La théorie des probabilités permet d'avoir une bonne compréhension théorique du terme de variance. En effet, notons $B_m = \{t \in S_m;\ \|t\| \leq 1\}$ et $\nu_n = P_n - P$ le processus empirique recentré. On montre par l'inégalité de Cauchy-Schwarz que

$$\|s_m - \hat{s}_m\|^2 = \left(\sup_{t \in B_m} \nu_n t\right)^2.$$

Le phénomène de concentration de la mesure (voir Ledoux & Talagrand [48], Ledoux [47] ou Massart [55]) permet de relier cette variable aléatoire à son espérance. Le résultat fondamental est l'inégalité de concentration du supremum du processus empirique de Talagrand. Elle a d'abord été prouvée par des méthodes d'isopérimétrie (voir Talagrand [65] pour un point de vue d'ensemble sur l'isopérimétrie) puis redémontrée par Ledoux [46] par la méthode d'entropie. Bousquet [18] a finalement utilisé la méthode d'entropie pour obtenir les constantes optimales dans cette inégalité.
Le terme de variance dans l'inégalité (1.1) est donc bien contrôlé en fonction de S_m, en particulier, plus S_m est grand, plus il se dégrade. À l'inverse, le terme de biais décroît avec la complexité de S_m. Le meilleur modèle optimise ces deux contraintes mais il est inconnu en pratique, c'est pourquoi les méthodes de sélection de modèles ont été introduites.

Sélection de modèles

L'histoire de la sélection de modèles remonte au moins aux travaux d'Akaike [1], [2] et Mallows [53]. Leur approche a été généralisée récemment par Birgé & Massart [15] et Barron, Birgé & Massart [10]. On pourra trouver une introduction beaucoup plus complète de cette théorie dans le livre de Massart [55]. Nous nous contentons

ici de rappeler quelques résultats récents et utiles dans la suite de la thèse.
L'idée de départ est que l'équilibre entre les termes de biais et de variances dans (1.1) est réalisé pour des modèles dépendant de propriétés inconnues de s. On ne peut donc pas choisir de modèles convenablement adaptés à s en l'absence d'hypothèses restrictives. En revanche, il est souvent possible de construire des collections de modèles $(S_m)_{m\in\mathcal{M}_n}$, souvent des espaces de dimension finie d_m, et les estimateurs des moindres carrés associés $(\hat{s}_m)_{m\in\mathcal{M}_n}$ de façon à ce qu'au moins l'un d'eux soit optimal. Le but est alors de choisir dans la collection $(\hat{s}_m)_{m\in\mathcal{M}_n}$, le meilleur estimateur de s. Formellement, on veut déterminer $\hat{m}$ dans $\mathcal{M}_n$ pour que l'estimateur final $\tilde{s} = \hat{s}_{\hat{m}}$ satisfasse une inégalité oracle

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

Birgé & Massart [15] et Barron, Birgé & Massart [10] voulaient construire des estimateurs adaptatifs à la régularité de s et ont fait le lien entre adaptativité et sélection de modèles. Rappelons brièvement la définition d'un estimateur minimax et d'un estimateur adaptatif. Pour un sous-ensemble $\mathcal{F}$ de $L^2(\mu)$, on dit que $\tilde{s}$ est minimax sur $\mathcal{F}$ quand il existe une constante C telle que

$$\sup_{s\in\mathcal{F}} \mathbb{E}\left(\|s-\tilde{s}\|^2\right) \le C \inf_{\hat{s}} \sup_{s\in\mathcal{F}} \mathbb{E}\left(\|s-\hat{s}\|^2\right).$$

L'infimum étant pris sur l'ensemble des estimateurs de s. On dit que $\tilde{s}$ est adaptatif sur une collection $(\mathcal{F}_t)_{t\in T}$ de sous-ensembles de $L^2(\mu)$ s'il est minimax sur chaque espace $\mathcal{F}_t$, c'est-à-dire que, pour tout t de T, il existe une constante C_t telle

$$\sup_{s\in\mathcal{F}_t} \mathbb{E}\left(\|s-\tilde{s}\|^2\right) \le C_t \inf_{\hat{s}} \sup_{s\in\mathcal{F}_t} \mathbb{E}\left(\|s-\hat{s}\|^2\right).$$

Nous renvoyons au livre de Tsybakov [66] pour une présentation plus complète du principe du minimax et de l'adaptativité ainsi que pour des résultats classiques de vitesse de convergence. En estimation de la densité, les classes d'intérêt sont souvent les classes de fonctions régulières, où la régularité est mesurée par les semi-normes de Hölder, de Sobolev ou de Besov. Le lien entre sélection de modèles et adaptativité repose alors sur la théorie de l'approximation développée par exemple dans le livre de Devore & Lorentz [26]. Celle-ci permet de déterminer des espaces de dimension finie bien adaptés aux différentes mesures de régularité, c'est-à-dire pour lesquels le biais est bien majoré. On obtient alors la preuve qu'un estimateur est adaptatif à partir d'une inégalité oracle de la forme

$$\mathbb{E}\left(\|\tilde{s}-s\|^2\right) \le C \inf_{m\in\mathcal{M}_n} \left\{\|s_m - s\|^2 + C_m\right\}, \tag{1.3}$$

où s_m est la projection orthogonale de s sur S_m et C_m est une borne supérieure de $\mathbb{E}\left(\|s_m - \hat{s}_m\|^2\right)$ qui mesure en quelque sorte la complexité de S_m vis-à-vis de l'estimation de s. Cette borne est souvent de la forme d_m/n. L'inégalité (1.3) est plus facile à obtenir que l'inégalité (1.2). Dans la suite, nous appellerons problème de sélection de modèles classique la recherche d'inégalités de la forme (1.3) et problème de sélection de modèles optimal la recherche de (1.2) Nous reviendrons sur cette théorie car les chapitres 2, 3 et 4 de la thèse y sont consacrés.

Régions de confiance pour s

Dans cette partie, α désigne un réel de $(0,1)$ et $\mathcal{F}$ un sous-ensemble de $L^2(\mu)$. Une région de confiance de niveau de confiance $1-\alpha$ sur $\mathcal{F}$ est un ensemble $\tilde{C}$ de fonctions de $L^2(\mu)$ mesurable par rapport à $\sigma(X_1,...X_n)$, satisfaisant la propriété suivante

$$\forall s \in \mathcal{F},\ \mathbb{P}(s \in \tilde{C}) \geq 1-\alpha, \tag{1.4}$$

La difficulté est de déterminer des régions aussi petites que possible. Nous mesurerons la taille d'une région de confiance par son diamètre L^2 défini par

$$\Delta(\tilde{C}) = \sup_{t,t' \in \tilde{C}} \|t-t'\|^2.$$

Comme dans le cadre de l'estimation, on peut définir des notions d'optimalité minimax et d'adaptativité pour des régions de confiance. On dit que l'ensemble $\tilde{C}$ est minimax sur $\mathcal{F}$ s'il vérifie la propriété suivante

$$\sup_{s\in\mathcal{F}} \Delta(\tilde{C}) \leq C \inf_{\hat{C}} \sup_{s\in\mathcal{F}} \Delta(\hat{C}) \tag{1.5}$$

L'infinimum dans (1.5) étant pris parmi les régions de confiance $\hat{C}$ de niveau de confiance $1-\alpha$ sur $\mathcal{F}$. Comme dans le cadre de l'estimation, cette approche présente l'inconvénient de devoir choisir un espace $\mathcal{F}$ a priori pour construire $\tilde{C}$, c'est pourquoi on préfère montrer une propriété d'adaptativité. On se donne une collection $(\mathcal{F}_t)_{t\in T}$ et on note $\mathcal{F} = \cup_{t\in T}\mathcal{F}_t$. On dit que $\tilde{C}$ est adaptatif sur $(\mathcal{F}_t)_{t\in T}$ s'il vérifie (1.4) et si, pour tout t de T, il existe une constante C_t telle que

$$\sup_{s\in\mathcal{F}_t} \Delta(\tilde{C}) \leq C_t \inf_{\hat{C}} \sup_{s\in\mathcal{F}_t} \Delta(\hat{C}) \tag{1.6}$$

L'infimum dans (1.6) étant pris sur les ensembles $\hat{C}$ satisfaisant la propriété

$$\forall s \in \mathcal{F}_t,\ \mathbb{P}\left(s \in \hat{C}\right) \geq 1-\alpha.$$

Contrairement au cas de l'estimation, l'adaptation pour les régions de confiance peut s'avérer impossible. Low [52] a montré que, pour l'estimation de la densité en un point, les données n'aidaient pas à réduire la taille des régions de confiance, par conséquent l'adaptativité est impossible pour l'estimation ponctuelle de la densité. Cette difficulté est liée à la possibilité de tester l'appartenance de s à une région de l'espace (voir par exemple les articles de Baraud [8], Juditsky & Lambert Lacroix [43] et de Robins & Van der Vaart [61]), ainsi que le chapitre 5). Nous y reviendrons dans la section 4.

Liens avec l'estimation adaptative

Le problème des régions de confiance adaptatives est lié aussi à l'apprentissage de la vitesse de convergence des estimateurs adaptatifs. En effet, comme le remarquent Hoffmann & Lepskii [38], la propriété d'adaptation ne donne pas en pratique de borne supérieure sur la vitesse de convergence de l'estimateur (car le meilleur espace $\mathcal{F}_t$ reste inconnu). Un moyen d'obtenir cette information serait, partant d'un estimateur adaptatif $\tilde{s}$, de fournir une majoration de son risque $\tilde{R}$ (l'ensemble

$\tilde{C} = \{t \in L^2(\mu),\ \|t - \tilde{s}\|^2 \leq \tilde{R}\}$ serait alors une région de confiance pour s), de façon à ce que

$$\forall t \in T,\ \forall s \in \mathcal{F}_t,\ \tilde{R} \leq C \inf_{\hat{s}} \sup_{s \in \mathcal{F}_t} \mathbb{E}\left(\|s - \hat{s}\|^2\right).$$

Comme la taille des régions de confiance est minorée par la vitesse minimax d'estimation sur cette classe, nous obtiendrions alors l'adaptativité de la région de confiance $\tilde{C}$. D'après la section précédente, cette propriété ne peut être garantie que pour certaines collections $(\mathcal{F}_t)_{t \in T}$. On ne peut donc pas connaître en pratique, sauf pour ces collections particulières, la vitesse de convergence réelle d'un estimateur adaptatif. Le pendant de cette remarque dans la théorie des tests est qu'on ne peut apprendre la régularité de s à partir des données (voir Ingster [40, 41, 42]).

1.2 Sélection de modèles par critères pénalisés

Nous nous intéressons ici au problème de la sélection d'un modèle efficient où nous cherchons $\hat{m}$ tel que le risque quadratique de l'estimateur final $\tilde{s} = \hat{s}_{\hat{m}}$ soit aussi faible que possible, c'est-à-dire qu'il se compare à celui de l'oracle $\inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2$. Birgé & Massart [15] et Barron, Birgé & Massart [10] proposent pour cela de définir une fonction pen sur $\mathcal{M}_n$ à valeurs réelles et de sélectionner le modèle $\hat{m}$ défini par

$$\hat{m} = \arg \min_{m \in \mathcal{M}_n} P_n Q(\hat{s}_m) + \mathrm{pen}(m) \text{ où } Q(t) = \|t\|^2 - 2t.$$

La raison de ce choix étant que l'oracle minimise le critère

$$PQ(\hat{s}_m) = P_n Q(\hat{s}_m) - \nu_n Q(\hat{s}_m).$$

Comme le remarque Arlot [5], la pénalité idéale est donc donnée par

$$\mathrm{pen}_{id}(m) = -\nu_n Q(\hat{s}_m) = 2\nu_n(\hat{s}_m) = 2\|s_m - \hat{s}_m\|^2 + 2\nu_n(s_m),$$

s_m étant la projection orthogonale de s sur S_m. Le problème est donc de bien estimer cette pénalité idéale.

Approche "classique"

Birgé & Massart [15] puis Barron, Birgé & Massart [10] proposent de définir la pénalité comme une borne supérieure de la pénalité idéale de la forme $L_n d_m / n$ où d_m désigne la dimension de l'espace vectoriel S_m et L_n dépend de la complexité de la collection $\mathcal{M}_n$. Ils montrent qu'alors l'estimateur $\tilde{s}$ satisfait une inégalité oracle, c'est-à-dire qu'il existe une quantité C_n telle que

$$\mathbb{E}\left(\|s - \tilde{s}\|^2\right) \leq C_n \inf_{m \in \mathcal{M}_n} \mathbb{E}\left(\|s - \hat{s}_m\|^2\right). \tag{1.7}$$

Ils montrent l'optimalité du point de vue du minimax de l'ordre de grandeur de C_n. Schématiquement, leur résultat peut se résumer ainsi:
-Pour des collections de modèles suffisamment régulières et pauvres, on peut choisir $L_n \leq C$ et obtenir une inégalité oracle avec une constante C devant l'infimum.
-Pour des collections de modèles plus riches ou plus irrégulières, on doit prendre $L_n \geq C \ln n$ pour obtenir une inégalité oracle et on a une perte logarithmique inévitable

$C_n \geq C \ln n$.
En particulier, ils réinterprètent des résultats de Donoho *et al.* [27] dans la théorie de la sélection de modèles et expliquent la perte logarithmique du risque quadratique de l'estimateur par seuillage d'ondelettes. Cette approche se généralise au cas de l'estimation de densité pour des données mélangeantes. C'est l'objet de l'article de Comte & Merlevède [23] pour des données β-mélangeantes et du Chapitre 3 pour des données τ-mélangeantes.

Sélection de modèles optimale

Comme le remarque Arlot [4], les inégalités oracles précédentes souffrent du fait que l'on compare le risque moyen de l'estimateur des moindres carrés pénalisés $\tilde{s}$ au meilleur estimateur $\hat{s}_m$ choisi de manière déterministe. Comme notre choix de pénalité (et donc de $\hat{m}$) peut être aléatoire, il est préférable de montrer des inégalités oracles de la forme

$$\mathbb{E}\left(\|\tilde{s} - s\|^2\right) \leq C\mathbb{E}\left(\inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2\right), \tag{1.8}$$

ou des probabilités de déviation

$$\mathbb{P}\left(\|s - \tilde{s}\|^2 > C \inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2\right) \leq \alpha_n. \tag{1.9}$$

On dit que $\tilde{s}$ satisfait une inégalité oracle en moyenne dans les cas (1.7) et (1.8) et une inégalité oracle trajectorielle dans le cas (1.9). Plus généralement, on dit que $\tilde{s}$ satisfait une inégalité oracle quand il vérifie une inégalité de la forme (1.7), (1.8) ou (1.9). Nous nous intéressons dans cette thèse au problème de la sélection de modèles optimale. Suivant Birgé & Massart [17], Arlot [5] ou Arlot & Massart [7], nous définissons

Définition: *Soit $(S_m)_{m \in \mathcal{M}_n}$ une collection de modèles et, pour tout m de $\mathcal{M}_n$ soit $\hat{s}_m$ un estimateur de s défini sur S_m. On dit que la pénalité $pen : \mathcal{M}_n \to \mathbb{R}^+$ est une procédure de sélection optimale si l'estimateur*

$$\tilde{s} = \hat{s}_{\hat{m}}, \text{ où } \hat{m} \in \arg\min_{m \in \mathcal{M}_n} \{P_n Q(\hat{s}_m) + pen(m)\} \tag{1.10}$$

satisfait l'une des deux inégalités oracles suivantes:
Il existe une suite $(\epsilon_n)_{n \in \mathbb{N}^*} \to 0$ telle que

$$\mathbb{E}\left(\|s - \tilde{s}\|^2\right) \leq (1 + \epsilon_n)\mathbb{E}\left(\inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2\right).$$

Il existe deux constantes $K > 0$ et $\gamma > 1$ et une suite $(\epsilon_n)_{n \in \mathbb{N}^*} \to 0$ telles que

$$\mathbb{P}\left(\|s - \tilde{s}\|^2 > (1 + \epsilon_n)\left(\inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2\right)\right) \leq \frac{K}{n^\gamma}.$$

Les résultats de Barron, Birgé & Massart [10] montrent que ces inéqualités oracles ne sont pas accessibles pour des collections trop complexes. Shibata [63] avait également montré qu'il était nécessaire que la collection $\mathcal{M}_n$ ne comporte pas de modèles de dimension finie et sans biais. Au chapitre 2, nous relierons la vitesse de convergence de ϵ_n vers 0 à l'ordre de grandeur de $R_{m_o} = \inf_{m\in\mathcal{M}_n}\left(n\|s-s_m\|^2+d_m\right)$ et à la complexité de $\mathcal{M}_n$. Introduisons maintenant deux méthodes récentes de calibration de la pénalité permettant d'obtenir des procédures optimales.

L'heuristique de pente

L'heuristique de pente est une méthode de calibration automatique de la pénalité introduite par Birgé & Massart [17] dans le cadre de la régression gaussienne et dans un cadre plus général de M-estimation par Arlot & Massart [7]. Elle repose sur une compréhension fine du comportement de la pénalité idéale. On montre la propriété suivante:

Propriété: *Soit $(S_m)_{m\in\mathcal{M}_n}$ une collection de modèles et, pour tout m de $\mathcal{M}_n$ soit $\hat{s}_m$ un estimateur de s défini sur S_m. Il existe une constante $K_{\min}>0$ et, pour tout m de $\mathcal{M}_n$, une quantité déterministe C_m telles que, pour toute procédure de sélection donnée par (1.10), on a*
-Si pen$(m)<K_{\min}C_m$, alors $C_{\hat{m}}$ est très grande.
-Si pen$(m)\simeq KC_m$ avec $K>K_{\min}$ alors $C_{\hat{m}}$ est beaucoup plus petite.
-Si, pour tout m de $\mathcal{M}_n$, pen$(m)\simeq 2K_{\min}C_m$, alors la procédure de sélection est optimale.

Quand la quantité C_m est connue en pratique, on peut l'utiliser pour calibrer la pénalité de façon optimale, il suffit de calculer, pour différentes valeurs de K, la valeur de $C_{\hat{m}}$ quand la pénalité vaut KC_m. On trace ensuite $C_{\hat{m}}$ en fonction de K. On repère $K_{\min}$ par une pente[1] fortement décroissante dans cette courbe. On choisit finalement la pénalité $\mathrm{pen}(m)=2K_{\min}C_m$ qui est optimale par la propriété précédente. Dans le modèle de densité, nous allons montrer que l'on peut prendre

$$C_m=\frac{1}{2}\mathbb{E}\left(\mathrm{pen}_{id}(m)\right)=\mathbb{E}\left(\|s_m-\hat{s}_m\|^2\right),\ K_{\min}=1.$$

C_m est donc inconnue en général. Toutefois, quand les modèles de $\mathcal{M}_n$ sont très réguliers, on retrouve que cette complexité est de l'ordre de grandeur de d_m/n. Prenons l'exemple simple où les données $X_1,...,X_n$ sont indépendantes, où s est à support dans $[0,1)$ et où $(S_m)_{m\in\mathcal{M}_n}$ est un ensemble d'histogrammes réguliers c'est-à-dire que pour tout entier m de $\mathbb{N}^*$, S_m est l'ensemble des fonctions constantes sur tous les ensembles $[k/m,(k+1)/m)$, $k=0,...,m-1$. On vérifie sans peine que pour tout m de $\mathcal{M}_n=\{1,...,n\}$, $\mathbb{E}\left(\sup_{t\in B_m}(\nu_n t(X))^2\right)=(d_m-\|s_m\|^2)/n$. Ainsi, la complexité de Barron, Birgé & Massart $C_m=d_m/n$ convient, l'heuristique de pente permettant alors de calibrer $\hat{K}_{\min}$ pour définir une procédure efficace. Pour traiter les modèles moins réguliers, on peut utiliser des estimateurs par rééchantillonnage de C_m. On retrouve alors les pénalités par rééchantillonnage définies par Arlot [5].

[1] Contrairement à ce que suggère cette présentation, le nom heuristique de pente ne vient pas du changement de "pente" de cette courbe. La "pente" originelle est celle de la courbe $C_m\to P_n(Q(\hat{s}_m))$ qui admet asymptotiquement pour asymptote la droite $y=-K_{\min}x-\|s\|^2$. Notre approche est celle de l'article de Arlot & Massart [7].

Le rééchantillonnage

En 1979, Efron [30] propose un algorithme général d'estimation de fonctionnelles du processus emprique. Son idée est que, pour estimer ces fonctionnelles, il suffit de tirer "au hasard", de nouvelles données parmi les observations $X_1, ..., X_n$. L'algorithme d'Efron fonctionne en deux étapes:
-On tire n "nouvelles" données $X_1^*, ..., X_n^*$, indépendantes et de même loi P_n conditionnellement au vecteur $(X_1, ..., X_n)$.
-On estime une fonctionnelle $F(P_n, P)$ par $\mathbb{E}^* \left(F(P_n^*, P_n) \right)$, où $P_n^* t = \sum_{i=1}^n t(X_i^*)/n$ et où E^* désigne l'espérance conditionnelle à $X_1, ..., X_n$.
Mason & Newton [54] et Praestgrad & Wellner [57] ont remarqué que l'on pouvait écrire $P_n^*(t) = \sum_{i=1}^n W_i t(X_i)/n$ où $(W_1, ..., W_n)$ est un vecteur indépendant de $X_1, .., X_n$ et de loi multinomiale $\mathcal{M}(n, 1/n, ..., 1/n)$. Ils ont alors proposé de généraliser l'algorithme d'Efron de la manière suivante:
-On se donne un vecteur $(W_1, ..., W_n)$ de variables aléatoires indépendantes de $X_1, ..., X_n$ et échangeables, ce qui signifie que, pour toute permutation τ de $\{1, .., n\}$,

$$(W_{\tau(1)}, ..., W_{\tau(n)}) \text{ a même loi que } (W_1, ..., W_n).$$

-L'estimateur par rééchantillonnage d'une fonctionnelle $F(P_n, P)$ est donné par

$$C_W \mathbb{E}^W \left(F(P_n^W, \bar{W}_n P_n) \right),$$

où $P_n^W t = \sum_{i=1}^n W_i t(X_i)/n$, $\bar{W}_n = \sum_{i=1}^n W_i/n$, E^W désigne l'espérance conditionnelle à $X_1, ..., X_n$ et où C_W est une constante ne dépendant que de la loi des poids $(W_1, ..., W_n)$ et de la fonctionnelle F.
Arlot [5] a utilisé cette heuristique pour définir un algorithme de calibration automatique de la pénalité dans les problèmes de M-estimation. Suivant l'heuristique d'Efron, il définit $\nu_n^W = P_n^W - \bar{W}_n P_n$, $\hat{s}_m^W = \arg\min_{t \in S_m} P_n^W Q(t)$ et

$$\begin{aligned} \text{pen}(m) &= -C_W \mathbb{E}^W \left(\nu_n^W (Q(\hat{s}_m^W)) \right) = C_W \mathbb{E}^W \left(\|\hat{s}_m^W - \bar{W}_n \hat{s}_m\|^2 \right) \\ &= C_W \mathbb{E}^W \left(\sum_{\lambda \in m} \left((P_n^W - \bar{W}_n P_n)(\psi_\lambda) \right)^2 \right). \end{aligned}$$

où C_W est une constante de renormalisation. Arlot a prouvé l'efficacité de cet algorithme pour sélectionner le meilleur histogramme en régression. Il a également obtenu de nombreux résultats numériques montrant les performances de cet algorithme pour différentes lois de rééchantillonnage. Il a aussi comparé cette approche avec d'autres méthodes de calibration automatiques de pénalités (notamment la validation croisée et l'heuristique de pente). De manière générale, dans [4], on trouvera de nombreuses références historiques ainsi que des résultats théoriques et pratiques sur l'utilisation de l'heuristique du rééchantillonnage en sélection de modèles.
Rappelons que l'idée d'utiliser ces pénalités en sélection de modèles vient d'Efron [31]. Elle a été ensuite utilisée par Fromont [32] pour définir des pénalités globales en classification. Dans le modèle de densité, nous citons ici Celisse [21] qui a étudié des méthodes de calibration par validation croisée.
Les pénalités par rééchantillonnage peuvent être optimisées grâce à l'heuristique de pente car elles fournissent de bons estimateurs de la complexité C_m du modèle S_m. Nous serons toujours capables dans nos exemples de calculer théoriquement la constante C_W optimisant asymptotiquement les performances de $\tilde{s}$. Il ne sera donc pas

nécessaire d'avoir recourt à cette heuristique. Toutefois, la méthode de la pente peut permettre dans certaines situations de surpénaliser légèrement, ce qui peut améliorer les performances des estimateurs d'un point de vue non asymptotique (voir Arlot & Massart [7]).

1.3 Données mélangeantes

Dans de nombreux modèles statistiques, les données ne peuvent raisonnablement être modélisées par des suites indépendantes. Pour affaiblir cette hypothèse et couvrir une gamme plus large de modèles, l'hypothèse de mélange a été introduite. Elle est particulièrement bien adaptée à la généralisation de méthodes définies dans le cadre indépendant. Pour définir une hypothèse de mélange de manière générale, on se donne une propriété vérifiée par des variables aléatoires indépendantes, par exemple:
Si X et Y sont deux variables aléatoires réelles, si P_Y désigne la loi de Y et si $P_{Y|\sigma(X)}$ désigne une loi de Y conditionnellement à $\sigma(X)$ alors, pour tout évènement A de $\sigma(Y)$,

$$P_{Y|\sigma(X)}(A) - P_Y(A) = 0 \text{ p.s., donc } \mathbb{E}\left(\sup_{A\in\sigma(Y)} P_{Y|\sigma(X)}(A) - P_Y(A)\right) = 0.$$

Pour deux variables aléatoires quelconques, on définit alors le coefficient de mélange associé à cette propriété

$$\beta(\sigma(X),\sigma(Y)) = \mathbb{E}\left(\sup_{A\in\sigma(Y)} P_{Y|\sigma(X)}(A) - P_Y(A)\right).$$

On s'intéresse à des suites de variables aléatoires strictement stationnaires, c'est à dire des processus $(X_n)_{n\in\mathbb{Z}}$ telles que, pour tous les entiers n de $\mathbb{Z}$ et k de $\mathbb{N}$, $(X_n,...,X_{n+k})$ a même loi que $(X_0,...,X_k)$.
À une telle suite $(X_n)_{n\in\mathbb{Z}}$, on associe les nombres

$$\beta_k = \sup_{l\geq k} \beta(\sigma((X_i)_{i\leq 0}), \sigma((X_i)_{i\geq l})).$$

On dit alors que la suite $(X_n)_{n\in\mathbb{Z}}$ est β-mélangeante si la suite β_k tend vers 0 quand k tend vers l'infini.
De nombreux coefficients ont ainsi été définis, citons les livres de Rio [60] et Dedecker *et.al* [24] pour une présentation des principaux coefficients et de leurs propriétés.

Les processus β et τ-mélangeants

Dans cette thèse, nous nous intéresserons particulièrement à deux coefficients. Le coefficient β dont nous venons de parler a été introduit par Rozanov & Vokonskii [71]. Berbee [13] a montré qu'il satisfait un lemme de couplage que l'on peut énoncer ainsi:

Lemme
Soit X une variable aléatoire définie sur un espace de probabilité $(\Omega,\mathcal{A},\mathbb{P})$, à valeurs

dans $\mathbb{R}^l$ et soit $\mathcal{M}$ une tribu. Supposons qu'il existe une variable aléatoire U uniformément distribuée sur $[0,1]$ et indépendante de $\mathcal{M}$ et de $\sigma(X)$. Alors, il existe une variable aléatoire $\mathcal{M} \vee \sigma(X) \vee \sigma(U)$-mesurable X^*, de même loi que X et indépendante de $\mathcal{M}$ telle que

$$\beta(\mathcal{M}, \sigma(X)) = \mathbb{P}(X \neq X^*)$$

Viennet [70] a déduit de ce lemme qu'une suite de variables aléatoires β-mélangeantes peut être approchée par une suite indépendante. Baraud *et.al.* [9] ont utilisé ce résultat pour construire un algorithme de sélection de modèles pour des données mélangeantes. Comte & Merlevède [23] ont ensuite appliqué cet algorithme dans le cadre de l'estimation de la densité et ont ainsi étendu des résultats de Barron, Birgé & Massart [10] au cadre β-mélangeant.
Le second coefficient auquel nous allons nous intéresser est le coefficient τ introduit par Dedecker & Prieur [25]. C'est le coefficient de mélange associé à la propriété suivante:
Si X est une variable aléatoire indépendante d'une tribu $\mathcal{M}$, alors pour toute fonction t 1-lipschitzienne

$$|\mathbb{P}_{Y|\mathcal{M}}(t) - \mathbb{P}_Y(t)| = 0 \text{ p.s.}.$$

Si λ_1 désigne l'ensemble des fonctions 1-lipschitziennes, on définit donc

$$\text{si } \mathbb{E}(|Y|) < \infty,\ \tau(\mathcal{M}, Y) = \mathbb{E}\left(\sup_{t \in \lambda_1} |\mathbb{P}_{Y|\mathcal{M}}(t) - \mathbb{P}_Y(t)|\right).$$

Si $(X_n)_{n\in\mathbb{Z}}$ est une suite de variables aléatoires strictement stationnaires, on définit, pour tout k et r de $\mathbb{N}^*$, les réels

$$\tau_{k,r} = \max_{1 \le l \le r} \frac{1}{l} \sup_{k \le i_1 < .. < i_l} \left\{\tau(\sigma(X_p, p \le 0), (X_{i_1}, ..., X_{i_l}))\right\},\ \tau_k = \sup_{r \in \mathbb{N}^*} \tau_{k,r}.$$

Dedecker & Prieur [25] ont montré que ce coefficient satisfaisait le lemme de couplage suivant.

Lemme
Soit X une variable aléatoire définie sur un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans $\mathbb{R}^l$ et soit $\mathcal{M}$ une tribu. Supposons qu'il existe une variable aléatoire U uniformément distribuée sur $[0,1]$ et indépendante de $\mathcal{M}$ et de $\sigma(X)$. Alors, il existe une variable aléatoire $\mathcal{M} \vee \sigma(X) \vee \sigma(U)$-mesurable X^*, de même loi que X et indépendante de $\mathcal{M}$ telle que

$$\tau(\mathcal{M}, X) = \mathbb{E}\left(|X - X^*|_l\right),$$

où, pour tout x et y de $\mathbb{R}^l$, $|x-y|_l = \sum_{i=1}^l |x_i - y_i|$.

Dans le chapitre 3, nous verrons comment utiliser ce lemme pour approcher les suites τ-mélangeantes par des suites indépendantes, étendre ainsi l'algorithme de Baraud *et.al.* [9] et en déduire une généralisation des résultats de Barron, Birgé & Massart [10] au cadre τ-mélangeant.

Exemples de processus mélangeant

Pour les exemples de processus β-mélangeants, nous renvoyons aux livres de Doukhan [28], Rio [60] ou Bradley [19]. Précisons seulement qu'une chaîne de Markov stationnaire, irréductible, apériodique et positivement récurrente est β-mélangeante. En revanche, de nombreuses chaînes de Markov ne sont pas β-mélangeantes mais sont τ-mélangeantes. Par exemple, si $(\epsilon_i)_{i\geq 1}$ sont des variables indépendantes de Bernoulli de paramètre $1/2$, alors la solution stationnaire de l'équation

$$X_n = \frac{1}{2}(X_{n-1} + \epsilon_n),\ X_0 \text{ indépendante de } (\epsilon_i)_{i\geq 1}$$

n'est pas β-mélangeante puisque $\beta_k = 1$ pour tout $k \geq 1$ (Andrews [3]) alors que $\tau_k \leq 2^{-k}$ (Dedecker & Prieur [25]). Un autre avantage du coefficient τ est qu'il est calculable dans de nombreuses situations. Rappelons ici quelques exemples tirés de Dedecker & Prieur [25].

Processus Linéaires.
Supposons que $X_i = \sum_{j\geq 0} a_j \xi_{n-j}$, où les $(\xi_i)_{i\in\mathbb{Z}}$ sont i.i.d. On a

$$\tau_k \leq 2\mathbb{E}|\xi_0| \sum_{j\geq k} |a_j|.$$

Chaines de Markov
Soit $(X_n)_{n\geq 0}$ une chaîne de Markov telle que $X_n = F(X_{n-1}, \xi_n)$ où F est une fonction mesurable et où $(\xi_i)_{i\geq 1}$ est une suite de variables aléatoires iid, indépendante de X_0. Supposons qu'il existe $\kappa < 1$ tel que

$$\mathbb{E}(|F(x,\xi_0) - F(y,\xi_0)|) \leq \kappa|x-y|.$$

Alors, on a

$$\tau_k \leq 2\mathbb{E}(|X_0|)\kappa^k.$$

Un exemple important est donné par les processus auto-régressifs $X_n = t(X_{n-1}) + \xi_n$ où t est une fonction κ-Lipschitzienne.

Systèmes dynamiques
Soit T une application mesurable de $[0,1]$ dans $[0,1]$. Si ν est une mesure de probabilité T-invariante, et si Y est une variable aléatoire de loi ν, la suite de variables aléatoires $(Y_i = T^i(Y))_{i\geq 0}$ définie sur $([0,1],\nu)$ est strictement stationnaire. Définissons l'opérateur K sur $L^1([0,1],\nu)$ à valeurs dans $L^1([0,1],\nu)$ *via* l'égalité

$$\forall k \in L^\infty([0,1],\nu),\ \int_0^1 (Kh)(x)k(x)\nu(dx) = \int_0^1 h(x)(k\circ T)(x)\nu(dx)$$

où $h \in L^1([0,1],\nu)$. On vérifie que $(Y_1,...,Y_n)$ a la même loi que $(X_n,...,X_1)$, où $(X_i)_{i\geq 1}$ est la chaîne de Markov de distribution invariante ν et de noyau de transition K. Si T est uniformément dilatante (voir les hypothèses page 218 dans Dedecker & Prieur [25]), il existe $C > 0$ et $\rho \in (0,1)$ tels que

$$\tau(\sigma(X_i, i\geq k), X_0) \leq C\rho^k$$

(voir Dedecker & Prieur [25] page 230). Remarquons que la chaîne de Markov $(X_i)_{i\geq 1}$ n'est pas β-mélangeante. En effet, $\beta(\sigma(X_1),\sigma(X_n)) = \beta(\sigma(T^n(Y)),\sigma(T(Y)))$. Comme $\sigma(T^n(Y)) \subset \sigma(T(Y))$, il vient, par stationnarité de Y_i,

$$\beta(\sigma(X_1),\sigma(X_n)) \geq \beta(\sigma(T^n(Y)),\sigma(T^n(Y))) = \beta(\sigma(T(Y)),\sigma(T(Y))).$$

Cette dernière borne est strictement positive dès que ν est non triviale.

Décomposition des données par blocs

Pour démontrer théoriquement les résultats d'estimation dans le cadre mélangeant, l'idée fondamentale est de découper le vecteur des données en blocs, pour utiliser la méthode de couplage. On se donne une partition $I_0, J_0, ..., I_{p-1}, J_{p-1}$ de $\{1, ..., n\}$ telle que

$$q = \inf_{k=0,...,p-2} (\min(I_{k+1}) - \max(I_k)) > 0.$$

Pour tout k, on note X_{I_k} le vecteur $(X_i)_{i \in I_k}$, soit aussi $I = \cup_{k=0}^{p-1} I_k$ et P_I le processus empirique $P_I = \sum_{i \in I} \delta_{X_i}/|I|$. Les lemmes de couplage évoqués à la section précédente servent à construire une suite de variables indépendantes $(X^*_{I_k})_{k=0,...,p-1}$ telles que, pour une certaine distance d, on ait, pour tout k $d(X_{I_k}, X^*_{I_k}) \leq \gamma(q)$ où γ est le coefficient de mélange. Les quantités construites à partir du processus empirique $F(P_n)$ sont alors contrôlées en deux étapes.
-Des inégalités algébriques permettent d'obtenir une constante $C \geq 1$ telle que

$$F(P_n) \leq CF(P_I).$$

-Ensuite, comme

$$F(P_I) \leq F(P^*_I) + |F(P^*_I) - F(P_I)|,$$

on peut contrôler $F(P_I)$ grâce aux techniques valables pour des données indépendantes (pour le terme $F(P^*_I)$) et aux coefficients de couplage (pour le terme $|F(P^*_I) - F(P_I)|$).
À notre connaissance, toutes les méthodes de sélection de modèles proposées pour étudier des processus mélangeants utilisent la méthode de couplage. Cette méthode est particulièrement efficace lorsqu'elle est associée à l'heuristique de rééchantillonnage. En effet, nous allons montrer pour nos fonctionnelles d'intérêt que $F(P^*_I)$ se comporte essentiellement comme son espérance. Comme l'ont remarqué Künsch [44], Liu & Singh [51], Radulovic [59], cette espérance est très bien approchée par son estimateur par rééchantillonnage, si on rééchantillonne les blocs (voir le chapitre 4 pour plus de détails).

Sélection de modèles en milieu mélangeant

Baraud *et.al.* [9] ont montré que les pénalités de Barron, Birgé & Massart [10] pouvaient être utilisées dans le modèle de régression quand les données sont β-mélangeantes. En utilisant l'indépendance du design et du bruit, ils obtiennent une pénalité calculable en pratique. Avec les mêmes outils, Comte & Merlevède [23] ont montré que les pénalités de la forme Kd_m/n sont aussi efficaces en estimation de la densité. Nous étendons cette approche au chapitre 3 au cas de données τ-mélangeantes. Nous en déduisons en particulier l'adaptativité de $\tilde{s}$ grâce au schéma de preuve de Barron, Birgé & Massart [10]. Ce dernier résultat est à rapprocher de ceux obtenus par Gannaz & Wintenberger [34]. Ils avaient en effet montré l'adaptativité de l'estimateur par seuillage de Donoho *et al.* [27] dans le cadre de données $\tilde{\phi}$-mélangeantes (le coefficient $\tilde{\phi}$ est défini dans Dedecker & Prieur [25]).
Le problème de l'algorithme de Baraud *et.al.* [9] dans le modèle de densité est que la constante K dans la pénalité dépend des coefficients de mélange. Cette pénalité est donc bien souvent incalculable en pratique. Comte & Merlevède [23]

ont proposé d'appliquer l'heuristique de pente dans le cadre mélangeant sans toutefois fournir de démonstration théorique de sa validité. Nous en proposons une dans le chapitre 4 dans les cadres β et τ-mélangeants. Nous étendons aussi dans ce chapitre l'approche par rééchantillonnage et obtenons ainsi des pénalités complètement calculables à partir des observations. À notre connaissance, nous avons obtenu la première preuve de leur validité théorique en estimation de la densité en milieu mélangeant. Nous obtenons des inégalités oracles trajectorielles pour des données β-mélangeantes et des inégalités oracles en moyenne pour les données τ-mélangeantes. Toutes les procédures sont optimales, ce qui est encore un résultat nouveau.
La grande différence avec le cas indépendant est qu'un terme de couplage apparait
-dans le contrôle de la probabilité de déviation pour les processus β-mélangeants
-dans le contrôle de l'espérance du risque pour les processus τ-mélangeants.
Ces termes de couplages sont contrôlés par les coefficients de mélange mais peuvent dégrader les performances des estimateurs. Nous ne savons pas si nos contrôles sont optimaux, mais les conditions de mélanges que nous demandons sont plus faibles que celles décrites par Comte & Merlevède [23] dans le cas β-mélangeant, ce sont les premières dans le cas τ-mélangeant.
Remarquons qu'il y a un paradoxe à utiliser l'heuristique du rééchantillonnage en milieu mélangeant. En effet, les poids sont échangeables, ce qui signifie que tout se passe de la même façon si on échange les données alors que par définition, un processus mélangeant est un processus qui oublie avec le temps, au sens où deux données éloignées sont presque indépendantes alors qu'on ne sait rien de la dépendance des données rapprochées. Cette difficulté disparait lorsque l'on rééchantillonne les blocs au lieu des données.

1.4 Présentation des résultats

Sélection de modèles optimale en estimation de la densité

Le chapitre 2 de la thèse est consacré à l'extension des résultats de Arlot [5] pour les pénalités par rééchantillonnage et à ceux de Arlot & Massart [7] sur l'heuristique de pente au cadre de l'estimation de la densité. L'hypothèse fondamentale que nous faisons (hypothèse **[V]** dans la section 2 du chapitre 2) est que les déviations de la pénalité idéale $\mathrm{pen}_{id}(m) = 2\nu_n(\hat{s}_m)$ sont uniformément petites devant l'espérance du risque $R_m/n = \mathbb{E}\left(\|s - \hat{s}_m\|^2\right)$. Cette hypothèse est légèrement différente de celles traditionnellement considérées en sélection de modèles. En effet, **[V]** ne suppose pas de forme particulière pour l'espérance du risque, en particulier pour le terme de variance, alors que celui-ci est souvent supposé borné par d_m/n. Cette borne est bien justifiée quand les modèles sont relativement réguliers (voir la section 4 du chapitre 2) mais elle peut être très mauvaise en général. Birgé [14] a exhibé des histogrammes de petites dimensions pour lesquels $R_m >> d_m$. Sous cette hypothèse, nous justifions d'abord l'utilisation de l'heuristique de pente dans l'estimation de la densité. La complexité C_m à considérer dans cette heuristique est l'espérance du terme de variance $D_m/n = \mathbb{E}\left(\|s_m - \hat{s}_m\|^2\right)$ et la constante $K_{\min} = 1$. En effet, on a (voir la Proposition 2.2.2 du chapitre 2)

Proposition 1.4.1 *Soit $\mathcal{M}_n$ une collection de modèles satisfaisant l'hypothèse* **[V]**. *Supposons qu'il existe une constante $0 < \delta < 1$ telle que $0 \leq pen(m) \leq (1-\delta)D_m/n$.*

Soit $\hat{m},\tilde{s}$ les variables aléatoires construites par la procédure de sélection de modèles associée. Alors, sur un évènement de probabilité supérieure à $1-c_1e^{-c_2(\ln n)^2}$, on a

$$D_{\hat{m}}\geq c_3\max_{m\in\mathcal{M}_n}D_m,\;\|s-\tilde{s}\|^2\geq c_4n^{c_5}\inf_{m\in\mathcal{M}_n}\|s-\hat{s}_m\|^2,$$

où c_1,c_2,c_3,c_4,c_5 sont des constantes positives, dépendantes de δ.

Cette proposition justifie le premier point de l'heuristique de pente. Si la pénalité est trop petite, $D_{\hat{m}}$ est aussi grand que possible (en général de l'ordre de n), de plus, on ne peut pas montrer d'inégalités oracles. L'heuristique de pente est complètement justifiée par la seconde proposition (voir Proposition 2.2.3 du chapitre 2)

Proposition 1.4.2 *Soit $\mathcal{M}_n$ une collection de modèles satisfaisant l'hypothèse* **[V]**. *Supposons qu'il existe $0<\delta<1$ telle que*

$$2\frac{D_m}{n}-\delta\frac{R_m}{n}\leq pen(m)\leq 2\frac{D_m}{n}+\delta\frac{R_m}{n}.$$

Alors, sur un évènement de probabilité plus grande que $1-c_1e^{-c_2(\ln n)^2}$, on a

$$D_{\hat{m}}\leq c_3\min_{m\in\mathcal{M}_n}R_m,\;\|s-\tilde{s}\|^2\leq\left(\frac{1+\delta}{1-\delta}+c_4n^{-c_5}\right)\inf_{m\in\mathcal{M}_n}\|s-\hat{s}_m\|^2,$$

où c_1,c_2,c_3,c_4,c_5 sont des constantes positives dépendantes de δ.

Le risque d'un oracle $\min_{m\in\mathcal{M}_n}R_m$ est en général de l'ordre de n^r avec $r<1$. On observe donc bien un saut de la complexité du modèle sélectionné quand la pénalité devient plus grande que D_m/n, ceci justifie le second point de l'heuristique de pente. Pour le dernier point, il suffit de faire tendre δ vers 0 en vérifiant que les constantes c_1,c_2,c_3,c_4,c_5 n'explosent pas. Le problème de cette heuristique en estimation de la densité est que la complexité D_m est inconnue en pratique, elle doit donc être correctement estimée pour calibrer les pénalités. Nous montrons en section 4 du chapitre 2 que, lorsque le modèle est suffisamment régulier, la dimension d_m peut être utilisée à la place de D_m. Nous montrons aussi que, sans autre hypothèse que **[V]**, on peut utiliser un estimateur par rééchantillonnage de D_m. Ce dernier choix revient à utiliser les pénalités par rééchantillonnage d'Arlot [5]. On montre en effet la proposition suivante (voir proposition 2.2.5 du chapitre 2)

Proposition 1.4.3 *Soient $X_1,...,X_n$ des variables aléatoires i.i.d de densité commune s. Soit $(W_1,...,W_n)$ un schéma de rééchantillonnage, soit $\bar{W}_n=\sum_{i=1}^nW_i/n$ et soit $v_W^2=Var(W_1-\bar{W}_n)$. Soit $\mathcal{M}_n$ une collection de modèles satisfaisant l'hypothèse* **[V]** *et soit $\tilde{s}$ l'estimateur sélectionné par la pénalité*

$$pen(m)=2(v_W^2)^{-1}\mathbb{E}^W\left(\sup_{t\in B_m}(\nu_n^W(t))^2\right).$$

Il existe des constantes c_1,c_2,c_3,c_4 strictement positives telles que

$$\mathbb{P}\left(\|s-\tilde{s}\|^2\leq\left(1+c_1n^{-c_2}\right)\inf_{m\in\mathcal{M}_n}\|s-\hat{s}_m\|^2\right)\geq 1-c_3e^{-c_4(\ln n)^2}.$$

Tous les résultats de ce chapitre sont non asymptotiques avec constantes explicites. Trois quantités nous permettent de décrire le saut de $D_{\hat{m}}$ dans l'heuristique de pente ainsi que la vitesse de convergence de C_n vers 1 dans les inégalités oracles. La première est le risque d'un oracle, $R_{m_o} = \min_{m\in\mathcal{M}_n} R_m$, la seconde est la plus grande des variances $D_{m*} = \max_{m\in\mathcal{M}_n} D_m$, la dernière est $\epsilon(R_{m_o})$ ou ϵ est la fonction introduite dans l'hypothèse **[V]**. Les deux premières quantités sont reliées par le paradigme de la sélection de modèles. Ce paradigme peut se résumer aux hypothèses suivantes

$$\frac{D_{m_o}}{n} \simeq \|s - s_{m_o}\|^2,\ \frac{D_{m^*}}{n} >> \|s - s_{m^*}\|^2,\ D_{m_o} << D_{m^*},\ R_{m_o} \to +\infty.$$

La dernière condition signifie que la vitesse d'estimation de s est non paramétrique. Essentiellement, ceci signifie qu'il n'existe pas de modèle de dimension finie expliquant s. Nos procédures ne sont optimales que sous cette hypothèse (en particulier, la condition $c_2 > 0$ dans la proposition 1.4.3 n'est vérifiée que s'il existe $r > 0$ tel que $R_{m_o} > Cn^r$). Toutefois, quand R_{m_o} est borné, nous obtenons des inégalités oracles avec une constante $C > 1$. Ainsi, même quand le problème de départ est très simple, nos procédures restent compétitives.

Sélection de modèles avec données mélangeantes

Le chapitre 3 est consacré au problème de sélection de modèles classique mais les données sont supposées mélangeantes. Ce chapitre s'inspire en grande partie de l'article de Comte & Merlevède [23]. Cet article traite le cas de données β-mélangeantes. La méthode de couplage que nous avons évoquée en section 3 est utilisée pour montrer qu'un estimateur sélectionné avec une pénalité de la forme Cd_m/n satisfait une inégalité oracle du type (1.3). Nous traitons d'abord des données β-mélangeantes. Nous prouvons une inégalité oracle avec grande probabilité plutôt qu'en espérance. Ce point de vue est mieux adapté au lemme de couplage de Berbee. Il nous permet d'affaiblir l'hypothèse de Comte & Merlevède [23] sur la vitesse de mélange. De plus, nous améliorons les constantes et nous ne supposons plus que s est bornée. Nous obtenons le résultat suivant (voir la proposition 3.3.1 du chapitre 3).

Théorème:
Supposons que le processus $(X_n)_{n\in\mathbb{Z}}$ est strictement stationnaire et β-mélangeant. Supposons de plus qu'il existe $\theta > 2$ tel que, pour tout entier l, $\beta_l \leq (1 + l)^{-(1+\theta)}$. Supposons que la collection de modèles $\mathcal{M}_n$ satisfait certaines hypothèses techniques et soit $\tilde{s}$ l'estimateur sélectionné par la pénalité

$$pen(m) = \frac{Kd_m}{n}.$$

Si K est assez grand, il existe des constantes positives c_1, c_2, c_3 telles que

$$\mathbb{P}\left(\|\tilde{s} - s\|^2 > (1 + c_1(\ln n)^{-c_2}) \inf_{m\in\mathcal{M}_n} \left(\|s - s_m\|^2 + \mathrm{pen}(m)\right)\right) \leq c_3 \frac{(\log n)^{(\theta+2)\kappa}}{n^{\theta/2}}.$$

Les hypothèses techniques de ce théorème sur la collection $\mathcal{M}_n$ sont les mêmes que dans les articles de Comte & Merlevède [23] et Birgé & Massart [15]. La constante

K dans la pénalité peut être choisie beaucoup plus petite que dans le théorème 3.1 de Comte & Merlevède [23], qui supposaient en outre l'hypothèse $\theta > 3$.
Le principal problème de ce résultat est que la constante K définissant la pénalité dépend des coefficients de mélange et est donc inconnue en pratique. La constante K est choisie de façon à assurer que $\text{pen}(m) \geq \mathbb{E}\left(\|s_m - \hat{s}_m\|^2\right)$. Ainsi, l'inégalité oracle obtenue est un peu plus faible que l'inégalité (1.3). Néanmoins, elle a la même forme que l'inégalité oracle obtenue pour des données indépendantes et permet d'obtenir l'adaptativité de l'estimateur $\tilde{s}$ à la régularité de s.
Nous étendons ensuite cette approche aux processus τ-mélangeants qui satisfont aussi un lemme de couplage et une inégalité de covariance. Toutefois, ces résultats sont plus faibles pour les processus τ-mélangeants et l'extension des résultats est techniquement difficile. En particulier, nous n'obtenons d'inégalité oracle que pour des modèles d'ondelettes régulières et bien localisées. L'avantage de ce résultat est que le coefficient τ est souvent plus facile à calculer que le coefficient β (voir les exemples de la section 3). Nous avons montré le résultat suivant (Théorème 4.1 dans le chapitre 3)

Théorème:
Soit $\mathcal{M}_n$ une collection de modèles engendrés par des ondelettes régulières à support compact. Supposons que le processus $(X_n)_{n\in\mathbb{Z}}$ est strictement stationnaire et τ-mélangeant. Supposons de plus qu'il existe $\theta > 5$ tel que, pour tout entier l, $\tau_l \leq (1+l)^{-(1+\theta)}$.
Soit $\tilde{s}$ l'estimateur sélectionné par la pénalité

$$pen(m) = \frac{Kd_m}{n}.$$

Si K est assez grand, il existe des constantes c_1, c_2 telles que

$$\mathbb{E}\left(\|\tilde{s} - s\|^2\right) \leq \left(1 + c_1(\ln n)^{-c_2}\right)\left(\inf_{m\in\mathcal{M}_n} \|s - s_m\|^2 + \text{pen}(m)\right).$$

Ainsi, l'estimateur vérifie le même type d'inégalités oracles que lorsque les données sont indépendantes. On peut en déduire qu'il est minimax sur des classes d'espaces de Besov en utilisant le procédé de Birgé & Massart [15]. Comme pour les données β-mélangeantes, la constante K dépend des coefficients de mélange, elle est calibrée de façon à assurer que $\text{pen}(m) \geq \mathbb{E}\left(\|s_m - \hat{s}_m\|^2\right)$ et que cette majoration soit la moins pessimiste possible.

Sélection de modèles optimale en milieu mélangeant

Le principal défaut des résultats précédents est que la pénalité dépend des coefficients de mélange qui sont inconnus en pratique. Pour calibrer de façon optimale la constante dans le terme de pénalité, Comte & Merlevède [23] suggéraient, suivant Birgé & Massart [16] d'utiliser l'heuristique de pente. Dans le chapitre 4, nous prouvons théoriquement le bien fondé de ce choix. Le lemme de couplage de Viennet [70] nous permet d'étendre assez facilement les méthodes de preuves du chapitres 2 à des données β-mélangeantes. Nous obtenons l'équivalent des Propositions 1.4.1, 1.4.2. Ces résultats améliorent ceux de la section précédente. En effet, les pénalités sont maintenant explicitement calculables. De plus, les inégalités oracles comparent le risque de $\tilde{s}$ à $\inf_{m\in\mathcal{M}_n}\|s-\hat{s}_m\|^2$ et plus à une borne supérieure de cet oracle. Comme dans le cadre indépendant, l'heuristique de pente ne peut être utilisée avec une complexité $C_m = d_m$ que si les modèles sont très réguliers, par exemple si la fonction $(x, m) \to \sup_{t\in B_m} t(x)/\sqrt{d_m}$ est presque constante. Cette dernière condition est vérifiée dans les exemples mentionnés dans Comte & Merlevède [23]. Pour pouvoir traiter des modèles un peu moins réguliers, nous avons aussi étendu l'approche par rééchantillonnage. Nous obtenons l'équivalent de la Proposition 1.4.3 dans le cas β-mélangeant. Regardons ce résultat plus précisément.

Proposition 1.4.4 *Soit $X_1, ..., X_n$ une suite strictement stationnaire de variables aléatoires de même densité s. Soit $(S_m)_{m\in\mathcal{M}_n}$ une collection de modèles satisfaisant l'hypothèse* **[V']**. *Supposons que le processus $(X_k)_{k=1,...,n}$ est β-mélangeant et qu'il existe $C > 0$ et $\theta > 2$ tels que, pour tout l, $\beta_l \le C(1 + l)^{-(1+\theta)}$. Soit $\tilde{s}$ l'estimateur sélectionné par la pénalité de rééchantillonnage.*
Pour tout $\kappa > 2$, il existe des constantes c_1 et c_2 telles que

$$\mathbb{P}\left(\|s-\tilde{s}\|^2 > (1 + c_1 f(n))\inf_{m\in\mathcal{M}_n}\|s-\hat{s}_m\|^2\right) \le c_2\frac{(\ln n)^{(\theta+2)\kappa}}{n^{\theta/2}},$$

où $f(n) = (\ln n)^{-2(\kappa-1)}$.

On observe deux grandes différences par rapport au cas indépendant. D'abord, un terme de couplage apparait dans le contrôle de la probabilité de déviation. Dans le cas indépendant, la borne est de l'ordre de $e^{-c(\ln n)^2}$, dans le cas mélangeant, elle devient $(\log n)^{c_1} n^{-\theta/2}$. Ensuite, l'hypothèse **[V]** est renforcée en une hypothèse **[V']**. La conséquence de cette transformation est que la vitesse de convergence de $f(n)$ vers 0, qui était de l'ordre de n^{-c} dans le cas indépendant devient $(\ln n)^{-c}$. Comme pour les données indépendantes, la seule hypothèse **[V']** suffit pour pouvoir appliquer les pénalités par rééchantillonnage.
Nous étendons aussi dans ce chapitre ces deux méthodes au cadre τ-mélangeant. Ce coefficient vérifie un lemme de couplage en distance L^1. Rappelons que dans la méthode de couplage, nous devons contrôler des termes de la forme $|F(P_I)-F(P_I^*)|$. Ces termes sont plus délicats à contrôler pour le coefficient τ et, comme dans la section précédente, nous choisissons des modèles d'ondelettes régulières. De plus, la forme du lemme de couplage nous a imposé de travailler en moyenne plutôt que sur des évènements de grandes probabilités. Les preuves du cas indépendant ne s'appliquaient plus directement. Nous obtenons une inégalité oracle de la forme (1.8) au lieu de (1.9). Néanmoins, ce résultat, comme en β-mélange, améliore les résulats du chapitre 3 car les pénalités sont explicitement calculables et le risque de $\tilde{s}$ est comparé à celui de l'oracle plutôt qu'à une borne supérieure de ce risque. Regardons ce que devient l'inégalité oracle obtenue par pénalité bootstrap.

Proposition 1.4.5 *Soit $X_1, ..., X_n$ une suite strictement stationnaire de variables aléatoires de densité s et soit $(S_m)_{m\in\mathcal{M}_n}$ une collection de modèles d'ondelettes régulières. Supposons que le processus $(X_k)_{k=1,...,n}$ est τ-mélangeant et qu'il existe $C > 0$ et $\theta > 5$ tels que, pour tout l, $\tau_l \le C(1 + l)^{-(1+\theta)}$. Soit $\tilde{s}$ l'estimateur sélectionné par la pénalité de rééchantillonnage. Alors, il existe des constantes c_1 et*

c_2 *telles que*

$$\mathbb{E}\left(\|s-\tilde{s}\|^2\right) \leq (1+c_1 f(n))\,\mathbb{E}\left(\inf_{m\in\mathcal{M}_n}\|s-\hat{s}_m\|^2\right)+c_2\frac{(\ln n)^{\kappa(1+\theta)}}{n^{(\theta-3)/2}},$$

où $f(n)=(\ln n)^{-2(\kappa-1)}$.

La condition de mélange est plus forte ($\theta > 5$) et le terme de couplage dans le contrôle de l'espérance est plus grand que dans le cas β-mélangeant.
À notre connaissance les résultats de ce chapitre sont nouveaux. En effet,
-l'utilisation de l'heuristique de pente n'avait jamais été justifiée théoriquement pour des données mélangeantes, dans aucun modèle, pour aucun coefficient de mélange,
-l'algorithme de preuve dans le cas τ-mélangeant est différent du cas indépendant, il ne nécessite pas seulement la prise en compte d'un terme de couplage. Les résultats obtenus pour ces processus sont plus faibles que dans le cas indépendant, mais les hypothèses pour les obtenir sont également un peu plus faibles,
-c'est la première fois que des méthodes de rééchantillonnage sont utilisées pour calibrer les pénalités quand les données sont mélangeantes, même si leurs bonnes propriétés avaient déjà été observées (voir Künsch [44], Liu & Singh [51], Radulovic [59] par exemple).

Régions de confiance pour s

Le chapitre 5 traite du problème de l'adaptativité pour les régions de confiance dans le modèle de densité. Ce problème est fondamentalement différent pour les régions de confiance et pour l'estimation. En effet, de nombreux estimateurs adaptatifs de la densité ont été proposés, voir par exemple l'article de Barron, Birgé & Massart [10] pour une construction par sélection de modèles. En revanche, comme le montre l'article de Robins & van der Vaart [61], on ne peut pas construire de régions de confiance adaptatives avec autant de généralité. Prenons l'exemple de fonctions s α-höldériennes. On peut construire des estimateurs de s adaptatifs à la régularité α, c'est-à-dire des estimateurs construits sans connaître le vrai α et se comportant aussi bien que possible quelle que soit cette régularité. En revanche, pour construire une région de confiance pour s, on doit préciser une borne inférieure α_-, sur l'ensemble des régularités admissibles. De plus, on n'obtient de régions de confiance adaptatives qu'aux régularités comprises entre α_- et $2\alpha_-$. La différence est que le biais du modèle doit être estimé quand on construit une région de confiance. On doit donc préciser un gros espace S auquel la densité est supposée appartenir. Si la fonction s appartient à un sous-espace S_m de S, la taille d'une région de confiance pour s est bornée inférieurement par le supremum de deux quantités. La première est la vitesse d'estimation dans le modèle S_m qui correspond à la vitesse de convergence du terme de variance en estimation. La seconde est la vitesse d'estimation du biais d'une fonction de S qui correspond à la vitesse de test de l'hypothèse $s \in S_m$ contre l'hypothèse $s \in S$ et $s \notin S_m$. Cette seconde vitesse est souvent définie par le gros espace S et limite l'adaptativité (voir Ingster [40, 41, 42]). Cette propriété négative avait déjà été isolée dans les articles de Juditsky & Lambert-Lacroix [43] et Baraud [8] dans le modèle de régression.
Les résultats de Robins & van der Vaart [61] sont asymptotiques et ils construisent

leurs régions de confiance en découpant l'échantillon en deux, la première partie servant à construire des estimateurs de s, la seconde servant à Ã©valuer l'erreur dans ces estimations.
Au chapitre 5, nous améliorons leurs résultats en prouvant des bornes non asymptotiques sur la taille des régions de confiance et en construisant, en nous inspirant des résultats de Baraud [8], des régions de confiance centrées sur des estimateurs utilisant toutes les données. Nous estimons l'erreur commise par rééchantillonnage pour ne pas avoir à découper l'échantillon en deux parties. Techniquement, les outils ressemblent à ceux du chapitre 2. Comme nous travaillons avec des données à valeurs réelles, les preuves sont simplifiées par l'utilisation de l'inégalité de concentration pour les U-statistiques de Houdré & Reynaud-Bouret [39].
Ces résultats négatifs ne nous ont pas incité à poursuivre dans cette direction, en particulier, nous n'avons pas cherché à étendre ces résultats au cas mélangeant.

Outils probabilistes

Pour finir, nous recensons quelques outils probabilistes utiles dans les différents chapitres. Les plus importants sont des inégalités de concentration. Les résultats de base que nous utilisons peuvent être trouvés dans les livres de Ledoux & Talagrand [48], Ledoux [47] ou Massart [55].

Processus empirique

Nous avons vu que le terme de variance d'un estimateur $\hat{s}_m$ s'exprime comme le supremum du carré du processus empirique sur l'ellipsoïde $B_m = \{t \in S_m,\ \|t\| \leq 1\}$. Notons $Z = \sup_{t\in B_m} \nu_n(t)$. L'inégalité de Talagrand (voir Talagrand [65]) décrit la concentration de Z autour de son espérance. Bousquet [18] a utilisé la méthode d'entropie de Ledoux [46] pour calculer les constantes optimales dans cette inégalité. Il a obtenu le résultat suivant.

Théorème:
Soient $X, X_1, ..., X_n$ des variables aléatoires i.i.d. à valeurs dans un espace mesurable $(\mathbb{X}, \mathcal{X})$ et soit S une classe de fonctions bornées par b. Soient $v^2 = \sup_{t\in S} Var(t(X))$ et soit $Z = \sup_{t\in S} \nu_n t$. Alors,

$$\forall x > 0,\ \mathbb{P}\left(Z > \mathbb{E}(Z) + \sqrt{\frac{2}{n}(v^2 + 2b\mathbb{E}(Z))x} + \frac{bx}{3n} \right) \leq e^{-x}.$$

C'est le résultat à la base de toutes les inégalités de concentration que nous avons montrées. Nous l'avons d'abord utilisé pour montrer la concentration de Z^2 autour de son espérance.

Corollaire 1:
Notons $D = \mathbb{E}(Z^2)$, $v^2 = \sup_{t\in S} Var(t(X))$, $\epsilon = b^2/n$, on a, pour tout $x > 0$,

$$\mathbb{P}\left(Z^2 - \frac{D}{n} > \frac{D^{3/4}(\epsilon(19x)^2)^{1/4} + 3\sqrt{Dv^2x} + 3v^2x + \epsilon(19x)^2}{n} \right) \leq e^{-x}.$$

$$\mathbb{P}\left(Z^2 - \frac{D}{n} < -\frac{8D^{3/4}(\epsilon x^2)^{1/4} + 7.61\sqrt{v^2Dx} + \epsilon(40.25x)^2}{n}\right) \leq e^{1-x}.$$

U-statistiques

En utilisant l'inégalité de Talagrand et des méthodes de martingales, Houdré & Reynaud-Bouret [39] ont montré une inégalité de concentration pour des U-statistiques construites à partir de variables aléatoires réelles. En évaluant les quantités d'intérêt de cette inégalité, nous avons obtenu les résultats suivants.

Corollaire 2:
Soient $X, X_1, ..., X_n$ des variables aléatoires réelles et i.i.d, de densité s par rapport à la mesure de lebesgue μ. Soit $(\psi_\lambda)_{\lambda\in\Lambda}$ un système orthonormal dans $L^2(\mu)$ et soit S l'espace vectoriel qu'il engendre. Soit $B = \{t \in S,\ \|t\| \leq 1\}$, $v^2 = \sup_{t\in B} Var(t(X))$ et $b = \sup_{t\in B} \|t\|_\infty$. Soit

$$U(\Lambda) = \frac{1}{n(n-1)} \sum_{i\neq j=1}^{n} \sum_{\lambda\in\Lambda} (\psi_\lambda(X_i) - P\psi_\lambda)(\psi_\lambda(X_j) - P\psi_\lambda).$$

Alors, pour tout $x > 0$ et tout ξ dans $\{-1, 1\}$, il existe un évènement Ω_x tel que $\mathbb{P}(\Omega_x^c) \leq 2.8e^{-x}$ sur lequel on a,

$$\xi U(\Lambda) \leq 5.7vb\frac{\sqrt{x}}{n} + 8v^2\frac{x}{n} + 384vb\left(\frac{x}{n}\right)^{3/2} + 1020\left(\frac{bx}{n}\right)^2.$$

Nous utilisons cette inégalité dans le chapitre 5 où nous nous intéressons à des variables aléatoires réelles. Elle aurait suffit dans le chapitre 2. Nous avons préféré montrer une inégalité un peu plus générale, valable en toutes dimensions. Nous avons ainsi pu travailler avec des blocs dans le cas mélangeant. En utilisant notre inégalité de concentration de Z^2 et l'inégalité de Cauchy-Schwarz, on montre le corollaire suivant.

Corollaire 3:
Soient $X, X_1, ..., X_n$ des variables aléatoires i.i.d à valeurs dans un espace mesurable $(\mathbb{X}, \mathcal{X})$ de loi P. Soit μ une mesure sur $(\mathbb{X}, \mathcal{X})$ et soit $(t_\lambda)_{\lambda\in\Lambda}$ un ensemble de fonctions de $L^2(\mu)$. Soit

$$B = \{t = \sum_{\lambda\in\Lambda} a_\lambda t_\lambda,\ \sum_{\lambda\in\Lambda} a_\lambda^2 \leq 1\},\ D = \mathbb{E}\left(\sup_{t\in B}(t(X) - Pt)^2\right),$$

$$v^2 = \sup_{t\in B} Var(t(X)),\ b = \sup_{t\in B} \|t\|_\infty \ and\ \epsilon = \frac{b^2}{n}.$$

Soit

$$U = \frac{1}{n(n-1)} \sum_{i\neq j=1}^{n} \sum_{\lambda\in\Lambda} (t_\lambda(X_i) - Pt_\lambda)(t_\lambda(X_j) - Pt_\lambda).$$

Alors, on a

$$\forall x > 0,\ \mathbb{P}\left(U > \frac{5.31D^{3/4}(\epsilon x^2)^{1/4} + 3\sqrt{v^2Dx} + 3v^2x + \epsilon(19.1x)^2}{n-1}\right) \leq 2e^{-x}.$$

$$\forall x>0,\ \mathbb{P}\left(U<-\frac{9D^{3/4}(\epsilon x^2)^{1/4}+7.61\sqrt{v^2Dx}+\epsilon(40.3x)^2}{n-1}\right)\leq 3.8e^{-x}.$$

Processus empirique rééchantillonné

Pour étudier les pénalités et les régions de confiance construites par rééchantillonnage, nous avons dû étudier la version rééchantillonnée du processus Z^2 introduit précédemment. Nous avons utilisé une remarque de Arlot [5] pour montrer le lemme suivant.

Lemme:
Soient $X_1,...,X_n$ des variables aléatoires i.i.d de loi P. Soit $(t_\lambda)_{\lambda\in\Lambda}$ une collection de fonctions de $L^2(\mu)$. Soit $p(\Lambda)=\sum_{\lambda\in\Lambda}(\nu_n(t_\lambda))^2$. Soit $(W_1,...,W_n)$ un schéma de rééchantillonnage, soit $\bar{W}_n=\sum_{i=1}^n W_i/n$ et soit $v_W^2=Var(W_1-\bar{W}_n)$. Soit

$$p^W(\Lambda)=(v_W^2)^{-1}\sum_{\lambda\in\Lambda}\mathbb{E}^W\left((\nu_n^W(t_\lambda))^2\right),$$

$T=\sum_{\lambda\in\Lambda}(t_\lambda-Pt_\lambda)^2$ et

$$U=\frac{1}{n(n-1)}\sum_{i\neq j=1}^{n}\sum_{\lambda\in\Lambda}(t_\lambda(X_i)-Pt_\lambda)(t_\lambda(X_j)-Pt_\lambda).$$

Alors, on a

$$p(\Lambda)=\frac{1}{n}P_nT+\frac{n-1}{n}U,\ p^W(\Lambda)=\frac{1}{n}P_nT-\frac{1}{n}U,\ p(\Lambda)-p^W(\Lambda)=U.$$

Ce lemme, les inégalités de concentration précédentes et l'inégalité de Cauchy-Schwarz suffisent à étudier les quantités définies par rééchantillonnage. En particulier, on déduit de ce lemme deux propriétés fondamentales des pénalités par rééchantillonnage pen.
-La moyenne de la pénalité idéale coïncide avec celle de pen.
-La concentration de pen autour de sa moyenne fait intervenir les mêmes quantités que celle de pen$_{id}$, elle est même souvent meilleure (effet régularisant du bootstrap).

Conclusion

Les principaux apports de la thèse concernent le problème de sélection de modèles optimale pour des données indépendantes (Chapitre 2) et mélangeantes (Chapitre 4). Dans ces deux chapitres, nous proposons des pénalités facilement utilisables en pratique et très performantes en théorie.
Dans les chapitres 3 et 5, nous démontrons de nombreux lemmes techniques, qui ont permis de comprendre la méthode de couplage pour les données mélangeantes (Chapitre 3) et le comportement du processus empiriques et du processus de rééchantillonnage sur les fonctionnelles d'intérêt dans le modèle de densité (Chapitre 5). Le Chapitre 5 peut ainsi être lu avant le chapitre 2 pour se familiariser avec le rééchantillonnage dans le modèle de densité. Le chapitre 3 fournit une bonne introduction à l'extension de preuves faites pour des données indépendantes par la technique de couplage. Chaque chapitre correspond à un article, ils peuvent donc être lus de façon indépendante.

Chapter 2

Optimal model selection in density estimation

Abstract

We build penalized least-squares estimators using the slope heuristic and resampling penalties. We prove oracle inequalities for the selected estimator with leading constant asymptotically equal to 1. We compare the practical performances of these methods in a short simulation study.

Key words: Density estimation, optimal model selection, resampling methods, slope heuristic.
2000 Mathematics Subject Classification: 62G07, 62G09.

2.1 Introduction

The aim of model selection is to construct data-driven criteria to select a model among a given list. The history of statistical model selection goes back at least to Akaike [1], [2] and Mallows [53]. They proposed to select among a collection of parametric models the one which minimizes an empirical loss plus some penalty term proportional to the dimension of the model. Birgé & Massart [15] and Barron, Birgé & Massart [10] generalized this approach, making in particular the link between model selection and adaptive estimation. They proved that previous methods, in particular cross-validation (see Rudemo [62]) and hard thresholding (see Donoho *et.al.* [27]) can be viewed as penalization methods. More recently, Birgé & Massart [17], Arlot & Massart [7] and Arlot [5], (see also [4]) arised the problem of optimal efficient model selection. Basically, the aim is to select an estimator satisfying an oracle inequality with leading constant asymptotically equal to 1. They obtained such procedures thanks to a sharp estimator of the ideal penalty pen_{id}. We will be interested in two natural ideas, that are used in practice to evaluate pen_{id} and proved to be efficient in other frameworks. The first one is the slope heuristic. It was introduced in Birgé & Massart [17] in Gaussian regression and developed in Arlot & Massart [7] in a M-estimation framework. It allows to optimize the choice of a leading constant in the penalty term, provided that we know the shape of pen_{id}. The other one is Efron's resampling heuristic. The basic idea comes from Efron [31] and was used by Fromont [32] in the classification framework. Then, Arlot

[5] made the link with ideal penalties and developed the general procedure. Up to our knowledge, these methods have only been theoretically validated in regression frameworks. We propose here to prove their efficiency in density estimation. Let us now explain more precisely our context.

2.1.1 Least-squares estimators

In this chapter, we define and study efficient penalized least-squares estimators in the density estimation framework when the error is measured with the L^2-loss. We observe n i.i.d random variables $X_1, ..., X_n$, defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, valued in a measurable space $(\mathbb{X}, \mathcal{X})$, with common law P. We assume that a measure μ on $(\mathbb{X}, \mathcal{X})$ is given and we denote by $L^2(\mu)$ the Hilbert space of square integrable real valued functions defined on $\mathbb{X}$. $L^2(\mu)$ is endowed with its classical scalar product, defined for all t, t' in $L^2(\mu)$ by

$$< t, t' >= \int_{\mathbb{X}} t(x)t'(x)d\mu(x)$$

and the associated L^2-norm $\|.\|$, defined for all t in $L^2(\mu)$ by $\|t\| = \sqrt{< t, t >}$. The parameter of interest is the density s of P with respect to μ, we assume that it belongs to $L^2(\mu)$. The risk of an estimator $\hat{s}$ of s is measured with the L^2-loss, that is $\|s - \hat{s}\|^2$, which is random when $\hat{s}$ is.
s minimizes the integrated quadratic contrast $PQ(t)$, where $Q : L^2(\mu) \to L^1(P)$ is defined for all t in $L^2(\mu)$ by $Q(t) = \|t\|^2 - 2t$. Hence, density estimation is a problem of M-estimation. These problems are classically solved in two steps. First, we choose a "model" S_m that should be close to the parameter s, which means that $\inf_{t \in S_m} \|s - t\|^2$ is "small". Then, we minimize over S_m the empirical version of the integrated contrast, that is, we choose

$$\hat{s}_m \in \arg\min_{t \in S_m} P_n Q(t). \tag{2.1}$$

This last minimization can be computationaly untractable for general sets S_m, leading to untractable procedures in practice. However, it can be easily solved when S_m is a linear subspace of $L^2(\mu)$ since, for all orthonormal basis $(\psi_\lambda)_{\lambda \in m}$, we have

$$\hat{s}_m = \sum_{\lambda \in m} (P_n \psi_\lambda)\psi_\lambda. \tag{2.2}$$

Thus, we will always assume that a model is a linear subspace in $L^2(\mu)$. The risk of the least-squares estimator $\hat{s}_m$ defined in (2.1) is then decomposed in two terms, called bias and variance, thanks to Pythagoras relation. Let s_m be the orthogonal projection of s onto S_m,

$$\|s - \hat{s}_m\|^2 = \|s - s_m\|^2 + \|s_m - \hat{s}_m\|^2.$$

The statistician should choose a space S_m realizing a trade-off between those terms. S_m must be sufficiently "large" to ensure a small bias $\|s - s_m\|^2$, but not too much, for the variance $\|s_m - \hat{s}_m\|^2$ not to explose. The best trade-off depends on unknown properties of s, since the bias is unknown, and on the behavior of the empirical minimizer $\hat{s}_m$ in the space S_m. Classically, S_m is a parametric space and the dimension

d_m of S_m as a linear space is used to give upper bounds on $D_m = n\mathbb{E}\left(\|s_m - \hat{s}_m\|^2\right)$. This approach is validated in regular models under the assumption that the support of s is a known compact, as mentioned in section 2.3. However, this definition can fail dramatically because there exist simple models (histograms with a small dimension d_m) where D_m is very large, and infinite dimensional models where D_m is easily upper bounded. This issue is extensively discussed in Birgé [14]. Birgé chooses to keep the dimension d_m of S_m as a complexity measure and build new estimators that achieve better risk bounds than the empirical minimizer. His procedures are unfortunatly untractable for the practical user because he can only prove the existence of his estimators. Even his bounds on the risk are only interesting theoretically because they involve constants which are not optimal. We will not take this point of view here and our estimator will always be the empirical minimizer, mainly because it can easily be computed, see (2.2). We will focus on the quantity D_m/n and introduce a general Assumption (namely Assumption [**V**]) that allows to work indifferently with D_m/n or with the actual risk $\|s_m - \hat{s}_m\|^2$. We will also provide and study an estimator of D_m/n based on the resampling heuristic.
We insist here on the fact that, unlike classical methods, we will not use in this chapter strong extra assumptions on s, like $\|s\|_\infty < \infty$ or assume that s is compactly supported.

2.1.2 Model selection

Recall that the choice of an optimal model S_m is impossible without strong assumptions on the function s, for example that we have precise informations on the regularity of s is regular. However, under less restrictive hypotheses, for example that s is regular with an unknown regularity, we can build a countable collection of models $(S_m)_{m\in\mathcal{M}_n}$, growing with the number of observations, such that the best estimator in the associated collection $(\hat{s}_m)_{m\in\mathcal{M}_n}$ realizes an optimal trade-off, see for example Birgé & Massart [15] and Barron, Birgé & Massart [10]. The aim is then to build an estimator $\hat{m}$ such that our final estimator, $\tilde{s} = \hat{s}_{\hat{m}}$ behaves almost as well as any model m_o in the set of oracles

$$\mathcal{M}_n^* = \{m_o \in \mathcal{M}_n,\ \|\hat{s}_{m_o} - s\|^2 = \inf_{m\in\mathcal{M}_n} \|\hat{s}_m - s\|^2\}.$$

This is the problem of model selection. More precisely, we want that $\tilde{s}$ satisfies an oracle inequality defined in general as follows.

Definition: *(Trajectorial oracle inequality) Let $(p_n)_{n\in\mathbb{N}}$ be a sequence such that $\sum_{n\in\mathbb{N}} p_n < \infty$, let $(C_n)_{n\in\mathbb{N}}$ and $(R_{m,n})_{n\in\mathbb{N}}$ be sequences of positive real numbers. The estimator $\tilde{s} = \hat{s}_{\hat{m}}$ satisfies a trajectorial oracle inequality $TO(C_n, (R_{m,n})_{m\in\mathcal{M}_n}, p_n)$ if*

$$\forall n \in \mathbb{N}^*,\ \mathbb{P}\left(\|\tilde{s} - s\|^2 > C_n \inf_{m\in\mathcal{M}_n} \left\{\|s - \hat{s}_m\|^2 + R_{m,n}\right\}\right) \le p_n. \qquad (2.3)$$

When $\tilde{s}$ satisfies an oracle inequality, C_n is called the leading constant.

In this chapter, we are interested in the problem of optimal model selection defined as follows.

Definition: *(Optimal model selection) We say that $\tilde{s}$ is optimal or that the procedure of selection $(X_1, ..., X_n) \mapsto \hat{m}$ is optimal when $\tilde{s}$ satisfies a trajectorial oracle* *inequality $TO(1+r_n, (R_{m,n})_{m\in\mathcal{M}_n}, p_n)$ with $r_n \to 0$ and for all n in $\mathbb{N}^*$ and m in $\mathcal{M}_n$ $R_{m,n}=0$. In order to simplify the notations, when $\tilde{s}$ is optimal we will say that $\tilde{s}$ satisfies an optimal oracle inequality $OTO(r_n, p_n)$.*

In order to build $\hat{m}$, we remark that, for all m in $\mathcal{M}_n$, we have

$$\|s-\hat{s}_m\|^2 = \|\hat{s}_m\|^2 - 2P\hat{s}_m + \|s\|^2 = P_nQ(\hat{s}_m) + 2\nu_n(\hat{s}_m) + \|s\|^2, \tag{2.4}$$

where $\nu_n = P_n - P$ is the centered empirical process. An oracle minimizes $\|s-\hat{s}_m\|^2$ and thus $P_nQ(\hat{s}_m) + 2\nu_n(\hat{s}_m)$. As we want to imitate the oracle, we will design a map pen : $\mathcal{M}_n \to \mathbb{R}^+$ and choose

$$\hat{m} \in \arg\min_{m\in\mathcal{M}_n} P_nQ(\hat{s}_m) + \text{pen}(m),\ \tilde{s} = \hat{s}_{\hat{m}}. \tag{2.5}$$

It is clear that the ideal penalty is $\text{pen}_{id}(m) = 2\nu_n(\hat{s}_m)$. For all m in $\mathcal{M}_n$, for all orthonormal basis $(\psi_\lambda)_{\lambda\in m}$, we have $\hat{s}_m = \sum_{\lambda\in m}(P_n\psi_\lambda)\psi_\lambda$ and $s_m = \sum_{\lambda\in m}(P\psi_\lambda)\psi_\lambda$, thus

$$\nu_n(\hat{s}_m - s_m) = \nu_n\left(\sum_{\lambda\in m}(\nu_n\psi_\lambda)\psi_\lambda\right) = \sum_{\lambda\in m}(\nu_n\psi_\lambda)^2 = \|\hat{s}_m - s_m\|^2.$$

Let us define, for all m in $\mathcal{M}_n$

$$p(m) = \nu_n(\hat{s}_m - s_m) = \|\hat{s}_m - s_m\|^2.$$

From (2.4), for all m in $\mathcal{M}_n$, we have

$$\begin{aligned}\|s-\tilde{s}\|^2 &= \|\tilde{s}\|^2 - 2P\tilde{s} + \|s\|^2 = \|\tilde{s}\|^2 - 2P_n\tilde{s} + 2\nu_n\tilde{s} + \|s\|^2\\ &\le P_nQ(\hat{s}_m) + \text{pen}(m) + (2\nu_n(\tilde{s}) - \text{pen}(\hat{m})) + \|s\|^2\\ &= \|s-\hat{s}_m\|^2 + (\text{pen}(m) - 2\nu_n(\hat{s}_m)) + (2\nu_n(\tilde{s}) - \text{pen}(\hat{m}))\end{aligned}$$

Hence, for all m in $\mathcal{M}_n$,

$$\|s-\tilde{s}\|^2 \le \|s-\hat{s}_m\|^2 + (\text{pen}(m) - 2p(m)) + (2p(\hat{m}) - \text{pen}(\hat{m})) + 2\nu_n(s_{\hat{m}} - s_m). \tag{2.6}$$

Let us define, for all $c_1, c_2 > 0$, the function

$$f_{c_1,c_2} : \mathbb{R}^+ \to \mathbb{R}^+,\ x \mapsto \left\{\begin{array}{ll} \frac{1+c_1x}{1-c_2x} - 1 & \text{if } x < 1/c_2\\ +\infty & \text{if } x \ge 1/c_2\end{array}\right. . \tag{2.7}$$

It comes from inequality (2.6) that $\tilde{s}$ satisfies an oracle inequality $OTO(f_{2,2}(\epsilon_n), p_n)$ as soon as we have, with probability larger than $1-p_n$

$$\forall m \in \mathcal{M}_n\ \frac{|2p(m) - \text{pen}(m)|}{\|s-\hat{s}_m\|^2} \le \epsilon_n \text{ and} \tag{2.8}$$

$$\forall (m,m') \in \mathcal{M}_n^2,\ \frac{2\nu_n(s_{m'} - s_m)}{\|s-\hat{s}_{m'}\|^2 + \|s-\hat{s}_m\|^2} \le \epsilon_n. \tag{2.9}$$

Inequality (2.9) does not depend on our choice of penalty, we will check that it can easily be satisfied in classical collections of models. In order to obtain inequality (2.8), we use two methods, defined in M-estimation, but studied only on some regression frameworks.

The slope heuristic

The first one is refered as the "slope heuristic". The idea has been introduced by Birgé & Massart [17] in the Gaussian regression framework and developed in a general algorithm by Arlot & Massart [7]. This heuristic states that there exist a sequence $(\Delta_m)_{m\in\mathcal{M}_n}$ and a constant $K_{\min}$ satisfying the following properties,

1. when $\text{pen}(m) < K_{\min}\Delta_m$, then $\Delta_{\hat{m}}$ is too large, typically $\Delta_{\hat{m}} \geq C\max_{m\in\mathcal{M}_n}\Delta_m$,
2. when $\text{pen}(m) \simeq (K_{\min}+\delta)\Delta_m$ for some $\delta > 0$, then $\Delta_{\hat{m}}$ is much smaller,
3. when $\text{pen}(m) \simeq 2K_{\min}\Delta_m$, the selected estimator is optimal.

Thanks to the third point, when Δ_m and $K_{\min}$ are known, this heuristic says that the penalty $\text{pen}(m) = 2K_{\min}\Delta_m$ selects an optimal estimator. When Δ_m only is known, the first and the second point can be used to calibrate $K_{\min}$ in practice. We use the following algorithm, see Arlot & Massart [7]:

Slope algorithm
For all $K > 0$, compute the selected model $\hat{m}(K)$ given by (2.5) with the penalty $\text{pen}(m) = K\Delta_m$ and the associated complexity $\Delta_{\hat{m}(K)}$.
Find the constant $K_{\min}$ such that $\Delta_{\hat{m}(K)}$ is large when $K < K_{\min}$, and "much smaller" when $K > K_{\min}$.
Take the final $\hat{m} = \hat{m}(2K_{\min})$.

We will justify the slope heuristic in the density estimation framework for $\Delta_m = \mathbb{E}(\|s_m - \hat{s}_m\|^2) = D_m/n$ and $K_{\min} = 1$. In general, D_m is unknown and has to be estimated, we propose a resampling estimator and prove that it can be used without extra assumptions to obtain optimal results.

Resampling penalties

Data-driven penalties have already been used in density estimation in particular cross-validation methods as in Stone [64], Rudemo [62] or Celisse [21]. We are interested here in the resampling penalties introduced by Arlot [5]. Let $(W_1, ..., W_n)$ be a resampling scheme, i.e. a vector of random variables independent of $X, X_1, ..., X_n$ and exchangeable, that is, for all permutations τ of $(1, ..., n)$, we have

$$(W_1, ..., W_n) \text{ has the same law as } (W_{\tau(1)}, ..., W_{\tau(n)}).$$

Hereafter, we denote by $\bar{W}_n = \sum_{i=1}^n W_i/n$ and by E^W and $\mathcal{L}^W$ respectively the expectation and the law conditionally to the data $X, X_1, ..., X_n$. Let $P_n^W = \sum_{i=1}^n W_i\delta_{X_i}/n$, $\nu_n^W = P_n^W - \bar{W}_nP_n$ be the resampled empirical processes. Arlot's procedure is based on the resampling heurististic of Efron (see Efron [30]), which states that the law of a functional $F(P, P_n)$ is close to its resampled counterpart, that is the conditional law $\mathcal{L}^W(C_WF(\bar{W}_nP_n, P_n^W))$. C_W is a renormalizing constant that depends only on the resampling scheme and on F. Following this heuristic, Arlot defines as a penalty the resampling estimate of the ideal penalty $2D_m/n$, that is

$$\text{pen}(m) = 2C_W\mathbb{E}^W(\nu_n^W(\hat{s}_m^W)), \tag{2.10}$$

where $\hat{s}_m^W$ minimizes $P_n^WQ(t)$ over S_m. We prove concentration inequalities for $\text{pen}(m)$ and deduce that $\text{pen}(m)$ provides an optimal procedure.

The chapter is organized as follows. In Section 2.2, we state our main results, we prove the efficiency of the slope algorithm and the resampling penalties.
In Section 2.3, we compute the rates of convergence in the oracle inequalities using classical collections of models. Section 2.4 is devoted to a short simulation study where we compare different methods in practice. The proofs are postponed to Section 2.5. Section 2.6 is an Appendix where we add some probabilistic material, we prove a concentration inequality for Z^2, where $Z = \sup_{t\in B} \nu_n(t)$ and B is symmetric. We deduce a simple concentration inequality for U-statistics of order 2 that extends a previous result by Houdré & Reynaud-Bouret [39].

2.2 Main results

Hereafter, we will denote by c, C, K, κ, L, α, with various subscripts some constants that may vary from line to line.

2.2.1 Concentration of the ideal penalty

Take an orthonormal basis $(\psi_\lambda)_{\lambda\in m}$ of S_m. Easy algebra leads to

$$s_m = \sum_{\lambda\in m}(P\psi_\lambda)\psi_\lambda,\ \hat{s}_m = \sum_{\lambda\in m}(P_n\psi_\lambda)\psi_\lambda,\ \text{thus } \|s_m - \hat{s}_m\|^2 = \sum_{\lambda\in m}(\nu_n(\psi_\lambda))^2.$$

$\hat{s}_m$ is an unbiased estimator of s_m and

$$\mathrm{pen}_{id}(m) = 2\nu_n(\hat{s}_m) = 2\nu_n(\hat{s}_m - s_m) + 2\nu_n(s_m) = 2\|s_m - \hat{s}_m\|^2 + 2\nu_n(s_m).$$

For all m, m' in $\mathcal{M}_n$, let

$$p(m) = \|s_m - \hat{s}_m\|^2 = \sum_{\lambda\in m}(\nu_n(\psi_\lambda))^2,\ \delta(m, m') = 2\nu_n(s_m - s_{m'}). \tag{2.11}$$

From (2.6), for all m in $\mathcal{M}_n$,

$$\|s - \tilde{s}\|_2^2 \leq \|s - \hat{s}_m\|_2^2 + (\mathrm{pen}(m) - 2p(m)) + (2p(\hat{m}) - \mathrm{pen}(\hat{m})) + \delta(\hat{m}, m). \tag{2.12}$$

In this section, we are interested in the concentration of $p(m)$ around $\mathbb{E}(p(m)) = D_m/n$. Let us first remark that, for all m in $\mathcal{M}_n$, $p(m)$ is the supremum of the centered empirical process over the ellipsoid $B_m = \{t \in S_m,\ \|t\| \leq 1\}$. From Cauchy-Schwarz inequality, for all real numbers $(b_\lambda)_{\lambda\in m}$,

$$\sum_{\lambda\in m} b_\lambda^2 = \left(\sup_{\sum a_\lambda^2\leq 1}\sum_{\lambda\in m} a_\lambda b_\lambda\right)^2. \tag{2.13}$$

We apply this inequality with $b_\lambda = \nu_n(\psi_\lambda)$. We obtain, since the system $(\psi_\lambda)_{\lambda\in m}$ is orthonormal,

$$\sum_{\lambda\in m}(\nu_n(\psi_\lambda))^2 = \sup_{\sum a_\lambda^2\leq 1}\left(\sum_{\lambda\in m} a_\lambda \nu_n(\psi_\lambda)\right)^2 = \sup_{\sum a_\lambda^2\leq 1}\left(\nu_n\left(\sum_{\lambda\in m} a_\lambda\psi_\lambda\right)\right)^2 = \sup_{t\in B_m}(\nu_n(t))^2.$$

Hence, $p(m)$ is bounded by a Talagrand's concentration inequality (see Talagrand [65]). This inequality involves $D_m = n\mathbb{E}\left(\|\hat{s}_m - s_m\|^2\right)$ and the constants

$$e_m = \frac{1}{n}\sup_{t\in B_m}\|t\|_\infty^2 \text{ and } v_m^2 = \sup_{t\in B_m}\mathrm{Var}(t(X)). \tag{2.14}$$

More precisely, the following proposition holds:

Proposition 2.2.1 *Let $X, X_1, ..., X_n$ be iid random variables with common density s with respect to a probability measure μ. Assume that s belongs to $L^2(\mu)$ and let S_m be a linear subspace in $L^2(\mu)$. Let s_m and $\hat{s}_m$ be respectively the orthogonal projection and the projection estimator of s onto S_m. Let $p(m) = \|s_m - \hat{s}_m\|^2$, $D_m = n\mathbb{E}(p(m))$ and let v_m, e_m be the constants defined in (2.14). Then, for all $x > 0$,*

$$\mathbb{P}\left(p(m) - \frac{D_m}{n} > \frac{D_m^{3/4}(e_m x^2)^{1/4} + 0.7\sqrt{D_m v_m^2 x} + 0.15 v_m^2 x + e_m x^2}{n}\right) \le e^{-x/20} \tag{2.15}$$

$$\mathbb{P}\left(\frac{D_m}{n} - p(m) > \frac{1.8 D_m^{3/4}(e_m x^2)^{1/4} + 1.71\sqrt{D_m v_m^2 x} + 4.06 e_m x^2}{n}\right) \le 2.8 e^{-x/20} \tag{2.16}$$

Comments : From (2.12), for all m in $\mathcal{M}_n$,

$$\begin{aligned}\|s-\tilde{s}\|_2^2 \;\le\; & \|s-\hat{s}_m\|_2^2 + \left(\mathrm{pen}(m) - 2\frac{D_m}{n}\right) + 2\left(\frac{D_m}{n} - p(m)\right) \\ & +2\left(p(\hat{m}) - \frac{D_{\hat{m}}}{n}\right) + \left(2\frac{D_{\hat{m}}}{n} - \mathrm{pen}(\hat{m})\right) + \delta(\hat{m}, m).\end{aligned} \tag{2.17}$$

It appears from (2.17) that we can obtain oracle inequalities with a penalty of order $2D_m/n$ if, uniformly over m, m' in $\mathcal{M}_n$,

$$p(m) - \frac{D_m}{n} << \|s - \hat{s}_m\|^2 \text{ and } \delta(m', m) << \|s - \hat{s}_m\|^2 + \|s - \hat{s}_{m'}\|^2.$$

Proposition 2.2.1 proves that the first part holds with large probability for all m in $\mathcal{M}_n$ such that $e_m \vee v_m^2 << n\mathbb{E}(\|s - \hat{s}_m\|^2)$. Actually, the other part also holds under the same kind of assumption.

2.2.2 Main assumptions

For all m, m' in $\mathcal{M}_n$, let

$$\frac{R_m}{n} = \mathbb{E}\left(\|s - \hat{s}_m\|^2\right) = \|s - s_m\|^2 + \frac{D_m}{n},$$

$$v_{m,m'}^2 = \sup_{t\in S_m + S_{m'}, \|t\|\le 1}\mathrm{Var}(t(X)), e_{m,m'} = \frac{1}{n}\sup_{t\in S_m+S_{m'},\|t\|\le 1}\|t\|_\infty^2 .$$

For all $k \in \mathbb{N}$, let $\mathcal{M}_n^k = \{m \in \mathcal{M}_n,\ R_m \in [k, k+1)\}$ and for all n in $\mathbb{N}$, for all $k > 0$, $k' > 0$ and $\gamma \ge 0$, let

$$l_{n,\gamma}(k,k') = \ln(1 + \mathrm{Card}(\mathcal{M}_n^{[k]})) + \ln(1 + \mathrm{Card}(\mathcal{M}_n^{[k']})) + \ln((k+1)(k'+1)) + (\ln n)^\gamma \tag{2.18}$$

Assumption [V]: *There exist $\gamma > 1$ and a sequence $(\epsilon_n)_{n\in\mathbb{N}}$, with $\epsilon_n \to 0$ such that, for all n in $\mathbb{N}$, we have*

$$\sup_{(k,k')\in(\mathbb{N}^*)^2}\sup_{(m,m')\in\mathcal{M}_n^k\times\mathcal{M}_n^{k'}}\left\{\left(\left(\frac{v_{m,m'}^2}{R_m\vee R_{m'}}\right)^2\vee\frac{e_{m,m'}}{R_m\vee R_{m'}}\right)l_{n,\gamma}^2(k,k')\right\}\le\epsilon_n^4.$$

[BR] *There exist two sequences $(h_n^*)_{n\in\mathbb{N}^*}$ and $(h_n^o)_{n\in\mathbb{N}^*}$ with $(h_n^o\vee h_n^*)\to 0$ as $n\to\infty$ such that, for all n in $\mathbb{N}^*$, for all $m_o\in\arg\min_{m\in\mathcal{M}_n}R_m$ and all $m^*\in\arg\max_{m\in\mathcal{M}_n}D_m$, we have*

$$\frac{R_{m_o}}{D_{m^*}}\le h_n^o,\ \frac{n\|s-s_{m^*}\|^2}{D_{m^*}}\le h_n^*.$$

Comments:

- Assumption **[V]** ensures that the fluctuations of the ideal penalty are uniformly small compared to the risk of the estimator $\hat{s}_m$. Note that for all k,k', $l_{n,\gamma}(k,k')\ge(\ln n)^\gamma$, thus, Assumption **[V]** holds only in typical non parametric situations where $R_n=\inf_{m\in\mathcal{M}_n}R_m\to\infty$ as $n\to\infty$.

- The slope heuristic states that the complexity $\Delta_{\hat{m}}$ of the selected estimator is too large when the penalty term is too small. A minimal assumption for this heuristic to hold with $\Delta_m=D_m$ would be that there exists a sequence $(\theta_n)_{n\in\mathbb{N}^*}$ with $\theta_n\to 0$ as $n\to\infty$ such that, for all n in $\mathbb{N}^*$, for all $m_o\in\arg\min_{m\in\mathcal{M}_n}\mathbb{E}\left(\|s-\hat{s}_m\|^2\right)$ and all $m^*\in\arg\max_{m\in\mathcal{M}_n}\mathbb{E}\left(\|s_m-\hat{s}_m\|^2\right)$, we have
$$D_{m_o}\le\theta_nD_{m^*}.$$
  Assumption **[BR]** is slightly stronger but will always hold in the examples (see Section 2.3).

In order to have an idea of the rates R_n,ϵ_n, h_n^*, h_n^o and θ_n, let us briefly consider the very simple following example:

Example HR: We assume that s is supported in $[0,1]$ and that $(S_m)_{m\in\mathcal{M}_n}$ is the collection of the regular histograms on $[0,1]$, with $d_m=1,...,n$ pieces. We will see in Section 2.3.2 that $D_m\sim d_m$ asymptotically, hence $D_{m^*}\simeq n$. Moreover, we assume that s is Hölderian and not constant so that there exist positive constants $c_l,c_u,\alpha_l,\alpha_u$ such that, for all m in $\mathcal{M}_n$, see for example Arlot [5],

$$c_ld_m^{-\alpha_l}\le\|s-s_m\|^2\le c_ud_m^{-\alpha_u}.$$

In Section 2.3.2, we prove that this assumption implies **[V]** with $\epsilon_n\le C\ln(n)n^{-1/(8\alpha_l+4)}$. Moreover, there exists a constant $C>0$ such that $R_{m_o}\le\inf_{m\in\mathcal{M}_n}(c_und_m^{-\alpha_u}+d_m)\le Cn^{-1/(2\alpha_u+1)}$, thus $R_{m_o}/D_m^*\le Cn^{1/(2\alpha_u+1)-1}=Cn^{-2\alpha_u/(2\alpha_u+1)}$. Since there exists $C>0$ such that $n\|s-s_{m^*}\|^2/D_{m^*}\le Cd_{m^*}^{-\alpha_u}=Cn^{-\alpha_u}$, **[BR]** holds with $h_n^o=Cn^{-2\alpha_u/(2\alpha_u+1)}$ and $h_n^*=Cn^{-\alpha_u}$.
Other examples can be found in Birgé & Massart [15], see also Section 2.3.

2.2.3 Results on the Slope Heuristic

Let us now turn to the slope heuristic presented in Section 2.1.2. The following theorem proves the first point of this heuristic.

Theorem 2.2.2 *(Minimal penalty) Let $\mathcal{M}_n$ be a collection of models satisfying* [**V**] *and* [**BR**] *and let $\epsilon_n^* = \epsilon_n \vee h_n^*$.*
Assume that there exists $0 < \delta_n < 1$ such that $0 \leq pen(m) \leq (1-\delta_n)D_m/n$. Let $\hat{m}, \tilde{s}$ be the random variables defined in (2.5) and let

$$c_n = \frac{\delta_n - 28\epsilon_n^*}{1+16\epsilon_n}.$$

There exists a constant $C > 0$ such that, with probability at least $1 - Ce^{-\frac{1}{2}(\ln n)^\gamma}$,

$$\mathbb{P}\left(D_{\hat{m}} \geq c_n D_{m^*},\ \|s-\tilde{s}\|^2 \geq \frac{c_n}{5h_n^o} \inf_{m\in\mathcal{M}_n} \|s-\hat{s}_m\|^2 \right) \geq 1 - Ce^{-\frac{1}{2}(\ln n)^\gamma}. \tag{2.19}$$

Comments: Assume that $\delta_n = \delta > 0$, then, for all $n \geq n_o$, $c_n \geq c > 0$, thus $c_n/h_n^o \to \infty$, as $n \to \infty$. Hence, inequality (2.19) proves that an oracle inequality can not be obtained if pen(m) is not larger than D_m/n. Moreover, $D_{\hat{m}} \geq cD_{m^*}$ is as large as possible. This proves the first step of the slope heuristic.
The following theorem justifies the points 2 and 3 with $\Delta_m = D_m/n$, $K_{\min} = 1$.

Theorem 2.2.3 *Let $\mathcal{M}_n$ be a collection of models satisfying Assumption* [**V**]. *Assume that there exist $\delta_- > -1$, $\delta^+ > -1$ and $0 \leq p' < 1$ such that, with probability at least $1-p'$,*

$$2\frac{D_m}{n} + \delta_- \frac{R_m}{n} \leq pen(m) \leq 2\frac{D_m}{n} + \delta^+ \frac{R_m}{n}.$$

Let $\hat{m}, \tilde{s}$ be the random variables defined in (2.5) and let

$$C_n(\delta_-, \delta^+) = \left(\frac{1+\delta_- - 46\epsilon_n}{1+\delta^+ + 26\epsilon_n} \vee 0 \right)^{-1}.$$

There exists a constant $C > 0$ such that, with probability larger than $1 - p' - Ce^{-\frac{1}{2}(\ln n)^\gamma}$,

$$D_{\hat{m}} \leq C_n(\delta_-, \delta^+) R_{m_o},\ \|s-\tilde{s}\|^2 \leq C_n(\delta_-, \delta^+) \inf_{m\in\mathcal{M}_n} \|s-\hat{s}_m\|^2. \tag{2.20}$$

Comments :

- Assume that, for all $n \geq n_o$, $1+\delta_- \geq c > 0$, $\delta_+ \leq C < \infty$, then there exists $K > 0$, $n_1 \in \mathbb{N}^*$ such that, for all $n \geq n_1$, $C_n(\delta_-, \delta_+) \leq K$, thus $D_{\hat{m}}$ jumps from D_{m^*} (Theorem 2.2.2) to R_{m_o} (2.20) when pen(m) is around D_m/n. This proves the second step of the slope heuristic and clarifies what we meant by “much smaller”.

- The fact that a penalty term of order $2D_m/n$ leads to an efficient model selection procedure comes from inequality (2.20) applied with small δ_- and δ^+. The rate of convergence of the leading constant to 1 is then given by the supremum between δ_-, δ^+ and ϵ_n.

- The condition on the penalty has the same form as the one given in Arlot & Massart [7]. It comes from the fact that we do not know D_m/n in many cases, therefore, it has to be estimated. We propose two alternatives to solve this issue. In Section 2.2.4, we give a resampling estimator of D_m. It can be used for all collection of models satisfying **[V]** and its error of approximation is upper bounded by $\epsilon_n R_m/n$. Thus Theorem 2.2.3 holds with $(\delta_- \vee \delta^+) \leq C\epsilon_n$. In Section 2.3.2, we will see that, in regular models, we can use d_m instead of D_m and the error is upper bounded by CR_m/R_{m_o}, thus Theorem 2.2.3 holds with $(\delta_- \vee \delta^+) \leq C/R_{m_o} << \epsilon_n$, $p' = 0$. In both cases, we deduce from Theorem 2.2.3 that the estimator $\tilde{s}$ given by the slope algorithm achieves an optimal oracle inequality $OTO(\kappa\epsilon_n, Ce^{-\frac{1}{2}(\ln n)^\gamma})$. In Example **HR**, for example, we obtain $\epsilon_n = Cn^{-1/(8\alpha_l+4)}\ln n$.

2.2.4 Resampling penalties

Optimal model selection is possible in density estimation provided that we have a sharp estimation of $D_m = n\mathbb{E}\left(\sup_{t\in B_m}(\nu_n(t))^2\right)$. We propose an estimator of this quantity based on the resampling heuristic. The model selection algorithm that we deduce is the same as the resampling penalization procedure introduced by Arlot [5]. Let F be a fixed functional. Efron's heuristic states that the law $\mathcal{L}(F(\nu_n))$ is close to the conditional law $\mathcal{L}^W(C_W F(\nu_n^W))$, where C_W is a normalizing constant depending only on the resampling scheme and the functional F. Let $P_n^W = \sum_{i=1}^n W_i\delta_{X_i}/n$ and $\nu_n^W = P_n^W - \bar{W}_n P_n$. The resampling estimator of D_m is $D_m^W = nC_W^2\mathbb{E}^W\left(\sup_{t\in B_m}(\nu_n^W(t))^2\right)$ and the resampling penalty associated is $\mathrm{pen}(m) = 2D_m^W/n$. Actually, the following result describes the concentration of D_m^W around its mean D_m and around $np(m)$.

Proposition 2.2.4 *Let $(W_1, ..., W_n)$ be a resampling scheme, let S_m be a linear space, $B_m = \{t \in S_m,\ \|t\| \leq 1$, $D_m = n\mathbb{E}\left(\sup_{t\in B_m}(\nu_n(t))^2\right)$ and let D_m^W be its resampling estimator based on $(W_1, ..., W_n)$, that is $D_m^W = nC_W^2\mathbb{E}^W\left(\sup_{t\in B_m}(\nu_n^W(t))^2\right)$, where $v_W^2 = Var(W_1 - \bar{W}_n)$ and $C_W^2 = (v_W^2)^{-1}$, then, for all m in $\mathcal{M}_n$, $\mathbb{E}(D_m^W) = D_m$. Let e_m, v_m be the quantities defined in (2.14). For all $x > 0$, on an event of probability larger than $1 - 7.8e^{-x}$,*

$$D_m^W - D_m \leq \sqrt{8e_m D_m x} + e_m\left(\frac{4x}{3} + \frac{(40.3x)^2}{n-1}\right) + \frac{9D_m^{3/4}(e_m x^2)^{1/4} + 7.61\sqrt{v_m^2 D_m x}}{n-1}. \tag{2.21}$$

$$D_m^W - D_m \geq -\sqrt{8e_m D_m x} - e_m\left(\frac{4x}{3} + \frac{(19.1x)^2}{n-1}\right) - \frac{5.31D_m^{3/4}(e_m x^2)^{1/4} + 3\sqrt{v_m^2 D_m x} + 3v_m^2 x}{n-1}. \tag{2.22}$$

Moreover, for all $x > 0$, we have

$$\mathbb{P}\left(p(m) - \frac{D_m^W}{n} > \frac{5.31D_m^{3/4}(e_m x^2)^{1/4} + 3\sqrt{v_m^2 D_m x} + 3v_m^2 x + e_m(19.1x)^2}{n-1}\right) \leq 2e^{-x} \tag{2.23}$$

$$\mathbb{P}\left(\frac{D_m^W}{n} - p(m) \leq \frac{9D_m^{3/4}(e_m x^2)^{1/4} + 7.61\sqrt{v_m^2 D_m x} + e_m(40.3x)^2}{n-1}\right) \leq 3.8e^{-x}. \tag{2.24}$$

Remark
The concentration of the resampling estimator involves the same quantities as the concentration of $p(m)$, thus, it can be used to estimate the ideal penalty in the slope heuristic's algorithm presented in the previous section without extra assumptions on the collection $\mathcal{M}_n$. Proposition 2.2.4 and Theorem 2.2.3 prove that this resampling penalty leads to an efficient model selection procedure. However, we do not need to use the slope heuristic in our framework to obtain an optimal model selection procedure as shown by the following theorem.

Theorem 2.2.5 *Let $X_1, ..., X_n$ be i.i.d random variables with common density s. Let $\mathcal{M}_n$ be a collection of models satisfying Assumption* **[V]**. *Let $W_1, ..., W_n$ be a resampling scheme, let $\bar{W}_n = \sum_{i=1}^n W_i/n$, $v_W^2 = Var(W_1 - \bar{W}_n)$ and $C_W = 2(v_W^2)^{-1}$. Let $\tilde{s}$ be the penalized least-squares estimator defined in (2.5) with*

$$pen(m) = C_W \mathbb{E}^W\left(\sup_{t \in B_m}(\nu_n^W(t))^2\right).$$

Then, there exists a constant $C > 0$ such that

$$\mathbb{P}\left(\|s - \tilde{s}\|^2 \leq (1 + 100\epsilon_n) \inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2\right) \geq 1 - Ce^{-\frac{1}{2}(\ln n)^\gamma}. \tag{2.25}$$

Comments : The main advantage of this results is that the penalty term is always totally computable. Unlike the penalties derived from the slope heuristic, it does not depend on an arbitrary choice of a constant $K_{\min}$ made by the observer, that may be hard to detect in practice (see the paper of Alot & Massart [7] for an extensive discussion on this important issue). However, C_W is only optimal asymptotically. It is sometimes useful to overpenalize a little in order to improve the non-asymptotic performances of our procedures (see Massart [55]) and the slope heuristic can be used to do it in an optimal way (see our short simulation study in Section 2.4).

2.2.5 A remarks on the "regularization phenomenon"

The regularization of the bootstrap phenomenon (see Arlot [4, 5] and the references therein) states that the resampling estimator $C_W\mathbb{E}^W(F(\nu_n^W))$ of a functional $F(\nu_n)$ concentrates around its mean better than $F(\nu_n)$. This phenomenon can be justified with our previous results for our functional F. Recall that we have proven in Proposition 2.2.1 that, for all $x > 0$, with probability larger than $1 - 3.8e^{-x/20}$,

$$\left|p(m) - \frac{D_m}{n}\right| \leq \frac{1.8D_m^{3/4}(e_m x^2)^{1/4} + 1.5\sqrt{D_m v_m^2 x} + 0.2v_m^2 x + 4.1e_m x^2}{n}.$$

In Example **HR**, we have the following upper bounds

$$D_m \leq d_m,\ e_m \leq \frac{d_m}{n},\ v_m^2 \leq c\|s\|\sqrt{d_m}.$$

Thus, there exists a constant C such that, for all $x>0$,

$$\mathbb{P}\left(|np(m)-D_m|>Cd_m\left(\sqrt{\frac{x}{\sqrt{n}}}+\left(\frac{x}{\sqrt{n}}\right)^2\right)\right)\leq 3.8e^{-x/20}. \tag{2.26}$$

On the other hand, it comes from Inequalities (2.21) and (2.22), that, for all $x>0$, we have, on an event of probability larger than $1-7.8e^{-x/20}$,

$$\begin{aligned}\left|D_m^W-D_m\right| \leq\ & \sqrt{0.4e_mD_mx}+e_m\left(\frac{x}{15}+\frac{4.1x^2}{n-1}\right)\\ &+\frac{1.8D_m^{3/4}(e_mx^2)^{1/4}+1.45\sqrt{v_m^2D_mx}+0.2v_m^2x}{n-1}.\end{aligned}$$

Thus, there exists a constant C such that, for all $x>0$,

$$\mathbb{P}\left(\left|D_m^W-D_m\right|>Cd_m\left(\sqrt{\frac{x}{n}}+\left(\frac{x}{n}\right)^2\right)\right)\leq 7.8e^{-x/20}.$$

The concentration of D_m^W is then much better than the one of $np(m)$. This implies that D_m^W is an estimator of D_m rather than an estimator of $np(m)$. Thus, the resampling penalty can be used when D_m/n is a good penalty for example, under **[V]**. When D_m/n is known to underpenalized, see the examples in Barron, Birgé & Massart [10], there is no chance that D_m^W/n can work.

2.3 Rates of convergence in classical collections

The aim of this section is to show that **[V]** can be derived from a more classical hypothesis in two classical collections of models: the histograms and Fourier spaces. We derive the rates ϵ_n under this new hypothesis.

2.3.1 Assumption on the risk of the oracle

As mentioned in Section 2.2.2, Assumption **[V]** can only hold if there exists $\gamma>1$ such that $R_n(\ln n)^{-\gamma}\to\infty$ as $n\to\infty$, where $R_n=\inf_{m\in\mathcal{M}_n}R_m$. In our example, we will make the following Assumption that ensures that this condition is always satisfied.

[BR] *(Bounds on the Risk) There exist constants $C_u>0$, $\alpha_u>0$, $\gamma>1$, and a sequence $(\theta_n)_{n\in\mathbb{N}}$ with $\theta_n\to\infty$ as $n\to\infty$ such that, for all n in $\mathbb{N}^*$, for all m in $\mathcal{M}_n$*

$$\theta_n^2(\ln n)^{2\gamma}\leq R_n\leq R_m\leq C_un^{\alpha_u}.$$

Comments: Assumption **[BR]** holds with $\theta_n=Cn^\alpha$ for the collection of regular histograms of example **HR**, provided that s is an Hölderian, non constant and compactly supported function (see for example Arlot [4]). It is also a classical result of minimax theory that there exist functions in Sobolev spaces satisfying this kind of Assumption when $\mathcal{M}_n$ is the collection of Fourier spaces that we will introduce below.

We want to check that these collections satisfy Assumption [**V**], i.e. that there exists $\gamma > 1$ such that

$$\sup_{(k,k')\in(\mathbb{N}^*)^2} \sup_{(m,m')\in\mathcal{M}_n^k\times\mathcal{M}_n^{k'}} \left\{ \left(\left(\frac{v_{m,m'}^2}{R_m \vee R_{m'}} \right)^2 \vee \frac{e_{m,m'}}{R_m \vee R_{m'}} \right) l_{n,\gamma}^2(k,k') \right\} \leq \epsilon_n^4.$$

For all $m \in \mathcal{M}_n$, $R_m \leq C_u n^{\alpha_u}$, thus for all $k > C_u n^{\alpha_u}$, $\mathrm{Card}(\mathcal{M}_n^k) = 0$. In particular, we can assume in the previous supremum that $k \leq C_u n^{\alpha_u}$ and $k' \leq C_u n^{\alpha_u}$. Hence, there exists a constant $\kappa > 0$ such that $\ln[(1+k)(1+k')] \leq \kappa \ln n$. We also add the following assumption that ensures that there exists a constant $\kappa > 0$ such that, for all $k \in \mathbb{N}$, we have $\ln(1 + Card(\mathcal{M}_n^k)) \leq \kappa \ln n$.

[**PC**] *(Polynomial collection) There exist constants $c_{\mathcal{M}} \geq 0$, $\alpha_{\mathcal{M}} \geq 0$, such that, for all n in $\mathbb{N}$,*

$$\mathrm{Card}(\mathcal{M}_n) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}.$$

Under Assumptions [**BR**] and [**PC**], there exists a constant $\kappa > 0$ such that, for all $\gamma > 1$ and $n \geq 3$,

$$\sup_{(k,k')\in(\mathbb{N}^*)^2} \sup_{(m,m')\in\mathcal{M}_n^k\times\mathcal{M}_n^{k'}} \left\{ \left(\left(\frac{v_{m,m'}^2}{R_m \vee R_{m'}} \right)^2 \vee \frac{e_{m,m'}}{R_m \vee R_{m'}} \right) l_{n,\gamma}^2(k,k') \right\}$$
$$\leq \sup_{(m,m')\in(\mathcal{M}_n)^2} \left\{ \left(\frac{v_{m,m'}^2}{R_m \vee R_{m'}} \right)^2 \vee \frac{e_{m,m'}}{R_m \vee R_{m'}} \right\} \kappa (\ln n)^{2\gamma}.$$

2.3.2 The histogram case

Let $(\mathbb{X}, \mathcal{X})$ be a measurable space. Let $(P_m)_{m\in\mathcal{M}_n}$ be a collection of measurable partitions $P_m = (I_\lambda)_{\lambda\in m}$ of subsets of $\mathbb{X}$ such that, for all $m \in \mathcal{M}_n$, for all $\lambda \in m$, we have $0 < \mu(I_\lambda) < \infty$. Let m in $\mathcal{M}_n$, the set S_m of histograms associated to P_m is the set of functions which are constant on each I_λ, $\lambda \in m$. S_m is a linear space. Setting, for all $\lambda \in m$, $\psi_\lambda = (\sqrt{\mu(I_\lambda)})^{-1} 1_{I_\lambda}$, the functions $(\psi_\lambda)_{\lambda\in m}$ form an orthonormal basis of S_m.
Let us recall that, for all m in $\mathcal{M}_n$,

$$D_m = \sum_{\lambda\in m} \mathrm{Var}(\psi_\lambda(X)) = \sum_{\lambda\in m} P(\psi_\lambda^2) - (P\psi_\lambda)^2 = \sum_{\lambda\in m} \frac{P(X \in I_\lambda)}{\mu(I_\lambda)} - \|s_m\|^2. \quad (2.27)$$

Moreover, from Cauchy-Schwarz inequality, for all x in $\mathbb{X}$, for all m, m' in $\mathcal{M}_n$

$$\sup_{t\in B_{m,m'}} t^2(x) \leq \sum_{\lambda\in m\cup m'} \psi_\lambda^2(x), \text{ thus } e_{m,m'} = \frac{1}{n} \sup_{\lambda\in m\cup m'} \frac{1}{\mu(I_\lambda)}. \quad (2.28)$$

Finally, it is easy to check that, for all m ,m' in $\mathcal{M}_n$

$$v_{m,m'}^2 = \sup_{\lambda\in m\cup m'} \mathrm{Var}(\psi_\lambda(X)) = \sup_{\lambda\in m\cup m'} \frac{P(X \in I_\lambda)(1 - P(X \in I_\lambda))}{\mu(I_\lambda)}. \quad (2.29)$$

We will consider two particular types of histograms.
Example 1 [Reg] : μ-regular histograms.

For all m in $\mathcal{M}_n$, P_m is a partition of $\mathbb{X}$ and there exist a family $(d_m)_{m\in\mathcal{M}_n}$ bounded by n and two constants c_{rh}, C_{rh} such that, for all m in $\mathcal{M}_n$, for all $\lambda \in \mathcal{M}_n$,

$$\frac{c_{rh}}{d_m} \leq \mu(I_\lambda) \leq \frac{C_{rh}}{d_m}.$$

The typical example here is the collection described in Example **HR**.

Example 2 [Ada]: Adapted histograms.
There exist positive constants c_r, C_{ah} such that, for all m in $\mathcal{M}_n$, for all $\lambda \in \mathcal{M}_n$, $\mu(I_\lambda) \geq c_r n^{-1}$ and

$$\frac{P(X \in I_\lambda)}{\mu(I_\lambda)} \leq C_{ah}.$$

[**Ada**] is typically satisfied when s is bounded on $\mathbb{X}$. Remark that the models satisfying [**Ada**] have finite dimension $d_m \leq Cn$ since

$$1 \geq \sum_{\lambda\in m} P(X \in I_\lambda) \geq c_{ah} \sum_{\lambda \in m} \mu(I_\lambda) \geq c_{ah} c_r d_m n^{-1}.$$

The example [Reg].
It comes from equations (2.27, 2.28, 2.29) and Assumption [**Reg**] that

$$C_{rh}^{-1} d_m - \|s_m\|^2 \leq D_m \leq c_{rh}^{-1} d_m - \|s_m\|^2.$$

$$e_{m,m'} \leq c_{rh}^{-1} \frac{d_m \vee d_{m'}}{n}, \; v_{m,m'}^2 \leq \sup_{t\in B_{m,m'}} \|t\|_\infty \|t\| \|s\| \leq c_{rh}^{-1/2} \|s\| \sqrt{d_m \vee d_{m'}}.$$

Thus

$$\frac{e_{m,m'}}{R_m \vee R_{m'}} \leq C_{rh} c_{rh}^{-1} \frac{(R_m \vee R_{m'}) + \|s\|^2}{n(R_m \vee R_{m'})} \leq Cn^{-1}.$$

If $D_m \vee D_{m'} \leq \theta_n^2 (\ln n)^{2\gamma}$,

$$\frac{v_{m,m'}^2}{R_m \vee R_{m'}} \leq \sqrt{C_{rh} c_{rh}^{-1}} \frac{\sqrt{(D_m \vee D_{m'}) + \|s\|^2}}{R_{m_o}} \leq \frac{C}{\theta_n (\ln n)^\gamma}.$$

If $D_m \vee D_{m'} \geq \theta_n^2 (\ln n)^{2\gamma}$,

$$\frac{v_{m,m'}^2}{R_m \vee R_{m'}} \leq \sqrt{C_{rh} c_{rh}^{-1}} \frac{\sqrt{(D_m \vee D_{m'}) + \|s\|^2}}{D_m \vee D_{m'}} \leq \frac{C}{\theta_n (\ln n)^\gamma}.$$

There exists $\kappa > 0$ such that $\theta_n^2 (\ln n)^{2\gamma} \leq \kappa n$ since for all m in $\mathcal{M}_n$, $R_m \leq n\|s - s_m\|^2 + c_{rh}^{-1} d_m \leq (\|s\|^2 + c_{rh}^{-1})n$. Hence Assumption [**V**] holds with γ given in Assumption [**BR**] and $\epsilon_n = C\theta_n^{-1/2}$.

The example [Ada].
It comes from inequalities (2.28), (2.29) and Assumption [**Ada**] that, for all m and m' in $\mathcal{M}_n$

$$e_{m,m'} \leq c_r^{-1} \text{ and } v_{m,m'}^2 \leq C_{ah}.$$

Thus, there exists a constant $\kappa > 0$ such that, for all m an m' in $\mathcal{M}_n$, we have

$$\sup_{(m,m')\in(\mathcal{M}_n)^2}\left\{\left(\frac{v_{m,m'}^2}{R_m \vee R_{m'}}\right)^2 \vee \frac{e_{m,m'}}{R_m \vee R_{m'}}\right\} \leq \frac{\kappa}{\theta_n^2(\ln n)^{2\gamma}}.$$

Therefore Assumption [**V**] holds also with γ given in Assumption [**BR**] and $\epsilon_n = \kappa\theta_n^{-1/2}$.

2.3.3 Fourier spaces

In this section, we assume that s is supported in $[0,1]$. We introduce the classical Fourier basis. Let $\psi_0 : [0,1] \to \mathbb{R},\ x \mapsto 1$ and, for all $k \in \mathbb{N}^*$, we define the functions

$$\psi_{1,k} : [0,1] \to \mathbb{R},\ x \mapsto \sqrt{2}\cos(2\pi kx),\ \psi_{2,k} : [0,1] \to \mathbb{R},\ x \mapsto \sqrt{2}\sin(2\pi kx).$$

For all j in $\mathbb{N}^*$, let

$$m_j = \{0\} \cup \{(i,k),\ i = 1,2,\ k = 1,...,j\} \text{ and } \mathcal{M}_n = \{m_j, j = 1,...,n\}.$$

For all m in $\mathcal{M}_n$, let S_m be the space spanned by the family $(\psi_\lambda)_{\lambda\in m}$. $(\psi_\lambda)_{\lambda\in m}$ is an orthonormal basis of S_m and for all j in $1,...,n$, $d_{m_j} = 2j+1$.
Let j in $1,...n$, for all x in $[0,1]$,

$$\sum_{\lambda\in m_j}\psi_\lambda^2(x) = 1 + 2\sum_{k=1}^{j}\cos^2(2\pi kx) + \sin^2(2\pi kx) = 1 + 2j = d_{m_j}.$$

Hence, for all m in $\mathcal{M}_n$,

$$D_m = P\left(\sum_{\lambda\in m_j}\psi_\lambda^2\right) - \|s_m\|^2 = d_m - \|s_m\|^2. \tag{2.30}$$

It is also clear that, for all m, m' in $\mathcal{M}_n$,

$$e_{m,m'} = \frac{d_m \vee d_{m'}}{n},\ v_{m,m'}^2 \leq \|s\|\sqrt{d_m \vee d_{m'}}. \tag{2.31}$$

The collection of Fourier spaces of dimension $d_m \leq n$ satisfies Assumption [**PC**], and the quantities D_m $e_{m,m'}$ and $v_{m,m'}^2$ satisfy the same inequalities as in the collection [**Reg**], therefore, [**V**] comes also in this collection from [**BR**]. We have obtained the following corollary of Theorem 2.2.5.

Corollary 2.3.1 *Let $\mathcal{M}_n$ be either a collection of histograms satisfying Assumptions* [**PC**]-[**Reg**] *or* [**PC**]-[**Ada**] *or the collection of Fourier spaces of dimension $d_m \leq n$. Assume that s satisfies Assumption* [**BR**] *for some $\gamma > 1$ and $\theta_n \to \infty$. Then, there exist constants $\kappa > 0$ and $C > 0$ such that the estimator $\tilde{s}$ selected by a resampling penalty satisfies*

$$\mathbb{P}\left(\|s-\tilde{s}\|^2 \leq (1+\kappa\theta_n^{-1/2})\inf_{m\in\mathcal{M}_n}\|s-\hat{s}_m\|^2\right) \geq 1 - Ce^{-\frac{1}{2}(\ln n)^\gamma}.$$

Comment: Assumption [**BR**] is hard to check in practice. We mentioned that it holds in Example **HR** provided that s is Hölderian, non constant and compactly supported (see Arlot [5]). It is also classical to build functions satisfying [**BR**] with the Fourier spaces in order to prove that the oracle reaches the minimax rate of convergence over some Sobolev balls, see for example Birgé & Massart [15], Barron, Birgé & Massart [10] or Massart [55]. In these cases, there exist $c > 0$, $\alpha > 0$ such that $\theta_n \geq cn^{\alpha}$. In more general situations, we can use the same trick as Arlot [5] and use our main theorem only for the models with dimension $d_m \geq (\ln n)^{4+2\gamma}$, they satisfy [**BR**] with $\theta_n = (\ln n)^2$, at least when n is sufficiently large, because

$$\|s\|^2 + R_m \geq \|s\|^2 + D_m \geq cd_m \geq c(\ln n)^4(\ln n)^{2\gamma}.$$

With our concentration inequalities, we can control easily the risk of the models with dimension $d_m \leq (\ln n)^{4+2\gamma}$ by $\kappa(\ln n)^{3+5\gamma/2}$ with probability larger than $1 - Ce^{-\frac{1}{2}(\ln n)^{\gamma}}$ and we can then deduce the following corollary.

Corollary 2.3.2 *Let $\mathcal{M}_n$ be either a collection of histograms satisfying Assumptions* [**PC**]-[**Reg**] *or* [**PC**]-[**Ada**] *or the collection of Fourier spaces of dimension $d_m \leq n$. There exist constants $\kappa > 0$, $\eta > 3 + 5\gamma/2$ and $C > 0$ such that the estimator $\tilde{s}$ selected by a resampling penalty satisfies*

$$\mathbb{P}\left(\|s - \tilde{s}\|^2 \leq (1 + \kappa(\ln n)^{-1})\left(\inf_{m\in\mathcal{M}_n} \|s - \hat{s}_m\|^2 + \frac{(\ln n)^{\eta}}{n}\right)\right) \geq 1 - Ce^{-\frac{1}{2}(\ln n)^{\gamma}}.$$

2.4 Simulation study

We propose in this section to show the practical performances of the slope algorithm and the resampling penalties on two examples. We estimate the density $s(x) = (3/4)x^{-1/4}1_{[0,1]}(x)$ and we compare the three following methods.

1. The first one is the slope heuristic applied with the linear dimension d_m of the models. We observe two main behaviors of $d_{\hat{m}(K)}$ with respect to K. Most of the times, we only observe one jump, as in Figure 2.1, and we find $K_{\min}$ easily.

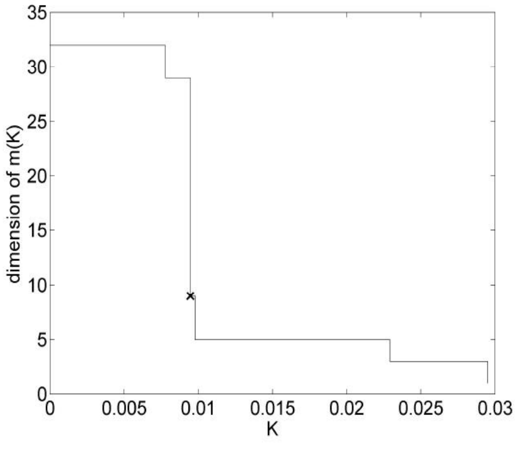


Figure 2.1: Classical behavior of $K \mapsto d_{\hat{m}(K)}$

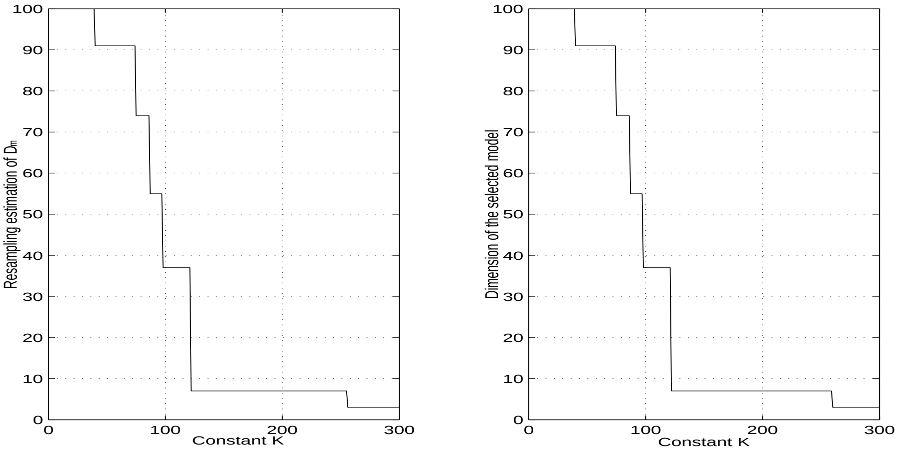


Figure 2.2: Comparison of d_m and D_m^W on the selected model

We also observe more difficult situations as the one of Figure 2.2 below, where we can see several jumps. In these cases, as prescribed in the regression framework by Arlot & Massart [7], we choose the constant $K_{\min}$ realizing the maximal jump of $d_{\hat{m}(K)}$. Arlot & Massart [7] also proposed to select $K_{\min}$ as the minimal K such that $d_{\hat{m}(K)} \leq d_{m^*}(\ln n)^{-1}$, but they obtained worse performances of the selected estimator in their simulations.
We justify this method only for collection of models where $d_m \simeq KD_m$ for some constant K. We will see that it gives really good performances when this condition is satisfied.

2. The second method is the resampling based penalization algorithm of Theorem 2.2.5. Note here that all the resampling penalties D_m^W/n can be easily computed, without any Monte Carlo approximations. Actually, for all resampling scheme,

$$\frac{D_m^W}{n} = \frac{1}{n}\sum_{\lambda\in m}\left(P_n\psi_\lambda^2 - \frac{1}{n(n-1)}\sum_{i\neq j=1}^{n}\psi_\lambda(X_i)\psi_\lambda(X_j)\right).$$

   Resampling penalties give always good approximations of D_m. However, in non asymptotic situations, it may be usefull to overpenalize a little bit in order to improve the leading constants in the oracle inequality (in Theorem 2.2.3, imagine that $46\epsilon_n$ is very close to 1).

3. In a third method, we propose therefore to use the slope algorithm applied with a complexity D_m^W. By this way, we hope to overpenalize a little bit the resampling penalty when it is necessary.

2.4.1 Example 1: regular case

In the first example, we consider the collection of regular histograms described in example **HR** and we observe $n = 100$ data. In this example, we saw that $D_m^W \simeq D_m \simeq d_m$. We can actually verify in Figure 2.2 that these quantities almost coincide for the selected model.

We compute $N = 1000$ times the oracle constant $c = \|s - \tilde{s}\|^2/(\inf_{m\in\mathcal{M}_n} \|s - \hat{s}_m\|^2)$ for the 3 methods. We put in the following array the mean, the median and the 0.95-quantile, $q_{0.95}$ of these quantities.

method	mean of the N constants c	median	$q_{0.95}$
slope + $\mathrm{d_m}$	3.56	2.30	10.07
resampling	4.43	2.52	15.47
resampling + slope	3.57	2.21	10.86

We observe that the slope algorithm allows to improve the resampling penalty in practice. This may be due to a little overpenalization even if it is not a straightforward consequence of our theoretical results. Note that, as $d_m \simeq D_m^W$, the slope algorithm leads to the same results when applied with d_m or with D_m^W. Although we have an explicite formula to compute the resampling penalties, the computation time is much longer if we use D_m^W. Therefore, we clearly recommand to use the slope algorithm with d_m for regular collections of model, as regular histograms or Fourier spaces described in Section 2.3.3.

2.4.2 Example 2: a more complicated collection

In the next example, we want to show that the linear dimension shall not be used in general. Let us consider a slightly more complicated collection. Let k, J_1, J_2, n be four non null integers satisfying $k \le n$, $J_1 \le k$, $J_2 \le n - k$. We denote by $S_{k,J_1,J_2,n}$ the linear space of histograms on the following partition.

$$\left\{\left[l\frac{k}{J_1 n}, (l+1)\frac{k}{J_1 n}\right[, \ l = 0, ..., J_1 - 1\right\}$$
$$\cup\left\{\left[\frac{k}{n} + l\frac{1-k/n}{J_2}, \frac{k}{n} + (l+1)\frac{1-k/n}{J_2}\right[, \ l = 0, ...J_2 - 1\right\}.$$

Let $n \in \mathbb{N}^*$ and let $\mathcal{M}_n = \{(k, J_1, J_2) \in (\mathbb{N}^*)^3;\ k \le n,\ J_1 \le k,\ J_2 \le n - k\}$. It is clear that $\mathrm{Card}(\mathcal{M}_n) \le n^3$. The oracle of this collection is better than the previous one since the regular histograms belongs to $(S_{m,n})_{m\in\mathcal{M}_n}$. It is easy to check that the dimension of $S_{k,J_1,J_2,n}$ is equal to $J_1 + J_2$ and that $D_{k,J_1,J_2,n}$ is equal to $(nJ_1/k)F(k/n) + (nJ_2/(n-k))(1 - F(k/n)) - \|s_{k,J_1,J_2,n}\|^2/n$, where F is the distribution function of the observations. Hence, there is no constant K_o such that $K_o d_{k,J_1,J_2,n} \simeq D_{k,J_1,J_2,n}$ as in the previous example. We also compute $N = 1000$ times the oracle constant $c = \|s - \tilde{s}\|^2/(\inf_{m\in\mathcal{M}_n} \|s - \hat{s}_m\|^2)$ for the 3 methods, taking $n = 100$ observations each time. The results are summarized in following array.

method	mean of the N constants c	median	$q_{0.95}$
slope + $\mathrm{d_m}$	8.30	7.01	19.73
resampling	6.11	5.08	13.52
resampling + slope	5.33	4.04	12.92

The slope heuristic gives bad results when applied with d_m. This is due to the fact that d_m is not proportional to D_m here. The resampling based penalty $2D_m^W/n$ is much better and, as in the regular case, it is well improved by the slope algorithm. Therefore, for general collections of models where we do not know an optimal shape of the ideal penalty, we recommand to apply the slope algorithm with a complexity equal to D_m^W.

2.5 Proofs

2.5.1 Proof of Proposition 2.2.1

It is a straightforward application of Corollary 2.6.6 in the appendix.

2.5.2 Technical lemmas

Before giving the proofs of the main theorems, we state and prove some important technical lemmas that we will use repeatedly all along the proofs. Let us recall here the main notations. For all m, m' in $\mathcal{M}_n$, we have

$$p(m) = \|s_m - \hat{s}_m\|^2,\ D_m = n\mathbb{E}(p(m)) = n\mathbb{E}\left(\|\hat{s}_m - s_m\|^2\right)$$

$$R_m = n\mathbb{E}\left(\|s - \hat{s}_m\|^2\right) = n\|s - s_m\|^2 + D_m,\ \delta(m, m') = \nu_n(s_m - s_{m'}).$$

For all $n \in \mathbb{N}^*,\ k > 0,\ k' > 0,\ \gamma > 0,$, let $[k]$ be the integer part of k and let

$$l_{n,\gamma}(k, k') = \ln((1 + \mathrm{Card}(\mathcal{M}_{\mathrm{n}}^{[\mathrm{k}]}))(1 + \mathrm{Card}(\mathcal{M}_{\mathrm{n}}^{[\mathrm{k}']}))) + \ln((1 + k)(1 + k')) + (\ln n)^\gamma.$$

Recall that, for all m, m' in $\mathcal{M}_n$, we have

$$\begin{aligned} v_{m,m'}^2 l_{n,\gamma}(R_m, R_{m'}) &\leq \epsilon_n^2(R_m \vee R_{m'}), \\ e_{m,m'}(l_{n,\gamma}(R_m, R_{m'}))^2 &\leq \epsilon_n^4(R_m \vee R_{m'}). \end{aligned} \tag{2.32}$$

Let us prove a simple result

Lemma 2.5.1 *For all $K > 1$,*

$$\Sigma(K) = \sum_{k\in\mathbb{N}} \sum_{m\in\mathcal{M}_n^k} e^{-K[\ln(1+\mathrm{Card}(\mathcal{M}_{\mathrm{n}}^{\mathrm{k}}))+\ln(1+k)]} < \infty. \tag{2.33}$$

For all m in $\mathcal{M}_n$, let $l_m = l_{n,\gamma}(R_m, R_m)$, then, for all $K > 1/\sqrt{2}$,

$$\sum_{m\in\mathcal{M}_n} e^{-K^2 l_m} = \Sigma(2K^2)e^{-K^2(\ln n)^\gamma}. \tag{2.34}$$

For all m, m' in $\mathcal{M}_n$, let $l_{m,m'} = l_{n,\gamma}(R_m, R_{m'})$, then, for all $K > 1$,

$$\sum_{(m,m')\in(\mathcal{M}_n)^2} e^{-K^2 l_{m,m'}} = (\Sigma(K^2))^2 e^{-K^2(\ln n)^\gamma}. \tag{2.35}$$

Proof :

Inequality (2.33) comes from the fact that, when $K > 1$,

$$\forall k \in \mathbb{N},\ \sum_{m\in\mathcal{M}_n^k} e^{-K[\ln(1+\mathrm{Card}(\mathcal{M}_n^k))]} \leq 1,\ \text{and} \sum_{k\in\mathbb{N}^*} e^{-K\ln k} < \infty.$$

For all k such that $\mathcal{M}_n^k \neq \emptyset$, for all m in $\mathcal{M}_n^k$, $l_m \geq 2[\ln(1 + \mathrm{Card}(\mathcal{M}_n^k)) + \ln(1 + k)] + (\ln n)^\gamma$, thus, for all $K > 1/\sqrt{2}$, it comes from (2.33) that

$$\sum_{m\in\mathcal{M}_n} e^{-K^2 l_m} \leq e^{-K^2(\ln n)^\gamma} \sum_{k\in\mathbb{N}} \sum_{m\in\mathcal{M}_n^k} e^{-2K^2[\ln(1+\mathrm{Card}(\mathcal{M}_n^k))+\ln(1+k)]} \leq \Sigma(2K^2)e^{-K^2(\ln n)^\gamma}.$$

Finally, for all (k,k') such that $\mathcal{M}_n^k \times \mathcal{M}_n^{k'} \neq \emptyset$,

$l_{m,m'} \geq \ln(1+\mathrm{Card}(\mathcal{M}_n^k)) + \ln(1+k) + \ln(1+\mathrm{Card}(\mathcal{M}_n^{k'})) + \ln(1+k') + (\ln n)^\gamma$,

thus, from (2.33),

$$\sum_{(m,m')\in(\mathcal{M}_n^2)} e^{-K^2 l_{m,m'}} = \left(\sum_{k\in\mathbb{N}}\sum_{m\in\mathcal{M}_n^k} e^{-K^2[\ln(1+\mathrm{Card}(\mathcal{M}_\mathrm{n}^\mathrm{k}))+\ln(1+k)]}\right)^2 e^{-K^2(\ln n)^\gamma}.$$

Lemma 2.5.2 *Let $\mathcal{M}_n$ be a collection of models satisfying Assumption* [**V**]*. We consider the following events.*

$$\begin{aligned}\Omega_\delta &= \left\{\forall (m,m')\in\mathcal{M}_n^2,\ \delta(m,m') \leq 6\epsilon_n \frac{R_m \vee R_{m'}}{n}\right\}\\ \Omega_p &= \bigcap_{m\in\mathcal{M}_n}\left\{\left\{p(m) - \frac{D_m}{n} \leq 10\epsilon_n \frac{R_m}{n}\right\} \cap \left\{p(m) - \frac{D_m}{n} \geq -20\epsilon_n\frac{R_m}{n}\right\}\right\}\end{aligned}$$

and $\Omega_T = \Omega_\delta \cap \Omega_p$. Then there exists a constant $C>0$ such that

$$\mathbb{P}(\Omega_\delta^c) \leq Ce^{-(\ln n)^\gamma},\ \mathbb{P}(\Omega_p^c) \leq Ce^{-\frac{1}{2}(\ln n)^\gamma},\ \mathbb{P}(\Omega_T^c) \leq Ce^{-\frac{1}{2}(\ln n)^\gamma}.$$

Proof :

Let $K>1$ be a constant to be chosen later. We apply Lemma 2.6.8 in the appendix to $u = s_m - s_{m'}$, $S = S_m + S_{m'}$, $L = id$, $x = K^2 l_{n,\gamma}(R_m, R_{m'})$. For all $\eta > 0$, for all m, m' in $\mathcal{M}_n$, on an event of probability larger than $1 - e^{-K^2 l_{n,\gamma}(R_m,R_{m'})}$,

$$\delta(m,m') \leq \frac{\eta}{2}\|s_m - s_{m'}\|^2 + \frac{2v_{m,m'}^2 K^2 l_{n,\gamma}(R_m,R_{m'}) + e_{m,m'}(K^2 l_{n,\gamma}(R_m,R_{m'}))^2/9}{\eta n}. \tag{2.36}$$

From [**V**], for all m, m' in $\mathcal{M}_n$,

$$2v_{m,m'}^2 K^2 l_{n,\gamma}(R_m,R_{m'})) + \frac{e_{m,m'}(K^2 l_{n,\gamma}(R_m,R_{m'}))^2}{9} \leq \left(2(K\epsilon_n)^2 + \frac{(K\epsilon_n)^4}{9}\right)\frac{R_m \vee R_{m'}}{n}.$$

Moreover, for all m, m' in $\mathcal{M}_n$, we have

$$\|s_m - s_{m'}\|^2 \leq 2(\|s - s_m\|^2 + \|s - s_{m'}\|^2) \leq 2(R_m + R_{m'}) \leq 4(R_m \vee R_{m'}).$$

Let $e_n(K) = \sqrt{(K\epsilon_n)^2 + (K\epsilon_n)^4/18}$. In (2.36) we take $\eta = e_n(K)$ and we obtain

$$\mathbb{P}\left(\delta(m,m') > 4e_n(K)\frac{R_m \vee R_{m'}}{n}\right) \leq e^{-Kl_{n,\gamma}(R_m,R_{m'})}. \tag{2.37}$$

From (2.35), for all $K>1$,

$$\mathbb{P}\left(\forall (m,m')\in\mathcal{M}_n^2,\ \delta(m,m') > 4e_n(K)\frac{R_m \vee R_{m'}}{n}\right) \leq (\Sigma(K))^2 e^{-K(\ln n)^2}.$$

Let $K = 1.1$ and take n sufficiently large so that $K^4\epsilon_n^2/18 \leq 1$, then $4e_n(K) \leq 6\epsilon_n$. Hence, the first conclusion of Lemma 2.5.2 holds for sufficiently large n, it holds in

general, provided that we increase the constant C if necessary.
We apply Assumption **[V]** (see (2.32)) with $m = m'$, let $l_m = l_{n,\gamma}(R_m, R_m)$, for all $K > 0$, for all n such that $4.06(K\epsilon_n)^3 \leq 2$, we have

$$\frac{D_m^{3/4}(e_m(K^2 l_m)^2)^{1/4} + 0.7\sqrt{D_m v_m^2 K^2 l_m} + 0.15 v_m^2 K^2 l_m + e_m(K^2 l_m)^2}{n}$$
$$\leq (1.7K\epsilon_n + 0.15(K\epsilon_n)^2 + (K\epsilon_n)^4)\frac{R_m}{n} \leq 3K\epsilon_n \frac{R_m}{n}.$$

$$\frac{1.8D_m^{3/4}(e_m(K^2 l_m)^2)^{1/4} + 1.71\sqrt{D_m v_m^2 (K^2 l_m)} + 4.06 e_m(K^2 l_m)^2}{n}$$
$$\leq (3.51K\epsilon_n + 4.06(K\epsilon_n)^4)\frac{R_m}{n} \leq 6K\epsilon_n \frac{R_m}{n}.$$

It comes then from Proposition 2.2.1 applied with $x = K^2 l_m$ that, for all m in $\mathcal{M}_n$

$$\mathbb{P}\left(p(m) - \frac{D_m}{n} > 3K\epsilon_n \frac{R_m}{n}\right) \leq e^{-\frac{K^2}{20} l_m}.$$

Thus, from (2.34), for all $K > \sqrt{10}$, and for all n sufficiently largege,

$$\mathbb{P}\left(\forall m \in \mathcal{M}_n,\ p(m) - \frac{D_m}{n} > 3K\epsilon_n \frac{R_m}{n}\right) \leq \Sigma(K^2/10) e^{-\frac{K^2}{20}(\ln n)^\gamma}.$$

We use the same arguments to prove that

$$\mathbb{P}\left(\forall m \in \mathcal{M}_n,\ p(m) - \frac{D_m}{n} < 6K\epsilon_n \frac{R_m}{n}\right) \leq \Sigma(K^2/10) e^{-\frac{K^2}{20}(\ln n)^\gamma}.$$

Fixe $K = \sqrt{10.5}$, then for all n sufficiently large , the conclusion of Lemma 2.5.2 holds. It holds in general provided that we increase the constant C if necessary.

Lemma 2.5.3 *Let $(\psi_\lambda)_{\lambda\in\Lambda}$ be an orthonormal system in $L^2(\mu)$ and let L be a linear functional defined on $L^2(\mu)$. Let $p(\Lambda) = \sum_{\lambda\in\Lambda}(\nu_n(L(\psi_\lambda)))^2$. Let $(W_1, ..., W_n)$ be a resampling scheme, let $\bar{W}_n = \sum_{i=1}^n W_i/n$ and let $v_W^2 = Var(W_1 - \bar{W}_n)$. Let*

$$D_\Lambda^W = n(v_W^2)^{-1} \sum_{\lambda\in\Lambda} \mathbb{E}^W\left((\nu_n^W(L(\psi_\lambda)))^2\right),$$

$T = \sum_{\lambda\in\Lambda}(L(\psi_\lambda) - PL(\psi_\lambda))^2$, $D = PT$ and

$$U = \frac{1}{n(n-1)} \sum_{i\neq j=1}^n \sum_{\lambda\in\Lambda} (L(\psi_\lambda)(X_i) - PL(\psi_\lambda))(L(\psi_\lambda)(X_j) - PL(\psi_\lambda)).$$

then

$$p(\Lambda) = \frac{1}{n}P_n T + \frac{n-1}{n}U,\ D_\Lambda^W = P_n T - U,\ p(\Lambda) - \frac{D_\Lambda^W}{n} = U,$$
$$\mathbb{E}(D_\Lambda^W) = D,\ D_\Lambda^W - D = \nu_n T - U.$$

Proof :

It is easy to check that

$$\begin{aligned} p(\Lambda) &= \sum_{\lambda\in\Lambda}(\frac{1}{n}\sum_{i=1}^{n} L(\psi_\lambda)(X_i) - PL(\psi_\lambda))^2 = \frac{1}{n^2}\sum_{i=1}^{n}(L(\psi_\lambda)(X_i) - PL(\psi_\lambda))^2 \\ &+\frac{1}{n^2}\sum_{i\neq j=1}^{n}\sum_{\lambda\in\Lambda}(L(\psi_\lambda)(X_i) - PL(\psi_\lambda))(L(\psi_\lambda)(X_j) - PL(\psi_\lambda)) \\ &= \frac{1}{n}P_nT + \frac{n-1}{n}U. \end{aligned}$$

Recall that $\nu_n^W = P_n^W - \bar{W}_nP_n$. For all λ in Λ, since $\sum_{i=1}^{n}(W_i - \bar{W}_n) = 0$, we have

$$\begin{aligned} \nu_n^W(L(\psi_\lambda)) &= \frac{1}{n}\sum_{i=1}^{n}(W_i - \bar{W}_n)L(\psi_\lambda)(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}(W_i - \bar{W}_n)(L(\psi_\lambda)(X_i) - PL(\psi_\lambda)). \end{aligned}$$

Thus, if $E_{i,j} = \mathbb{E}\left((W_i - \bar{W}_n)(W_j - \bar{W}_n)\right)/v_W^2$, we have

$$\begin{aligned} D_\Lambda^W &= n(v_W^2)^{-1}\sum_{\lambda\in\Lambda}\mathbb{E}^W\left(\left(\frac{1}{n}\sum_{i=1}^{n}(W_i - \bar{W}_n)(L(\psi_\lambda)(X_i) - PL(\psi_\lambda))\right)^2\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{E}\left((W_i - \bar{W}_n)^2\right)}{v_W^2}(L(\psi_\lambda)(X_i) - PL(\psi_\lambda))^2 + \\ &\frac{1}{n}\sum_{i\neq j=1}^{n}\sum_{\lambda\in\Lambda}E_{i,j}(L(\psi_\lambda)(X_i) - PL(\psi_\lambda))(L(\psi_\lambda)(X_j) - PL(\psi_\lambda)). \end{aligned}$$

Since the weights are exchangeable, for all $i = 1,..,n$, we have $\mathbb{E}((W_i - \bar{W}_n)^2) = \mathrm{Var}(W_1 - \bar{W}_n) = v_W^2$ and for all $i \neq j = 1,...,n$,

$$v_W^2E_{i,j} = \mathbb{E}\left((W_i - \bar{W}_n)(W_j - \bar{W}_n)\right) = \mathbb{E}\left((W_1 - \bar{W}_n)(W_2 - \bar{W}_n)\right).$$

Moreover, since $\sum_{i=1}^{n}(W_i - \bar{W}_n) = 0$, we have

$$\begin{aligned} 0 &= E\left[\left(\sum_{i=1}^{n}(W_i - \bar{W}_n)\right)^2\right] = \sum_{i=1}^{n}\mathbb{E}\left((W_i - \bar{W}_n)^2\right) + \sum_{i\neq j=1}^{n}v_W^2E_{i,j} \\ &= n\mathbb{E}((W_1 - \bar{W}_n)^2) + n(n-1)\mathbb{E}\left((W_1 - \bar{W}_n)(W_2 - \bar{W}_n)\right). \end{aligned}$$

Hence, for all $i \neq j = 1,...,n$, $E_{i,j} = -1/(n-1)$, thus

$$D_\Lambda^W = P_nT - U.$$

The last inequalities of Lemma 2.5.3 follow from the fact that $\mathbb{E}(U) = 0$. Finally,

$$p(\Lambda) - \frac{D_\Lambda^W}{n} = \frac{1}{n}P_nT + \frac{n-1}{n}U - \left(\frac{1}{n}P_nT - \frac{1}{n}U\right) = U.$$

Lemma 2.5.4 *Let*

$$\begin{aligned}\Omega_u &= \bigcap_{m\in\mathcal{M}_n}\left\{\frac{D_m^W}{n}-p(m)\leq 10\epsilon_n\frac{R_m}{n}\right\}\\ \Omega_l &= \bigcap_{m\in\mathcal{M}_n}\left\{\frac{D_m^W}{n}-p(m)\geq -12\epsilon_n\frac{R_m}{n}\right\}\end{aligned}$$

and $\tilde{\Omega}_p=\Omega_u\cap\Omega_l$. *There exists a constant* $C>0$ *such that* $\mathbb{P}(\tilde{\Omega}_p^c)\leq Ce^{-\frac{1}{2}(\ln n)^\gamma}$.

Proof :

From Assumption **[V]** applied with $m=m'$, (see (2.32)), if $l_m=l_{n,\gamma}(R_m,R_m)$, we have, for all $K>0$,

$$D_m^{3/4}(e_m(K^2l_m)^2)^{1/4}\leq K\epsilon_nR_m,\ \sqrt{v_m^2D_m(K^2l_m)}\leq K\epsilon_nR_m,$$

$$v_m^2(K^2l_m)\leq(K\epsilon_n)^2R_m,\ e_m(Kl_m)^2\leq(K\epsilon_n)^4R_m.$$

We apply Proposition 2.2.4 with $x=K^2l_m$ and we have

$$\mathbb{P}\left(\frac{D_m^W}{n}-p(m)>\left(8.31K\epsilon_n+3(K\epsilon_n)^2+(19.1)^2(K\epsilon_n)^4\right)\frac{R_m}{n-1}\right)\leq 2e^{-K^2l_m}.$$

Thus, for all $K>1/(\sqrt{2})$, if $e_n(K)=n\left(8.31K\epsilon_n+3(K\epsilon_n)^2+(19.1)^2(K\epsilon_n)^4\right)/(n-1)$, we have from (2.34)

$$\mathbb{P}\left(\forall m\in\mathcal{M}_n,\ \frac{D_m^W}{n}-p(m)>e_n(K)\frac{R_m}{n}\right)\leq 2\Sigma(2K^2)e^{-K^2(\ln n)^\gamma}.$$

Take $K=8/8.31$ and $n\geq 10$ sufficiently large to ensure that $3K^2\epsilon_n+(19.1)^2K^4\epsilon_n^3\leq 1$, then we have

$$e_n(K)\leq\frac{10}{9}\left(8\epsilon_n+\epsilon_n\right)\leq 10\epsilon_n.$$

We deduce that, for sufficiently large n, we have

$$\mathbb{P}(\Omega_u^c)\leq 2\Sigma(2K^2)e^{-K^2(\ln n)^\gamma}.$$

We also apply Proposition 2.2.4 with $x=K^2l_m$, and we use the same arguments to prove that, for $K=16/16.61$, for all $n\geq 10$ sufficiently large to ensure that $(40.3)^2K^4\epsilon_n^3\leq 2$

$$\mathbb{P}\left(\forall m\in\mathcal{M}_n,\ \frac{D_m^W}{n}-p(m)<-20\epsilon_n\frac{R_m}{n}\right)\leq 3.8\Sigma(2K^2)e^{-K^2(\ln n)^\gamma}.$$

Hence, the conclusion of Lemma 2.5.4 holds for sufficiently large n. It holds in general, provided that we increase the constant C if necessary.

2.5.3 Proof of Theorem 2.2.2

If $c_n < 0$, there is nothing to prove. We can then assume that $c_n \geq 0$, this implies in particular that

$$28\epsilon_n \leq \delta_n < 1.$$

We use the notations of Lemma 2.5.2. From Lemma 2.5.2, the inequalities (2.19) will be proved if, on Ω_T, we have $D_{\hat{m}} \geq c_n D_{m^*}$ and

$$\|s - \tilde{s}\|^2 \geq \frac{c_n}{5h_n^o} \inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2.$$

Let $m_o \in \arg\min_{m \in \mathcal{M}_n} R_m$, $\hat{m}$ minimizes over $\mathcal{M}_n$ the following criterion.

$$\begin{aligned} \mathrm{Crit}(m) &= P_n Q(\hat{s}_m) + \mathrm{pen}(m) + \|s\|^2 + 2\nu_n(s_{m_o}) \\ &= \|s - s_m\|^2 - p(m) + \delta(m_o, m) + \mathrm{pen}(m). \end{aligned}$$

Recall that $0 \leq \mathrm{pen}(m) \leq (1-\delta_n)D_m/n$. On Ω_T, for all m in $\mathcal{M}_n$, since $R_m \geq R_{m_o}$, we have the following inequalities

$$\begin{aligned} \mathrm{Crit}(m) &\geq \|s - s_m\|^2 - \frac{D_m}{n} - 16\epsilon_n \frac{R_m}{n} \geq -(1+16\epsilon_n)\frac{D_m}{n}. \\ \mathrm{Crit}(m) &\leq \|s - s_m\|^2 + 26\epsilon_n \frac{R_m}{n} - \delta_n \frac{D_m}{n} = (1+26\epsilon_n)\|s - s_m\|^2 - (\delta_n - 26\epsilon_n)\frac{D_m}{n}. \end{aligned}$$

When $D_m \leq c_n D_{m^*}$, we have

$$(1+16\epsilon_n)D_m \leq D_{m^*}\left((\delta_n - 26\epsilon_n) - (1+26\epsilon_n)\frac{n\|s - s_{m^*}\|^2}{D_{m^*}}\right).$$

Thus $\mathrm{Crit}(m) \geq \mathrm{Crit}(m^*)$. This implies that $D_{\hat{m}} \geq c_n D_{m^*}$.
Moreover, on Ω_T, we also have, for all m in $\mathcal{M}_n$

$$\|s - \tilde{s}\|^2 = \frac{R_{\hat{m}}}{n} + \left(p(\hat{m}) - \frac{D_{\hat{m}}}{n}\right) \geq (1 - 20\epsilon_n)\frac{R_{\hat{m}}}{n},$$

and

$$\inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2 \leq \inf_{m \in \mathcal{M}_n} \frac{R_m}{n}(1+10\epsilon_n) \leq \frac{R_{m_o}}{n}(1+10\epsilon_n).$$

Thus

$$\begin{aligned} \|s - \tilde{s}\|^2 &\geq (1-20\epsilon_n)\frac{R_{\hat{m}}}{n} \geq (1-20\epsilon_n)\frac{D_{\hat{m}}}{n} \geq (1-20\epsilon_n)c_n\frac{D_{m^*}}{n} \\ &\geq c_n \frac{1-20\epsilon_n}{h_n^o}\frac{R_{m_o}}{n} \geq \frac{c_n}{h_n^o}\frac{1-20\epsilon_n}{1+10\epsilon_n} \inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2. \end{aligned}$$

We conclude the proof, saying that $\epsilon_n \leq 1/28$ implies that $(1-20\epsilon_n)(1+10\epsilon_n)^{-1} \geq 8/38 \geq 1/5$.

2.5.4 Proof of Theorem 2.2.3

If $\delta_- - 46\epsilon_n < -1$, there is nothing to prove, hence, we can assume in the following that $\delta_- - 46\epsilon_n > -1$.
We keep the notation Ω_T introduced in Lemma 2.5.2. Let

$$\Omega_{\text{pen}} = \bigcap_{m\in\mathcal{M}_n} \left\{ \frac{2D_m}{n} + \delta_- \frac{R_m}{n} \leq \text{pen}(m) \leq \frac{2D_m}{n} + \delta^+ \frac{R_m}{n} \right\},$$

$\Omega = \Omega_T \cap \Omega_{\text{pen}}$ and $m_o \in \arg\min_{m\in\mathcal{M}_n} R_m$. Recall that $\mathbb{P}(\Omega_{\text{pen}}) \geq 1 - p'$ and that, $\hat{m}$ minimizes over m the following criterion.

$$\begin{aligned} \text{Crit}(m) &= P_n Q(\hat{s}_m) + \text{pen}(m) + \|s\|^2 + 2\nu_n(s_{m_o}) \\ &= \|s - s_m\|^2 - p(m) + \delta(m_o, m) + \text{pen}(m). \end{aligned}$$

Therefore, on Ω, for all m in $\mathcal{M}_n$, since $R_m \geq R_{m_o}$,

$$\begin{aligned} \text{Crit}(m) &\geq (1+\delta_-)\frac{R_m}{n} + \left(\frac{D_m}{n} - p(m)\right) - 6\epsilon_n \frac{R_m}{n} \\ &\geq (1+\delta_- - 16\epsilon_n)\|s - s_m\|^2 + (1+\delta_- - 16\epsilon_n)\frac{D_m}{n} \geq (1+\delta_- - 16\epsilon_n)\frac{D_m}{n} \\ \text{Crit}(m) &\leq (1+\delta^+ + 26\epsilon_n)\frac{R_m}{n}. \end{aligned}$$

If $D_m > C_n(\delta_-, \delta^+)R_{m_o}$,

$$(1+\delta_- - 16\epsilon_n)D_m > (1+\delta^+ + 26\epsilon_n)R_{m_o},$$

Thus $\text{Crit}(m) > \text{Crit}(m_o)$, hence $D_{\hat{m}} \leq C_n(\delta_-, \delta^+)R_{m_o}$.
Moreover, from (2.6), for all m in $\mathcal{M}_n$

$$\begin{aligned} \|s - \tilde{s}\|^2 &\leq \|s - \hat{s}_m\|^2 + (\text{pen}(m) - 2p(m)) + (2p(\hat{m}) - \text{pen}(\hat{m})) + \delta(\hat{m}, m) \\ &\leq \|s - \hat{s}_m\|^2 + 2\left(\frac{D_m}{n} - p(m)\right) + (\delta^+ + 6\epsilon_n)\frac{R_m}{n} \\ &\quad +2\left(p(\hat{m}) - \frac{D_{\hat{m}}}{n}\right) + (-\delta_- + 6\epsilon_n)\frac{R_{\hat{m}}}{n} \\ &\leq \|s - \hat{s}_m\|^2 + (46\epsilon_n + \delta^+)\frac{R_m}{n} + (26\epsilon_n - \delta_-)\frac{R_{\hat{m}}}{n}. \end{aligned}$$

For all m in $\mathcal{M}_n$, we have, on Ω_T,

$$\|s - \hat{s}_m\|^2 = \frac{R_m}{n} + \left(p(m) - \frac{D_m}{n}\right) \geq (1 - 20\epsilon_n)\frac{R_m}{n}.$$

Hence, for all $m \in \mathcal{M}_n$,

$$\|s - \tilde{s}\|^2 \leq \|s - \hat{s}_m\|^2 \left(1 + \frac{46\epsilon_n + \delta^+}{1 - 20\epsilon_n}\right) + \frac{26\epsilon_n - \delta_-}{1 - 20\epsilon_n}\|s - \tilde{s}\|^2.$$

This concludes the proof of Proposition 2.2.3.

2.5.5 Proof of Proposition 2.2.4

We apply Lemma 2.5.3 with $L = id$ and $\Lambda = m$. By definition of $p(m)$ and D_m^W, we have

$$p(m) - \frac{D_m^W}{n} = \frac{1}{n(n-1)} \sum_{i \neq j = 1}^{n} \sum_{\lambda \in m} (\psi_\lambda(X_i) - P\psi_\lambda)(\psi_\lambda(X_j) - P\psi_\lambda).$$

Thus, from Lemma 2.6.7 in the appendix, we have, for all $x > 0$,

$$\mathbb{P}\left(p(m) - \frac{D_m^W}{n} > \frac{5.31 D_m^{3/4}(e_m x^2)^{1/4} + 3\sqrt{v_m^2 D_m x} + 3v_m^2 x + e_m(19.1x)^2}{n-1} \right) \leq 2e^{-x}.$$

$$\mathbb{P}\left(\frac{D_m^W}{n} - p(m) > \frac{9 D_m^{3/4}(e_m x^2)^{1/4} + 7.61\sqrt{v_m^2 D_m x} + e_m(40.3x)^2}{n-1} \right) \leq 3.8e^{-x}.$$

This proves (2.23) and (2.24).
In order to obtain (2.21) and (2.22), we introduce, for all m in $\mathcal{M}_n$, the function $T_m = \sum_{\lambda \in m} (\psi_\lambda - P\psi_\lambda)^2$ and the random variable

$$U_m = \frac{1}{n(n-1)} \sum_{i \neq j = 1}^{n} \sum_{\lambda \in m} (\psi_\lambda(X_i) - P\psi_\lambda)(\psi_\lambda(X_j) - P\psi_\lambda).$$

We apply Lemma 2.5.3 with $L = id$, we have

$$D_m^W - D_m = \nu_n(T_m) - U_m.$$

From Bernstein's inequality (see Proposition 2.6.3), we have, for all $x > 0$ and all ξ in $\{-1, 1\}$,

$$\mathbb{P}\left(\xi \nu_n(T_m) > \sqrt{\frac{2\mathrm{Var}(T_m(X))x}{n}} + \frac{\|T_m\|_\infty x}{3n} \right) \leq e^{-x}.$$

From Cauchy-Schwarz inequality, we have $T_m = \sup_{t \in B_m} (t - Pt)^2$, thus $\|T_m\|_\infty / n = 4e_m$ and $\mathrm{Var}(T_m(X))/n \leq \|T_m\|_\infty PT_m/n = 4e_m D_m$, therefore, for all $x > 0$ and all ξ in $\{-1, 1\}$,

$$\mathbb{P}\left(\xi \nu_n(T_m) > \sqrt{8 e_m D_m x} + \frac{4 e_m x}{3} \right) \leq e^{-x}.$$

Moreover, from Lemma 2.6.7 in the appendix, we have, for all $x > 0$,

$$\mathbb{P}\left(U_m > \frac{5.31 D_m^{3/4}(e_m x^2)^{1/4} + 3\sqrt{v_m^2 D_m x} + 3v_m^2 x + e_m(19.1x)^2}{n-1} \right) \leq 2e^{-x}.$$

$$\mathbb{P}\left(U_m < -\frac{9 D_m^{3/4}(e_m x^2)^{1/4} + 7.61\sqrt{v_m^2 D_m x} + e_m(40.3x)^2}{n-1} \right) \leq 3.8e^{-x}.$$

We deduce that, for all $x > 0$, with probability larger than $1 - 4.8e^{-x}$, we have

$$\begin{aligned} D_m^W - D_m &\leq \sqrt{8e_m D_m x} + e_m\left(\frac{4x}{3} + \frac{(40.3x)^2}{n-1}\right) \\ &\quad + \frac{9D_m^{3/4}(e_m x^2)^{1/4} + 7.61\sqrt{v_m^2 D_m x}}{n-1}. \end{aligned}$$

Moreover, for all $x > 0$, we have, on an event of probability larger than $1 - 3e^{-x}$,

$$\begin{aligned} D_m^W - D_m &\geq -\sqrt{8e_m D_m x} - e_m\left(\frac{4x}{3} + \frac{(19.1x)^2}{n-1}\right) \\ &\quad - \frac{5.31D_m^{3/4}(e_m x^2)^{1/4} + 3\sqrt{v_m^2 D_m x} + 3v_m^2 x}{n-1}. \end{aligned}$$

2.5.6 Proof of Theorem 2.2.5

Recall that $\mathbb{P}\left(\Omega_T^c\right) \leq Ce^{-\frac{1}{2}(\ln n)^\gamma}$, and that, on Ω_T, we have

$$\forall m \in \mathcal{M}_n, (1 - 20\epsilon_n)\frac{R_m}{n} \leq \|s - \hat{s}_m\|^2,$$

$$\forall m, m' \in \mathcal{M}_n^2,\ \delta(m, m') \leq 6\epsilon_n \frac{R_m \vee R_{m'}}{n}.$$

Let $\tilde{\Omega}_p$ be the event defined in Lemma 2.5.4 and let $\Omega = \tilde{\Omega}_p \cap \Omega_T$, from Lemma 2.5.2, $\mathbb{P}\left(\Omega^c\right) \leq Ce^{-\frac{1}{2}(\ln n)^\gamma}$. Recall that $\mathrm{pen}(m) = 2D_m^W/n$. On Ω, from (2.6), for all n such that $20\epsilon_n < 1$, for all m in $\mathcal{M}_n$,

$$\begin{aligned} \|s - \tilde{s}\|^2 &\leq \|s - \hat{s}_m\|^2 + 26\epsilon_n \frac{R_m}{n} + 16\epsilon_n \frac{R_{\hat{m}}}{n} \\ &\leq \|s - \hat{s}_m\|^2 + \frac{26\epsilon_n}{1 - 20\epsilon_n}\|s - \hat{s}_m\|^2 + \frac{16\epsilon_n}{1 - 20\epsilon_n}\|s - \tilde{s}\|^2. \end{aligned}$$

Hence, for all n such that $20\epsilon_n < 1$, on Ω,

$$(1 - 36\epsilon_n)\|s - \tilde{s}\|^2 \leq (1 + 6\epsilon_n) \inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2.$$

For all n such that $42/(1 - 36\epsilon_n) < 100$,

$$\|s - \tilde{s}\|^2 \leq \left(1 + \frac{42\epsilon_n}{1 - 36\epsilon_n}\right) \inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2 \leq (1 + 100\epsilon_n) \inf_{m \in \mathcal{M}_n} \|s - \hat{s}_m\|^2.$$

Hence (2.25) holds for sufficiently large n, it holds in general provided that we enlarge the constant C if necessary..

2.6 Appendix

In this Section, we state and prove some technical lemmas that are useful in the proofs. The main tool is the first Lemma based on Bousquet's version of Talagrand's inequality. It is a concentration inequality for the square of the supremum of the empirical process over a uniformly bounded class of functions. Recall first Bousquet's and Klein versions of Talagrand's inequality.

Theorem 2.6.1 *(Bousquet's bound) Let $X_1, ..., X_n$ be i.i.d. random variables valued in a measurable space $(\mathbb{X}, \mathcal{X})$ and let S be a class of real valued functions bounded by b. Let $v^2 = \sup_{t\in S} Var(t(X))$ and let $Z = \sup_{t\in S} \nu_n t$. Then*

$$\forall x > 0,\ \mathbb{P}\left(Z > \mathbb{E}(Z) + \sqrt{\frac{2}{n}(v^2 + 2b\mathbb{E}(Z))x} + \frac{bx}{3n}\right) \leq e^{-x}.$$

Theorem 2.6.2 *(Klein's bound) Let $X_1, ..., X_n$ be i.i.d. random variables valued in a measurable space $(\mathbb{X}, \mathcal{X})$ and let S be a class of real valued functions bounded by b. Let $v^2 = \sup_{t\in S} Var(t(X))$ and let $Z = \sup_{t\in S} \nu_n t$. Then*

$$\forall x > 0,\ \mathbb{P}\left(Z < \mathbb{E}(Z) - \sqrt{\frac{2}{n}(v^2 + 2b\mathbb{E}(Z))x} - \frac{8bx}{3n}\right) \leq e^{-x}.$$

Let us now also recall Bernstein's inequality.

Proposition 2.6.3 *Bernstein's inequality*
Let $X_1, ..., X_n$ be iid random variables valued in a measurable space $(X, \mathcal{X})$ and let t be a measurable real valued function. Then, for all $x > 0$, we have

$$\mathbb{P}\left(\nu_n(t) > \sqrt{\frac{2\,Var(t(X_1))x}{n}} + \frac{\|t\|_\infty x}{3n}\right) \leq e^{-x}.$$

We derive from these bounds the following useful corollary. Hereafter, S denotes a symetric class of real valued functions upper bounded by b, $v^2 = \sup_{t\in S} \mathrm{Var}(t(X))$, $Z = \sup_{t\in S} \nu_n t$, $n\mathbb{E}(Z^2) = D$. Since S is symetric, we always have $Z \geq 0$.

Corollary 2.6.4 *Let S be a symetric class of real valued functions upper bounded by b, $v^2 = \sup_{t\in S} Var(t(X))$, $Z = \sup_{t\in S} \nu_n t$, $n\mathbb{E}(Z^2) = D$, $e_b = b^2/n$ and*

$$nE_m = 225e_b + \left(2.1 + \sqrt{2\pi}\right)\sqrt{v^2 D} + \sqrt{15}D^{3/4}e_b^{1/4},$$

then we have

$$\mathbb{E}(Z^2\mathbf{1}_{Z\geq\mathbb{E}(Z)}) \leq (\mathbb{E}(Z))^2\mathbb{P}\left(Z \geq \mathbb{E}(Z)\right) + E_m. \tag{2.38}$$

In particular, we have

$$(\mathbb{E}(Z))^2 \leq \mathbb{E}(Z^2) \leq (\mathbb{E}(Z))^2 + E_m. \tag{2.39}$$

Proof :
We have

$$\begin{aligned}\mathbb{E}(Z^2\mathbf{1}_{Z\geq\mathbb{E}(Z)}) &= \int_0^\infty \mathbb{P}(Z^2\mathbf{1}_{Z\geq\mathbb{E}(Z)} > x)dx = \int_0^\infty \mathbb{P}(Z\mathbf{1}_{Z\geq\mathbb{E}(Z)} > \sqrt{x})dx \\ &= (\mathbb{E}(Z))^2\mathbb{P}\left(Z \geq \mathbb{E}(Z)\right) + \int_{(\mathbb{E}(Z))^2}^\infty \mathbb{P}(Z > \sqrt{x})dx\end{aligned}$$

Take $x = (\mathbb{E}(Z) + \sqrt{2(v^2 + 2b\mathbb{E}(Z))y/n} + by/(3n))^2$ in the previous integral, from Bousquet's version of Talagrand's inequality, we have

$$\begin{aligned}\mathbb{E}(Z^2\mathbf{1}_{Z\geq\mathbb{E}(Z)}) &\leq \mathbb{E}(Z)\sqrt{\frac{2}{n}(v^2 + 2b\mathbb{E}(Z))}\int_0^\infty \frac{e^{-y}}{\sqrt{y}}dy + \frac{2v^2 + 14b\mathbb{E}(Z)/3}{n}\int_0^\infty e^{-y}dy \\ &+\frac{b}{n}\sqrt{\frac{2}{n}(v^2 + 2b\mathbb{E}(Z))}\int_0^\infty e^{-y}\sqrt{y}dy + \frac{2b^2}{9n^2}\int_0^\infty ye^{-y}dy.\end{aligned}$$

Classical computations lead to

$$\int_0^\infty \frac{e^{-y}}{\sqrt{y}}dy = 2\int_0^\infty e^{-y}\sqrt{y}dy = \sqrt{\pi}, \int_0^\infty e^{-y}dy = \int_0^\infty ye^{-y}dy = 1.$$

Therefore, if $e_b = b^2/n$, using repeatedly the inequalities

$$a^\alpha b^{1-\alpha} \le \alpha a + (1-\alpha)b \tag{2.40}$$

and $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, we obtain, for all $\eta > 0$,

$$\sqrt{ne_b}\mathbb{E}(Z) \le \frac{e_b}{3\eta^2} + \frac{2\eta}{3}e_b^{1/4}(\sqrt{n}\mathbb{E}(Z))^{3/2},$$

$$(\sqrt{n}\mathbb{E}(Z))^{1/2}e_b^{3/4} \le \frac{\eta}{3}e_b^{1/4}(\sqrt{n}\mathbb{E}(Z))^{3/2} + \frac{2e_b}{3\sqrt{\eta}}.$$

Thus

$$\begin{aligned}\mathbb{E}(Z^2\mathbf{1}_{Z\ge\mathbb{E}(Z)}) &\le \left(2v^2 + \frac{2}{9}e_b + v\frac{\sqrt{2\pi e_b}}{2}\right)\frac{1}{n} + \sqrt{\pi}\frac{\sqrt{\sqrt{n}\mathbb{E}(Z)}\,(e_b)^{3/4}}{n} \\ &\quad + \left(\frac{14}{3}\sqrt{e_b} + v\sqrt{2\pi}\right)\frac{\sqrt{n}\mathbb{E}(Z)}{n} + 2\sqrt{\pi}\frac{(\sqrt{n}\mathbb{E}(Z))^{3/2}\,(e_b)^{1/4}}{n} \\ &\le \left(2 + \eta\frac{\sqrt{2\pi}}{4}\right)\frac{v^2}{n} + \sqrt{\frac{2\pi}{n}}v\mathbb{E}(Z) + \left(\frac{2}{9} + \frac{\sqrt{2\pi}}{4\eta} + \frac{2\sqrt{\pi}}{3\sqrt{\eta}} + \frac{14}{9\eta^2}\right)\frac{e_b}{n} \\ &\quad + \left(\eta\left(\frac{\sqrt{\pi}}{3} + \frac{28}{9}\right) + 2\sqrt{\pi}\right)\frac{(\sqrt{n}\mathbb{E}(Z))^{3/2}\,(e_b)^{1/4}}{n}.\end{aligned}$$

Therefore, taking $\eta = 0.088$, we obtain

$$\mathbb{E}(Z^2\mathbf{1}_{Z\ge\mathbb{E}(Z)}) \le 2.1\frac{v^2}{n} + 15^2\frac{e_b}{n} + \sqrt{2\pi}v\frac{\sqrt{n}\mathbb{E}(Z)}{n} + \sqrt{15}\frac{(\sqrt{n}\mathbb{E}(Z))^{3/2}\,(e_b)^{1/4}}{n}.$$

Finally, we use Cauchy-Schwarz inequality to obtain that $\sqrt{n}\mathbb{E}(Z) \le (n\mathbb{E}(Z^2))^{1/2} = (D)^{1/2}$. Since $v^2 \le D$, we get (2.38).

We deduce from this result the following concentration inequalities for Z^2

Corollary 2.6.5 *Let $e_b = b^2/n$. We have, for all $x > 0$,*

$$\mathbb{P}\left(Z^2 - \frac{D}{n} > \frac{D^{3/4}(e_b(19x)^2)^{1/4} + 3\sqrt{Dv^2x}) + 3v^2x + e_b(19x)^2}{n}\right) \le e^{-x}.$$

Moreover, for all $x > 0$, we have, with probability larger than $1 - e^{-x}$,

$$\frac{D}{n} - Z^2 \le \frac{D^{3/4}e_b^{1/4}(\sqrt{15} + 4.127\sqrt{x}) + \sqrt{v^2D}(4.61 + 3\sqrt{x}) + 225e_b(6.2x^2+1)}{n}. \tag{2.41}$$

Proof :

From Bousquet's version of Talagrand's inequality and from $(\mathbb{E}(Z))^2 \leq \mathbb{E}(Z^2)$, we obtain that, for all $x > 0$, we have, with probability larger than $1-e^{-x}$, Z^2-D/n is not larger than

$$\frac{4D^{3/4}(e_b x^2)^{1/4} + \sqrt{D}(14\sqrt{e_b x^2}/3 + 2\sqrt{2v^2 x}) + 4D^{1/4}(e_b x^2)^{3/4}/3 + 3v^2 x + e_b x^2/3}{n}.$$

We use repeatedly the inequality $a^\alpha b^{1-\alpha} \leq \alpha a + (1-\alpha)b$ to obtain that, with probability at least $1-e^{-x}$, $Z^2 - D/n$ is not larger than

$$\frac{(4+32\eta/9)D^{3/4}(e_b x^2)^{1/4} + 2\sqrt{2}\sqrt{Dv^2 x} + 3v^2 x + (3 + 14/\eta^2 + 8/\sqrt{\eta})e_b x^2/9}{n}.$$

For $\eta = 0.07$, this gives

$$Z^2 - \frac{D}{n} > \frac{D^{3/4}(e_b(19x)^2)^{1/4} + 2\sqrt{2}\sqrt{Dv^2 x} + 3v^2 x + e_b(19x)^2}{n}.$$

For the second one we use Klein's version of Talagrand's inequality to obtain, for all $x > 0$ such that $r(x) = \sqrt{2(v^2 + 2b\mathbb{E}(Z))x/n} + 8bx/3n < \mathbb{E}(Z)$,

$$\mathbb{P}\left(Z^2 < (\mathbb{E}(Z) - r(x))^2\right) \leq e^{-x}.$$

We have $(\mathbb{E}(Z) - r(x))^2 = (\mathbb{E}(Z))^2 - 2\mathbb{E}(Z)r(x) + r(x)^2 \geq (\mathbb{E}(Z))^2 - 2\mathbb{E}(Z)r(x)$, thus

$$\mathbb{P}\left(Z^2 < (\mathbb{E}(Z))^2 - 2\mathbb{E}(Z)r(x)\right) \leq e^{-x}.$$

From the previous corollary, we have $(\mathbb{E}(Z))^2 \geq \mathbb{E}(Z^2) - E_m$, thus

$$\mathbb{P}\left(Z^2 < \mathbb{E}(Z^2) - E_m - 2\mathbb{E}(Z)r(x)\right) \leq e^{-x}.$$

In order to conclude the proof of 2.6.5, just remark that

$$\begin{aligned} 2\mathbb{E}(Z)r(x) &\leq \frac{4D^{3/4}(e_b x^2)^{1/4} + 3\sqrt{Dv^2 x} + 16\sqrt{De_b x^2}/3}{n} \\ &\leq \frac{(4+32\eta/9)D^{3/4}(e_b x^2)^{1/4} + 3\sqrt{Dv^2 x} + 16/(9\eta^2)e_b x^2}{n}. \end{aligned}$$

For $\eta = 0,0357$, we obtain (2.41).

Finally, we have obtained the following result for the concentration of Z^2 around its mean

Corollary 2.6.6 *For all $x > 0$, we have*

$$\mathbb{P}\left(Z^2 - \frac{D}{n} > \frac{D^{3/4}(e_b(19x)^2)^{1/4} + 3\sqrt{Dv^2 x} + 3v^2 x + e_b(19x)^2}{n}\right) \leq e^{-x}.$$

$$\mathbb{P}\left(Z^2 - \frac{D}{n} < -\frac{8D^{3/4}(e_b x^2)^{1/4} + 7.61\sqrt{v^2 Dx} + e_b(40.25x)^2}{n}\right) \leq ee^{-x}.$$

Proof :

In order to obtain the second inequality, we remark that the inequality is trivial when $x \leq 1$, thus we only have to use (2.41) for $x > 1$ and then we have $\sqrt{x} > 1$ and $x^2 > 1$.

We will use this lemma to obtain a concentration inequality for totally degenerate U-statistics of order 2. The following result generalizes a previous inequality due to Houdré & Reynaud-Bouret [39] to random variables taking values in a measurable space.

Lemma 2.6.7 *Let $X, X_1, ..., X_n$ be i.i.d random variables taking value in a measurable space $(\mathbb{X}, \mathcal{X})$ with common law P. Let μ be a measure on $(\mathbb{X}, \mathcal{X})$ and let $(t_\lambda)_{\lambda\in\Lambda}$ be a set of functions in $L^2(\mu)$. Let*

$$B = \{t = \sum_{\lambda\in\Lambda} a_\lambda t_\lambda,\ \sum_{\lambda\in\Lambda} a_\lambda^2 \leq 1\},\ D = \mathbb{E}\left(\sup_{t\in B}(t(X) - Pt)^2\right),$$

$$v^2 = \sup_{t\in B} Var(t(X)),\ b = \sup_{t\in B} \|t\|_\infty \text{ and } e_b = \frac{b^2}{n}.$$

Let

$$U = \frac{1}{n(n-1)} \sum_{i\neq j=1}^{n} \sum_{\lambda\in\Lambda} (t_\lambda(X_i) - Pt_\lambda)(t_\lambda(X_j) - Pt_\lambda).$$

Then the following inequality holds

$$\forall x > 0,\ \mathbb{P}\left(U > \frac{5.31 D^{3/4}(e_b x^2)^{1/4} + 3\sqrt{v^2 Dx} + 3v^2 x + e_b(19.1x)^2}{n-1}\right) \leq 2e^{-x}. \tag{2.42}$$

$$\forall x > 0,\ \mathbb{P}\left(U < -\frac{9D^{3/4}(e_b x^2)^{1/4} + 7.61\sqrt{v^2 Dx} + e_b(40.3x)^2}{n-1}\right) \leq 3.8e^{-x}. \tag{2.43}$$

Proof :

Remark that, from Cauchy-Schwarz inequality, we have

$$\sup_{t\in B}(\nu_n(t))^2 = \left(\sup_{\sum a_\lambda^2 \leq 1} \sum_{\lambda\in\Lambda} a_\lambda \nu_n(t_\lambda)\right)^2 = \sum_{\lambda\in\Lambda}(\nu_n(t_\lambda))^2.$$

For all x in $\mathbb{X}$, from Cauchy-Schwarz inequality, we have

$$\sup_{t\in B}(t(x) - Pt)^2 = \sum_{\lambda}(t_\lambda(x) - Pt_\lambda)^2,$$

in particular, we have $D = \sum_{\lambda\in\Lambda} \mathrm{Var}(\psi_\lambda(X))$. Moreover, easy algebra leads to

$$\begin{aligned}
\sum_{\lambda\in\Lambda}(\nu_n(t_\lambda))^2 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\lambda\in\Lambda}(t_\lambda(X_i) - Pt_\lambda)^2 \\
&\quad + \frac{1}{n^2}\sum_{i\neq j=1}^{n}\sum_{\lambda\in\Lambda}(t_\lambda(X_i) - Pt_\lambda)(t_\lambda(X_j) - Pt_\lambda) \\
&= \frac{1}{n}P_n\left(\sum_{\lambda\in\Lambda}(t_\lambda - Pt_\lambda)^2\right) + \frac{n-1}{n}U.
\end{aligned}$$

Let $Z^2 = \sup_{t\in B}(\nu_n(t))^2$, $T_\Lambda = \sum_{\lambda\in\Lambda}(t_\lambda - Pt_\lambda)^2$, we have

$$\mathbb{E}(Z^2) = \mathbb{E}\left(\frac{1}{n}P_n T_\Lambda\right) = \frac{D}{n}.$$

Hence

$$U = \frac{n}{n-1}\left(Z^2 - \mathbb{E}(Z^2) - \frac{1}{n}\nu_n(T_\Lambda)\right).$$

From Corollary 2.6.6, we have, for all $x > 0$,

$$\mathbb{P}\left(Z^2 - \frac{D}{n} > \frac{D^{3/4}(e_b(19x)^2)^{1/4} + 3\sqrt{v^2Dx} + 3v^2x + e_b(19x)^2}{n}\right) \leq e^{-x}.$$

$$\mathbb{P}\left(Z^2 - \frac{D}{n} < -\frac{8D^{3/4}(e_b(x)^2)^{1/4} + 7.61\sqrt{v^2Dx} + e_b(40.25x)^2}{n}\right) \leq 2.8e^{-x}.$$

Moreover, from Bernstein inequality, we have, for all $x > 0$,

$$\mathbb{P}\left(-\nu_n T_\Lambda > \sqrt{2De_bx} + \frac{e_bx}{3}\right) \leq e^{-x}.$$

$$\mathbb{P}\left(\nu_n T_\Lambda > \sqrt{2De_bx} + \frac{e_bx}{3}\right) \leq e^{-x}.$$

We apply inequality (2.40) with $a = D^{3/4}(e_bx^2)^{1/4}$, $b = e_b\sqrt{x}$, $\alpha = 2/3$ and we obtain

$$\mathbb{P}\left(-\nu_n T_\Lambda > \frac{2\sqrt{2}}{3}D^{3/4}(e_bx^2)^{1/4} + e_b\left(\frac{x+\sqrt{2x}}{3}\right)\right) \leq e^{-x}.$$

$$\mathbb{P}\left(\nu_n T_\Lambda > \frac{2\sqrt{2}}{3}D^{3/4}(e_bx^2)^{1/4} + e_b\left(\frac{x+\sqrt{2x}}{3}\right)\right) \leq e^{-x}.$$

Therefore, for all $x > 0$, we have

$$\mathbb{P}\left(U > \frac{5.31D^{3/4}(e_bx^2)^{1/4} + 3\sqrt{v^2Dx} + 3v^2x + e_b\left((19x)^2 + (x+\sqrt{2x})/3\right)}{n-1}\right) \leq 2e^{-x}.$$

$$\mathbb{P}\left(U < -\frac{9D^{3/4}(e_bx^2)^{1/4} + 7.61\sqrt{v^2Dx} + e_b\left((40.25x)^2 + (x+\sqrt{2x})/3\right)}{n-1}\right) \leq 3.8e^{-x}.$$

These inequalities are trivial when $x < 1$. We only use them when $x > 1$ and we obtain (2.42) and (2.43) since $x < x^2$ and $\sqrt{x} < x^2$ when $x > 1$.

Let us now state the corollary of Bernstein's inequality that we used repeatedly in the article.

Lemma 2.6.8 *Let $X, X_1, ..., X_n$ be i.i.d random variables taking value in a measurable space $(\mathbb{X}, \mathcal{X})$ with common law P. Let μ be a measure on $(\mathbb{X}, \mathcal{X})$ and let $(\psi_\lambda)_{\lambda\in\Lambda}$ be an orthonormal system in $L^2(\mu)$. Let L be a linear functional in $L^2(\mu)$ and let $B = \{t = \sum_{\lambda\in\Lambda} a_\lambda L(\psi_\lambda),\ \sum_{\lambda\in\Lambda} a_\lambda^2 \leq 1\}$, $v^2 = \sup_{t\in B} Var(t(X))$, $b = \sup_{t\in B}\|t\|_\infty$ and $e_b = b^2/n$. Let u be a function in S, the linear space spanned by the functions $(\psi_\lambda)_{\lambda\in\Lambda}$ and let $\eta > 0$. Then the following inequality holds*

$$\forall x > 0,\ \mathbb{P}\left(\nu_n(L(u)) > \frac{\eta}{2}\|u\|^2 + \frac{2v^2x + e_bx^2/9}{\eta n}\right) \leq e^{-x}. \qquad (2.44)$$

Proof :

From Bernstein's inequality, we have

$$\forall x > 0,\ \mathbb{P}\left(\nu_n(L(u)) > \sqrt{\frac{2\mathrm{Var}(L(u)(X))x}{n}} + \frac{\|L(u)\|_\infty\, x}{3n}\right) \leq e^{-x}.$$

Since $t = L(u/\|u\|)$ belongs to B, we have

$$\begin{aligned}\sqrt{\frac{2\mathrm{Var}(L(u)(X))x}{n}} + \frac{\|L(u)\|_\infty\, x}{3n} &= \|u\|\left(\sqrt{\frac{2\mathrm{Var}(t(X))x}{n}} + \frac{\|t\|_\infty\, x}{3n}\right)\\ &\leq \frac{\eta}{2}\|u\|^2 + \frac{1}{2\eta}\left(\sqrt{\frac{2v^2x}{n}} + \frac{bx}{3n}\right)^2.\end{aligned}$$

We conclude the proof using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$.

Chapter 3

Adaptive density estimation of stationary β-mixing and τ-mixing processes

Abstract:

We propose an algorithm to estimate the common density s of a stationary process $X_1, ..., X_n$. We suppose that the process is either β or τ-mixing. We provide a model selection procedure based on a generalization of Mallows' C_p and we prove oracle inequalities for the selected estimator under a few prior assumptions on the collection of models and on the mixing coefficients. We prove that our estimator is adaptive over a class of Besov spaces, namely, we prove that it achieves the same rates of convergence as in the i.i.d framework.

Key words: Density estimation, weak dependence, model selection.
2000 Mathematics Subject Classification: 62G07, 62M99.

3.1 Introduction

We consider the problem of estimating the unknown density s of P, the law of a random variable X, based on the observation of n (possibly) dependent data $X_1, ..., X_n$ with common law P. We assume that X is real valued, that s belongs to $L^2(\mu)$ where μ denotes the Lebesgue measure on $\mathbb{R}$ and that s is compactly supported, say in $[0, 1]$. Throughout the chapter, we consider least-squares estimators $\hat{s}_m$ of s on a collection $(S_m)_{m\in\mathcal{M}_n}$ of linear subspaces of $L^2(\mu)$. Our final estimator is chosen through a model selection algorithm.

Model selection has received much interest in the last decades. When its final goal is prediction, it can be seen more generally as the question of choosing between the outcomes of several prediction algorithms. With such a general formulation, a very natural answer is the following. First, estimate the prediction error for each model, that is $\|s - \hat{s}_m\|_2^2$. Then, select the model which minimizes this estimate.

It is natural to think of the empirical risk as an estimator of the prediction error. This can fail dramatically, because it uses the same data for building predictors and for comparing them, making these estimates strongly biased for models involving a number of parameters growing with the sample size.

In order to correct this drawback, penalization's methods state that a good choice can be made by minimizing the sum of the empirical risk (how do algorithms fit the data) and some complexity measure of the algorithms (called the penalty). This method was first developped in the work of Akaike [1] and [2] and Mallows [53].

In the context of density estimation, with independent data, Birgé & Massart [15] used penalties of order $L_n D_m/n$, where D_m denotes the dimension of S_m and L_n is a constant depending on the complexity of the collection $\mathcal{M}_n$. They used Talagrand's inequality (see for example Talagrand [65] for an overview) to prove that this penalization procedure is efficient *i.e.* the integrated quadratic risk of the selected estimator is asymptotically equivalent to the risk of the oracle (see Section 2 for a precise definition). They also proved that the selected estimator achieves adaptive rates of convergence over a large class of Besov spaces. Moreover, they showed that some methods of adaptive density estimation like the unbiased cross validation (Rudemo [62]) or the hard thresholded estimator of Donoho *et al.* [27] can be viewed as special instances of penalized projection estimators.

More recently, Arlot [5] introduced new measures of the quality of penalized least-squares estimators (PLSE). He proved pathwise oracle inequalities, that is deviation bounds for the PLSE that are harder to prove but more informative from a practical point of view (see also Section 2 for details).

When the process $(X_i)_{i=1,\dots,n}$ is β-mixing (Rozanov & Volkonskii [71] and Section 2), Talagrand's inequality can not be used directly. Baraud *et al.* [9] used Berbee's coupling lemma (see Berbee ([13]) and Viennet's covariance inequality (Viennet [70]) to overcome this problem and build model selection procedure in the regression problem. Then Comte & Merlevède [23] used this algorithm to investigate the problem of density estimation for a β-mixing process. They proved that under reasonable assumptions on the collection $\mathcal{M}_n$ and on the coefficients β, one can recover the results of Birgé & Massart [15] in the i.i.d. framework.

The main drawback of those results is that many processes, even simple Markov chains are not β-mixing. For instance, if $(\epsilon_i)_{i\geq 1}$ is iid with marginal $\mathcal{B}(1/2)$, then the stationary solution $(X_i)_{i\geq 0}$ of the equation

$$X_n = \frac{1}{2}(X_{n-1} + \epsilon_n),\ X_0 \text{ independent of } (\epsilon_i)_{i\geq 1} \tag{3.1}$$

is not β-mixing (Andrews [3]). More recently, Dedecker & Prieur [25] introduced new mixing-coefficients, in particular the coefficients τ, $\tilde{\phi}$ and $\tilde{\beta}$ and proved that many processes like (3.1) happen to be τ, $\tilde{\phi}$ and $\tilde{\beta}$-mixing. They proved a coupling lemma for the coefficient τ and covariance inequalities for $\tilde{\phi}$ and $\tilde{\beta}$. Gannaz & Wintenberger [34] used the covariance inequality to extend the result of Donoho *et al.* [27] for the wavelet thresholded estimator to the case of $\tilde{\phi}$-mixing processes. They recovered (up to a $\log(n)$ factor) the adaptive rates of convergence over Besov spaces.

In this article, we first investigate the case of β-mixing processes. We prove a pathwise oracle inequality for the PLSE. We extend the result of Comte & Merlevède [23] under weaker assumptions on the mixing coefficients. Then, we consider τ-mixing processes. The problem is that the coupling result is weaker for the coefficient τ than for β. Moreover, in order to control the empirical process we use a covariance inequality that is harder to handle. Hence, the generalization of the procedure of Baraud *et al.* [9] to the framework of τ-mixing processes is not straightforward.

We recover the optimal adaptive rates of convergence over Besov spaces (that is the same as in the independent framework) for τ-mixing processes, which is new as far as we know.

The chapter is organized as follows. In Section 2, we give the basic material that we will use throughout the chapter. We recall the definition of some mixing coefficients and we state their properties. We define the penalized least-squares estimator (PLSE). Sections 3 and 4 are devoted to the statement of the main results, respectively in the β-mixing case and in the τ-mixing case. In Section 5, we derive the adaptive properties of the PLSE. Finally, Section 6 is devoted to the proofs. Some additional material has been reported in the Appendix in Section 7.

3.2 Preliminaries

3.2.1 Notation.

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. Let μ be the Lebesgue measure on $\mathbb{R}$, let $\|.\|_p$ be the usual norm on $L^p(\mu)$ for $1 \leq p \leq \infty$. For all $y \in \mathbb{R}^l$, let $|y|_l = \sum_{i=1}^l |y_i|$. Denote by λ_κ the set of κ-Lipschitz functions, *i.e.* the functions t from $(\mathbb{R}^l, |.|_l)$ to $\mathbb{R}$ such that $\text{Lip}(t) \leq \kappa$ where

$$\text{Lip}(t) = \sup\left\{\frac{|t(x)-t(y)|}{|x-y|_l}, x, y \in \mathbb{R}^l, x \neq y\right\} \leq \kappa.$$

Let BV and BV_1 be the set of functions t supported on $\mathbb{R}$ satisfying respectively $\|t\|_{BV} < \infty$ and $\|t\|_{BV} \leq 1$ where

$$\|t\|_{BV} = \sup_{n \in \mathbb{N}^*} \sup_{-\infty < a_1 < \ldots < a_n < \infty} |t(a_{i+1}) - t(a_i)|.$$

3.2.2 Some measures of dependence.

Definitions and assumptions

Let $Y = (Y_1, ..., Y_l)$ be a random variable defined on $(\Omega, \mathcal{A}, \mathbb{P})$ with values in $(\mathbb{R}^l, |.|_l)$. Let $\mathcal{M}$ be a σ-algebra of $\mathcal{A}$. Let $\mathbb{P}_{Y|\mathcal{M}}$, $\mathbb{P}_{Y_1|\mathcal{M}}$ be conditional distributions of Y and Y_1 given $\mathcal{M}$, let $\mathbb{P}_Y$, $\mathbb{P}_{Y_1}$ be the distribution of Y and Y_1 and let $F_{Y_1|\mathcal{M}}$, F_{Y_1} be distribution functions of $\mathbb{P}_{Y_1|\mathcal{M}}$ and P_{Y_1}. Let $\mathcal{B}$ be the Borel σ-algebra on $(\mathbb{R}^l, |.|_l)$. Define now

$$\begin{aligned} \beta(\mathcal{M}, \sigma(Y)) &= \mathbb{E}\left(\sup_{A \in \mathcal{B}} |\mathbb{P}_{Y|\mathcal{M}}(A) - \mathbb{P}_Y(A)|\right), \\ \tilde{\beta}(\mathcal{M}, Y_1) &= \mathbb{E}\left(\sup_{x \in \mathbb{R}} \left|F_{Y_1|\mathcal{M}}(x) - F_{Y_1}(x)\right|\right), \\ \text{and if } \mathbb{E}(|Y|) < \infty,\ \tau(\mathcal{M}, Y) &= \mathbb{E}\left(\sup_{t \in \lambda_1} |\mathbb{P}_{Y|\mathcal{M}}(t) - \mathbb{P}_Y(t)|\right). \end{aligned}$$

The coefficient $\beta(\mathcal{M}, \sigma(Y))$ is the mixing coefficient introduced by Rozanov & Volkonskii [71]. The coefficients $\tilde{\beta}(\mathcal{M}, Y_1)$ and $\tau(\mathcal{M}, Y)$ have been introduced by Dedecker & Prieur [25].

Let $(X_k)_{k\in\mathbb{Z}}$ be a stationary sequence of real valued random variables defined on $(\Omega,\mathcal{A},\mathbb{P})$. For all $k\in\mathbb{N}^*$, the coefficients β_k, $\tilde{\beta}_k$ and τ_k are defined by

$$\beta_k=\beta(\sigma(X_i,i\leq 0),\sigma(X_i,i\geq k)),\ \tilde{\beta}_k=\sup_{j\geq k}\{\tilde{\beta}(\sigma(X_p,p\leq 0),X_j)\}.$$

If $\mathbb{E}(|X_1|)<\infty$, for all $k\in\mathbb{N}^*$ and all $r\in\mathbb{N}^*$, let

$$\tau_{k,r}=\max_{1\leq l\leq r}\frac{1}{l}\sup_{k\leq i_1<..<i_l}\{\tau(\sigma(X_p,p\leq 0),(X_{i_1},...,X_{i_l}))\},\ \tau_k=\sup_{r\in\mathbb{N}^*}\tau_{k,r}.$$

Moreover, we set $\beta_0=1$. In the sequel, the processes of interest are either β-mixing or τ-mixing, meaning that, for $\gamma=\beta$ or τ, the γ-mixing coefficients $\gamma_k\to 0$ as $k\to+\infty$. For $p\in\{1,2\}$, we define κ_p as:

$$\kappa_p=p\sum_{l=0}^{\infty}l^{p-1}\beta_l, \tag{3.2}$$

where $0^0=1$, when the series are convergent. Besides, we consider two kinds of rates of convergence to 0 of the mixing coefficients, that is for $\gamma=\beta$ or τ,
[**AR**] arithmetical γ-mixing with rate θ if there exists some $\theta>0$ such that $\gamma_k\leq(1+k)^{-(1+\theta)}$ for all k in $\mathbb{N}$,
[**GEO**] geometrical γ-mixing with rate θ if there exists some $\theta>0$ such that $\gamma_k\leq e^{-\theta k}$ for all k in $\mathbb{N}$.

Properties

Coupling
Let X be an $\mathbb{R}^l$-valued random variable defined on $(\Omega,\mathcal{A},\mathbb{P})$ and let $\mathcal{M}$ be a σ-algebra. Assume that there exists a random variable U uniformly distributed on $[0,1]$ and independent of $\mathcal{M}\vee\sigma(X)$. There exist two $\mathcal{M}\vee\sigma(X)\vee\sigma(U)$-measurable random variables X_1^* and X_2^* distributed as X and independent of $\mathcal{M}$ such that

$$\beta(\mathcal{M},\sigma(X))=\mathbb{P}(X\neq X_1^*)\text{ and} \tag{3.3}$$

$$\tau(\mathcal{M},X)=\mathbb{E}\left(|X-X_2^*|_l\right). \tag{3.4}$$

Equality (3.3) has been established by Berbee [13], Equality (3.4) has been established in Dedecker & Prieur [25], Section 7.1.
Covariance inequalities
Let X,Y be two real valued random variables and let f,h be two measurable functions from $\mathbb{R}$ to $\mathbb{C}$. Then, there exist two measurable functions $b_1:\mathbb{R}\to\mathbb{R}$ and $b_2:\mathbb{R}\to\mathbb{R}$ with $\mathbb{E}\left(b_1(X)\right)=\mathbb{E}(b_2(Y))=\beta(\sigma(X),\sigma(Y))$ such that, for any conjugate $p,q\geq 1$ (see Viennet [70] Lemma 4.1)

$$|\mathrm{Cov}(f(X),h(Y))|\leq 2\mathbb{E}^{1/p}\left(|f(X)|^p b_1(X)\right)\mathbb{E}^{1/q}(|h(Y)|^q b_2(Y)).$$

There exists a random variable $b(\sigma(X),Y)$ such that $\mathbb{E}(b(\sigma(X),Y))=\tilde{\beta}(\sigma(X),Y)$ and such that, for all Lipschitz functions f and all h in BV (Dedecker & Prieur [25] Proposition 1)

$$|\mathrm{Cov}(f(X),h(Y))|\leq\|h\|_{BV}\,\mathbb{E}\left(|f(X)|b(\sigma(X),Y)\right)\leq\|h\|_{BV}\,\|f\|_{\infty}\tilde{\beta}(\sigma(X),Y). \tag{3.5}$$

Comparison results
Let $(X_k)_{k\in\mathbb{Z}}$ be a sequence of identically distributed real random variables. If the marginal distribution satisfies a concentration's condition $|F_X(x)-F_X(y)|\leq K|x-y|^a$ with $a\leq 1$, $K>0$, then (Dedecker *et al.* [24] Remark 5.1 p 104)

$$\tilde{\beta}_k \leq 2K^{1/(1+a)}\tau_{k,1}^{a/(a+1)} \leq 2K^{1/(1+a)}\tau_k^{a/(a+1)}.$$

In particular, if $\mathbb{P}_X$ has a density s with respect to the Lebesgue measure μ and if $s\in L^2(\mu)$, we have from Cauchy-Schwarz inequality

$$|F_X(x)-F_X(y)| = |\int \mathbf{1}_{[x,y]}sd\mu| \leq \|s\|_2\left(\int \mathbf{1}_{[x,y]}d\mu\right)^{1/2} = \|s\|_2\,|x-y|^{1/2},$$

thus

$$\tilde{\beta}_k \leq 2\,\|s\|_2^{2/3}\,\tau_k^{1/3}.$$

In particular, for any arithmetically [**AR**] τ-mixing process with rate $\theta>2$, we have

$$\tilde{\beta}_k \leq 2\,\|s\|_2^{2/3}\,(1+k)^{-(1+\theta)/3}. \tag{3.6}$$

Examples

Examples of β-mixing and τ-mixing sequences are well known, we refer to the books of Doukhan [28] and Bradley [19] for examples of β-mixing processes and to the book of Dedecker *et. al* [24] or the articles of Dedecker & Prieur [25], Prieur [58], and Comte *et. al* [22] for examples of τ-mixing sequences. One of the most important example is the following: a stationary, irreducible, aperiodic and positively recurent Markov chain $(X_i)_{i\geq 1}$ is β-mixing. However, many simple Markov chains are not β-mixing but are τ-mixing. For instance, it is known for a long time that if $(\epsilon_i)_{i\geq 1}$ are i.i.d Bernoulli $\mathcal{B}(1/2)$, then a stationary solution $(X_i)_{i\geq 0}$ of the equation

$$X_n = \frac{1}{2}(X_{n-1}+\epsilon_n),\ X_0 \text{ independent of } (\epsilon_i)_{i\geq 1}$$

is not β-mixing since $\beta_k = 1$ for any $k\geq 1$ whereas $\tau_k\leq 2^{-k}$ (see Dedecker & Prieur [25] Section 4.1). Another advantage of the coefficient τ is that it is easy to compute in many situations (see Dedecker & Prieur [25] Section 4).

3.2.3 Collections of models

We observe n identically distributed real valued random variables $X_1,...,X_n$ with common density s with respect to the Lebesgue measure μ. We assume that s belongs to the Hilbert space $L^2(\mu)$ endowed with norm $\|.\|_2$. We consider an orthonormal system $\{\psi_{j,k}\}_{(j,k)\in\Lambda}$ of $L_2(\mu)$ and a collection of models $(S_m)_{m\in\mathcal{M}_n}$ indexed by subsets $m\subset\Lambda$ for which we assume that the following assumptions are fulfilled:
$[M_1]$ for all $m\in\mathcal{M}_n$, S_m is the linear span of $\{\psi_{j,k}\}_{(j,k)\in m}$ with finite dimension $D_m = |m|\geq 2$ and $N_n = \max_{m\in\mathcal{M}_n} D_m$ satisfies $N_n\leq n$;
$[M_2]$ there exists a constant Φ such that

$$\forall m,m'\in\mathcal{M}_n, \forall t\in S_m, \forall t'\in S_{m'}, \|t+t'\|_\infty \leq \Phi\sqrt{\dim(S_m+S_{m'})}\|t+t'\|_2;$$

$[M_3]$ $D_m \leq D_{m'}$ implies that $m \subset m'$ and so $S_m \subset S_{m'}$.
As a consequence of Cauchy-Schwarz inequality, we have

$$\left\| \sum_{(j,k)\in m\cup m'} \psi_{j,k}^2 \right\|_\infty = \sup_{t\in S_m+S_{m'},t\neq 0} \frac{\|t\|_\infty^2}{\|t\|_2^2} \tag{3.7}$$

see Birgé & Massart [15] p 58. Three examples are usually developed as fulfilling this set of assumptions:
[T] trigonometric spaces: $\psi_{0,0}(x) = 1$ and for all $j \in \mathbb{N}^*$, $\psi_{j,1}(x) = \cos(2\pi jx)$, $\psi_{j,2}(x) = \sin(2\pi jx)$. $m = \{(0,0),(j,1),(j',2),\ 1 \leq j,j' \leq J_m\}$ and $D_m = 2J_m + 1$;
[P] regular piecewise polynomial spaces: S_m is generated by r polynomials $\psi_{j,k}$ of degree $k = 0,...,r-1$ on each subinterval $[(j-1)/J_m, j/J_m]$ for $j = 1,...,J_m$, $D_m = rJ_m$, $\mathcal{M}_n = \{m = \{(j,k),\ j = 1,...,J_m,\ k = 0,...,r-1\}, 1 \leq J_m \leq [n/r]\}$;
[W] spaces generated by dyadic wavelet with regularity r as described in Section 4.
For a precise description of those spaces and their properties, we refer to Birgé & Massart [15].

3.2.4 The estimator

Let $(X_n)_{n\in\mathbb{Z}}$ be a real valued stationary process and let P denote the law of X_0. Assume that P has a density s with respect to the Lebesgue measure μ and that $s \in L_2(\mu)$. Let $(S_m)_{m\in\mathcal{M}_n}$ be a collection of models satisfying assumptions $[M_1]$-$[M_3]$. We define $S_n = \cup_{m\in\mathcal{M}_n} S_m$, s_m and s_n the orthogonal projections of s onto S_m and S_n respectively, let $\mathbb{P}$ be the joint distribution of the observations $(X_n)_{n\in\mathbb{Z}}$ and let $\mathbb{E}$ be the corresponding expectation. We define the operators P_n, P and ν_n on $L^2(\mu)$ by

$$P_n t = \frac{1}{n}\sum_{i=1}^n t(X_i),\ Pt = \int t(x)s(x)d\mu(x),\ \nu_n(t) = (P_n - P)t.$$

All the real numbers that we shall introduce and which are not indexed by m or n are fixed constants. In order to define the penalized least-squares estimator, let us consider on $\mathbb{R} \times S_n$ the contrast function $\gamma(x,t) = -2t(x) + \|t\|_2^2$ and its empirical version $\gamma_n(t) = P_n\gamma(.,t)$. Minimizing $\gamma_n(t)$ over S_m leads to the classical projection estimator $\hat{s}_m$ on S_m. Let $\hat{s}_n$ be the projection estimator on S_n. Since $\{\psi_{j,k}\}_{(j,k)\in m}$ is an orthonormal basis of S_m one gets

$$\hat{s}_m = \sum_{(j,k)\in m} (P_n\psi_{j,k})\psi_{j,k} \text{ and } \gamma_n(\hat{s}_m) = -\sum_{(j,k)\in m} (P_n\psi_{j,k})^2.$$

Now, given a penalty function pen : $\mathcal{M}_n \to \mathbb{R}^+$, we define a selected model $\hat{m}$ as any element

$$\hat{m} \in \arg\min_{m\in\mathcal{M}_n} (\gamma_n(\hat{s}_m) + \mathrm{pen}(m)) \tag{3.8}$$

and a PLSE is defined as any $\tilde{s} \in S_{\hat{m}} \subset S_n$ such that

$$\gamma_n(\tilde{s}) + \mathrm{pen}(\hat{m}) = \inf_{m\in\mathcal{M}_n} (\gamma_n(\hat{s}_m) + \mathrm{pen}(m)). \tag{3.9}$$

3.2.5 Oracle inequalities

An ideal procedure for estimation chooses an oracle

$$m_o \in \operatorname{Arg}\min_{m\in\mathcal{M}_n}\{\|s-\hat{s}_m\|_2\}.$$

An oracle depends on the unknown s and on the data so that it is unknown in practice. In order to validate our procedure, we try to prove:
-non asymptotic oracle inequalities for the PLSE:

$$\mathbb{E}\left(\|s-\tilde{s}\|_2^2\right) \le L\inf_{m\in\mathcal{M}_n}\{\mathbb{E}\left(\|s-\hat{s}_m\|_2^2 + R(m,n)\right)\}, \tag{3.10}$$

for some constant $L \ge 1$ (as close to 1 as possible) and a remainder term $R(m,n) \ge 0$ possibly random, and small compared to $\mathbb{E}\left(\|s-\tilde{s}\|_2^2\right)$ if possible. This inequality compares the risk of the PLSE with the best deterministic choice of m. Since $\hat{m}$ is random, we prefer to prove a stronger form of oracle inequality :

$$\mathbb{E}\left(\|s-\tilde{s}\|_2^2\right) \le L\mathbb{E}\left(\inf_{m\in\mathcal{M}_n}\{\|s-\hat{s}_m\|_2^2 + R(m,n)\}\right), \tag{3.11}$$

or, when it is possible, deviation bounds for the PLSE:

$$\mathbb{P}\left(\|s-\tilde{s}\|_2^2 > L\inf_{m\in\mathcal{M}_n}\left(\|s-\hat{s}_m\|_2^2 + R(m,n)\right)\right) \le c_n, \tag{3.12}$$

where typically $c_n \le C/n^{1+\gamma}$ for some $\gamma > 0$. Inequality (3.12) proves that, asymptotically, the risk $\|s-\tilde{s}\|_2^2$ is almost surely the one of the oracle. Let

$$\Omega = \left\{\|s-\tilde{s}\|_2^2 > L\inf_{m\in\mathcal{M}_n}\left(\|s-\hat{s}_m\|_2^2 + R(m,n)\right)\right\}.$$

We have

$$\mathbb{E}\left(\|s-\tilde{s}\|_2^2\right) = \mathbb{E}\left(\|s-\tilde{s}\|_2^2 1_\Omega\right) + \mathbb{E}\left(\|s-\tilde{s}\|_2^2 1_{\Omega^c}\right).$$

It is clear that $\mathbb{E}\left(\|s-\tilde{s}\|_2^2 1_{\Omega^c}\right) \le L\mathbb{E}\left(\inf_{m\in\mathcal{M}_n}\left\{\|s-\hat{s}_m\|_2^2 + R(m,n)\right\}\right)$. Moreover, we have $\|s-\tilde{s}\|^2 = \|s-s_{\hat{m}}\|^2 + \|s_{\hat{m}}-\tilde{s}\|^2 \le \|s\|^2 + \Phi^2 D_{\hat{m}} \le \|s\|^2 + \Phi^2 n$, thus, when (3.12) holds, we have

$$\mathbb{E}\left(\|s-\tilde{s}\|_2^2 1_{\Omega^c}\right) \le (\|s\|^2 + \Phi^2 n)c_n \le \frac{C}{n^\gamma}.$$

Therefore, inequality (3.12) implies

$$\mathbb{E}\left(\|s-\tilde{s}\|_2^2\right) \le \mathbb{E}\left(\inf_{m\in\mathcal{M}_n}\{\|s-\hat{s}_m\|_2^2 + R(m,n)\}\right) + \frac{C}{n^\gamma}.$$

We can derive from these inequalities adaptive rates of convergence of the PLSE on Besov spaces (see Birgé & Massart [15] for example). In order to achieve this goal, we only have to prove a weaker form of oracle inequality where the remainder term $R(m,n) \le LD_m/n$ for some constant L, for all the models m with sufficiently large dimension. This will be detailed in Section 5.

3.3 Results for β-mixing processes

From now on, the letters κ, L and K, with various sub- or supscripts, will denote some constants which may vary from line to line. One shall use $L_{.}$ to indicate more precisely the dependence on various quantities, especially those which are related to the unknown s.
In this section, we give the following theorem for β-mixing sequences. It can be seen as a pathwise version of Theorem 3.1 in Comte & Merlevède [23].

Theorem 3.3.1 *Consider a collection of models satisfying* $[M_1]$, $[M_2]$ *and* $[M_3]$. *Assume that the process* $(X_n)_{n\in\mathbb{Z}}$ *is strictly stationary and arithmetically* $[\mathbf{AR}]$ *β-mixing with mixing rate* $\theta > 2$ *and that its marginal distribution admits a density* s *with respect to the Lebesgue measure* μ, *with* $s \in L_2(\mu)$.
Let κ_1 *be the constant defined in (3.2) and let* $\tilde{s}$ *be the PLSE defined by (3.9) with*

$$pen(m) = \frac{K\Phi^2\kappa_1 D_m}{n}, \text{ where } K > 4.$$

Then, for all $\kappa > 2$ *there exist* $c_0 > 0, L_s > 0$, $\gamma_1 > 0$ *and a sequence* $\epsilon_n \to 0$, *such that*

$$\mathbb{P}\left(\|\tilde{s}-s\|_2^2 > (1+\epsilon_n)\inf_{m\in\mathcal{M}_n, D_m\geq c_0(\log n)^{\gamma_1}}\left(\|s-s_m\|_2^2 + pen(m)\right)\right) \leq L_s\frac{(\log n)^{(\theta+2)\kappa}}{n^{\theta/2}}. \tag{3.13}$$

Remark: The term $K\Phi^2\kappa_1$ is the same as in Theorem 3.1 of Comte & Merlevède [23] but with a constant $K > 4$ instead of 320. The main drawback of this result is that the penalty term involves the constant κ_1 which is unknown in practice. However, Theorem 3.3.1 ensures that penalties proportional to the linear dimension of S_m lead to efficient model selection procedures. Thus we can use this information to apply the slope heuristic algorithm introduced by Birgé & Massart [17] in a Gaussian regression context and generalized by Arlot & Massart [7] to more general M-estimation frameworks. This algorithm calibrates the constant in front of the penalty term when the shape of an ideal penalty is available. The result of Arlot & Massart is proven for independent sequences, in a regression framework, but it can be generalized to the density estimation framework, for independent as well as for β or τ dependent data. This result is beyond the scope of this chapter and will be proved in chapter 4.
We have to consider the infimum in equation (3.13) over the models with sufficiently large dimensions. However, as noted by Arlot [5] (Remark 9 p 43), we can take the infimum over all the models in (3.13) if we add an extra term in (3.13). More precisely, we can prove that, with probability larger than $1 - L_s(\log n)^{(\theta+2)\kappa}/n^{\theta/2}$

$$\|\tilde{s}-s\|_2^2 \leq (1+\epsilon_n)\inf_{m\in\mathcal{M}_n}\left(\|s-\hat{s}_m\|_2^2 + \mathrm{pen}(m)\right) + L\frac{(\log n)^{\gamma_2}}{n}, \tag{3.14}$$

where $L > 0$ and $\gamma_2 > 0$.
Remark : The main improvement of Theorem 3.3.1 is that it gives an oracle inequality in probability, with a deviation bound of order $o(1/n)$ as soon as $\theta > 2$ instead of $\theta > 3$ in Comte & Merlevède [23]. Moreover, we do not require s to be bounded to prove our result.

Remark: When the data are independent, the proof of Theorem 3.3.1 can be used to obtain that the estimator $\tilde{s}$ chosen with a penalty term of order $K\Phi D_m/n$ satisfy an oracle inequality as (3.13). The main difference would be that $\kappa_1 = 1$, thus it can be used without a slope heuristic (even if this algorithm can be used also in this context to optimize the constant K) and the control of the probability would be $L_s e^{-\ln(n)^2/C_s}$ for some constants L_s, C_s instead of $L_s(\log n)^{(\theta+2)}\kappa n^{-\theta/2}$ in our theorem.

3.4 Results for τ-mixing sequences

In order to deal with τ-mixing sequences, we need to specify the basis $(\psi_{j,k})_{(j,k)\in\Lambda}$.

3.4.1 Wavelet basis

Throughout this section, r is a real number, $r \geq 1$ and we work with an r-regular orthonormal multiresolution analysis of $L_2(\mu)$, associated with a compactly supported scaling function ϕ and a compactly supported mother wavelet ψ. Without loss of generality, we suppose that the support of the functions ϕ and ψ is an interval $[A_1, A_2)$ where A_1 and A_2 are integers such that $A_2 - A_1 = A \geq 1$. Let us recall that ϕ and ψ generate an orthonormal basis by dilatations and translations.
For all $k \in \mathbb{Z}$ and $j \in \mathbb{N}^*$, let $\psi_{0,k} : x \to \sqrt{2}\phi(2x-k)$ and $\psi_{j,k} : x \to 2^{j/2}\psi(2^j x - k)$. The family $\{(\psi_{j,k})_{j\geq 0, k\in\mathbb{Z}}\}$ is an orthonormal basis of $L_2(\mu)$. Let us recall the following inequalities: for all $p \geq 1$, let $K_p = (\sqrt{2}\|\phi\|_p) \vee \|\psi\|_p$, $K_L = (2\sqrt{2}\text{Lip}(\phi)) \vee \text{Lip}(\psi)$, $K_{BV} = AK_L$.
Then for all $j \geq 0$, we have $\|\psi_{j,k}\|_\infty \leq K_\infty 2^{j/2}$,

$$\begin{aligned} \left\| \sum_{k\in\mathbb{Z}} |\psi_{j,k}| \right\|_\infty &\leq AK_\infty 2^{j/2} && (3.15) \\ \text{Lip}(\psi_{j,k}) &\leq K_L 2^{3j/2}, && (3.16) \\ \|\psi_{j,k}\|_{BV} &\leq K_{BV} 2^{j/2}. && (3.17) \end{aligned}$$

We assume that our collection $(S_m)_{m\in\mathcal{M}_n}$ satisfies the following assumption:
[W] dyadic wavelet generated spaces: let $J_n = [\log(n/2(A+1))/\log(2)]$ and for all $J_m = 1, ..., J_n$, let

$$m = \{(0,k), -A_2 < k < 2 - A_1\} \cup \{(j,k),\ 1 \leq j \leq J_m,\ -A_2 < k < -A_1 + 2^j\}$$

and S_m the linear span of $\{\psi_{j,k}\}_{(j,k)\in m}$. In particular, we have $D_m = (A-1)(J_m + 1) + 2^{J_m+1}$ and thus $2^{J_m+1} \leq D_m \leq (A-1)(J_m+1) + 2^{J_m+1} \leq A2^{J_m+1}$.

3.4.2 The τ-mixing case

The following result proves that we keep the same rate of convergence for the PLSE based on τ-mixing processes.

Theorem 3.4.1 *Consider the collection of models* **[W]**. *Assume that $(X_n)_{n\in\mathbb{Z}}$ is strictly stationary and arithmetically* **[AR]** *τ-mixing with mixing rate $\theta > 5$ and that* *its marginal distribution admits a density s with respect to the Lebesgue measure μ. Let $\tilde{s}$ be the PLSE defined by (3.9) with*

$$pen(m) = KAK_\infty K_{BV}\left(\sum_{l=0}^{\infty}\tilde{\beta}_l\right)\frac{D_m}{n},\ where\ K \geq 8.$$

Then there exist constants $c_0 > 0, \gamma_1 > 0$ and a sequence $\epsilon_n \to 0$ such that

$$\mathbb{E}\left(\|\tilde{s}-s\|_2^2\right) \leq (1+\epsilon_n)\left(\inf_{m\in\mathcal{M}_n,\ D_m\geq c_0(\log n)^{\gamma_1}}\|s-s_m\|_2^2 + pen(m)\right). \tag{3.18}$$

Remark : As in Theorem 3.3.1, the penalty term involves an unknown constant and we have a condition on the dimension of the models in (3.18). However, the slope heuristic can also be used in this context to calibrate the constant and a careful look at the proof shows that we can take the infimum over all models $m \in \mathcal{M}_n$ provided that we increase the constant K in front of the penalty term. Our result allows to derive rates of convergence in Besov spaces for the PLSE that correspond to the rates in the i.i.d. framework (see Proposition 3.5.2).

Remark : Theorem 3.4.1 gives an oracle inequality for the PLSE built on τ-mixing sequences. This inequality is not pathwise and the constants involved in the penalty term are not optimal. This is due to technical reasons, mainly because we use the coupling result (3.4) instead of (3.3). However, we recover the same kind of oracle inequality as in the i.i.d. framework (Birgé and Massart [15]) under weak assumptions on the mixing coefficients since we only require arithmetical **[AR]** τ-mixing assumptions on the process $(X_n)_{n\in\mathbb{Z}}$. This is the first result for these processes up to our knowledge.

Let us mention here Theorem 4.1 in Comte & Merlevède [23]. They consider α-mixing processes (for a definition of the coefficient α and its properties, we refer to Rio [60]). They make geometrical **[GEO]** α-mixing assumptions on the processes and consider penalties of order $L\log(n)D_m/n$ to get an oracle inequality. This leads to a logarithmic loss in the rates of convergence. They get the optimal rate under an extra assumption (namely Assumption [Lip] in Section 3.2). There exist random processes that are τ-mixing and not α-mixing (see Dedecker & Prieur [25]), however, the comparison of these coefficients is difficult in general and our method can not be applied in this context.

The constants c_0, γ_1, n_o are given in the end of the proof.

Remark : Inequality (2.6) can be improved under stronger assumptions on s. For example, when s is bounded, we have $\tilde{\beta}_k \leq C\sqrt{\tau_k}$. Under this assumption and $\theta > 3$, we can prove that the estimator $\tilde{s}$ satisfies the inequality

$$\mathbb{E}\left(\|\tilde{s}-s\|_2^2\right) \leq (1+\epsilon_n)\left(\inf_{m\in\mathcal{M}_n,\ D_m\geq c_0(\log n)^{\gamma_1}}\|s-s_m\|_2^2 + \mathrm{pen}(m)\right) + \frac{(\log n)^{\kappa(\theta+1)}}{n^{(\theta-3)/2}}.$$

When $\theta < 5$, the extra term $(\log n)^{\kappa(\theta+1)}/n^{(\theta-3)/2}$ may be larger than the main term $\inf_{m\in\mathcal{M}_n,\ D_m\geq c_0(\log n)^{\gamma_1}}\|s-s_m\|_2^2+\mathrm{pen}(m)$. In this case, we don't know if our control remains optimal. On the other hand, Proposition 3.5.2 ensures that $\tilde{s}$ is adaptive over the class of Besov balls when $\theta \geq 5$.

3.5 Minimax results

3.5.1 Approximation results on Besov spaces

Besov balls.
Throughout this section, $\Lambda = \{(j,k),\ j \in \mathbb{N},\ k \in \mathbb{Z}\}$ and $\{\psi_{j,k},\ (j,k) \in \Lambda\}$ denotes an r-regular wavelet basis as introduced in Section 4.1. Let α, p be two positive numbers such that $\alpha+1/2-1/p > 0$. For all functions $t \in L_2(\mu)$, $t = \sum_{(j,k)\in\Lambda} t_{j,k}\psi_{j,k}$, we say that t belongs to the Besov ball $B_{\alpha,p,\infty}(M_1)$ on the real line if $\|t\|_{\alpha,p,\infty} \leq M_1$ where

$$\|t\|_{\alpha,p,\infty} = \sup_{j\in\mathbb{N}} 2^{j(\alpha+1/2-1/p)} \left(\sum_{k\in\mathbb{Z}} |t_{j,k}|^p\right)^{1/p}.$$

It is easy to check that if $p \geq 2$ $B_{\alpha,p,\infty}(M_1) \subset B_{\alpha,2,\infty}(M_1)$ so that upper bounds on $B_{\alpha,2,\infty}(M_1)$ yield upper bounds on $B_{\alpha,p,\infty}(M_1)$.

Approximation results on Besov spaces.
We have the following result (Birgé & Massart [15] Section 4.7.1). Suppose that the support of s equals $[0,1]$ and that s belongs to the Besov ball $B_{\alpha,2,\infty}(1)$, then whenever $r > \alpha - 1$,

$$\|s - s_m\|_2^2 \leq \frac{\|s\|_{\alpha,2,\infty}^2}{4(4^\alpha - 1)} 2^{-2J_m\alpha} \leq \frac{(2A)^{2\alpha}\|s\|_{\alpha,2,\infty}^2}{4(4^\alpha - 1)} D_m^{-2\alpha} \tag{3.19}$$

3.5.2 Minimax rates of convergence for the PLSE

We can derive from Theorems 3.3.1 and 3.4.1 adaptation results to unknown smoothness over Besov Balls.

Proposition 3.5.1 *Assume that the process $(X_n)_{n\in\mathbb{Z}}$ is stricly stationary and arithmetically* **[AR]** *β-mixing with mixing rate $\theta > 2$ and that its marginal distribution admits a density s with respect to the Lebesgue measure μ, that s is supported in $[0,1]$ and that $s \in L^2(\mu)$. For all $\alpha, M_1 > 0$, the PLSE $\tilde{s}$ defined in Theorem 3.3.1 for the collection of models* **[W]** *satisfies*

$$\forall \kappa > 2, \sup_{s\in B_{\alpha,2,\infty}(M_1)} \mathbb{P}\left(\|\tilde{s} - s\|_2^2 > L_{M_1,\alpha,\theta} n^{-2\alpha/(2\alpha+1)}\right) \leq \frac{L_{M_1}(\log n)^{(\theta+2)\kappa}}{n^{\theta/2}}.$$

Proposition 3.5.2 *Assume that the process $(X_n)_{n\in\mathbb{Z}}$ is stricly stationary and arithmetically* **[AR]** *τ-mixing with mixing rate $\theta > 5$ and that its marginal distribution admits a density s with respect to the Lebesgue measure μ, that s is supported in $[0,1]$ and that $s \in L^2(\mu)$. For all $\alpha, M_1 > 0$, the PLSE $\tilde{s}$ defined in Theorem 3.4.1 satisfies*

$$\sup_{s\in B_{\alpha,2,\infty}(M_1)} \mathbb{E}\left(\|\tilde{s} - s\|_2^2\right) \leq L_{M_1,\alpha,\theta} n^{-2\alpha/(2\alpha+1)}.$$

Remark: Proposition 3.5.2 can be compared to Theorem 3.1 in Gannaz & Wintenberger [34]. They prove near minimax results for the thresholded wavelet estimator introduced by Donoho *et al.* [27] in a $\tilde{\phi}$-dependent setting (for a definition of the coefficient $\tilde{\phi}$, we refer to Dedecker & Prieur [25]). Basically, with our notations, their result can be stated as follows: if $(X_n)_{n\in\mathbb{Z}}$ is $\tilde{\phi}$-mixing with $\tilde{\phi}_1(r)\leq Ce^{-ar^b}$ for some constants C,a,b, then the thresholded wavelet estimator $\hat{s}$ of s satisfies

$$\forall\alpha>0,\ \forall p>1,\qquad \sup_{s\in B_{\alpha,p,\infty}(M_1)\cap L^\infty(M)}\mathbb{E}\left(\|\hat{s}-s\|_2^2\right)\leq L_{M,M_1,\alpha,p}\left(\frac{\log n}{n}\right)^{2\alpha/(2\alpha+1)}.$$

The main advantage of their result is that they can deal with Besov balls with regularity $1<p<2$. However, in the regular case, when $p\geq 2$, we have been able to remove the extra $\log n$ factor. Moreover, our result only requires arithmetical **[AR]** rates of convergence for the mixing coefficients and we do not have to suppose that s is bounded.

3.6 Proofs.

3.6.1 Proofs of the minimax results.

Proof of Proposition 3.5.1:

Let $\alpha>0$ and $M_1>0$ and assume that $s\in B_{\alpha,2,\infty}(M_1)$. Let $\tilde{\mathcal{M}}_n=\{m\in\mathcal{M}_n, D_m>c_0(\log n)^{\gamma_1}\}$. By Theorem 3.3.1, there exists a constant $L_\theta>0$ such that

$$\mathbb{P}\left(\|\tilde{s}-s\|_2^2>L_\theta\inf_{m\in\tilde{\mathcal{M}}_n}\left\{\|s-s_m\|_2^2+\frac{D_m}{n}\right\}\right)\leq\frac{L_s(\log n)^{(\theta+2)\kappa}}{n^{\theta/2}}.\qquad(3.20)$$

It appears from the proof of Theorem 3.3.1 that the constant L_s depends only on $\|s\|_2$ and that it is a nondecreasing function of $\|s\|_2$ so that L_s can be uniformly bounded over $B_{\alpha,2,\infty}(M_1)$ by a constant L_{M_1} so that, by (3.20)

$$\mathbb{P}\left(\|\tilde{s}-s\|_2^2>L_\theta\inf_{m\in\tilde{\mathcal{M}}_n}\left\{\|s-s_m\|_2^2+\frac{D_m}{n}\right\}\right)\leq\frac{L_{M_1}(\log n)^{(\theta+2)\kappa}}{n^{\theta/2}}.$$

In particular, for a model m in $\mathcal{M}_n$ with dimension D_m such that

$$c_0(\log n)^{\gamma_1}\leq L_1n^{1/(2\alpha+1)}\leq D_m\leq L_2n^{1/(2\alpha+1)},$$

we have

$$\mathbb{P}\left(\|\tilde{s}-s\|_2^2>L_\theta\left(\|s-s_m\|_2^2+\frac{D_m}{n}\right)\right)\leq\frac{L_{M_1}(\log n)^{(\theta+2)\kappa}}{n^{\theta/2}}.$$

Since s belongs to $B_{\alpha,2,\infty}(M_1)$, we can use Inequality (3.19) to get

$$\|s-s_m\|_2^2\leq L_{\alpha,M_1}D_m^{-2\alpha}.$$

Thus we obtain

$$\mathbb{P}\left(\|\tilde{s}-s\|_2^2>L_{M_1,\alpha,\theta}n^{-2\alpha/(2\alpha+1)}\right)\leq\frac{L_{M_1}(\log n)^{(\theta+2)\kappa}}{n^{\theta/2}}.\square$$

Proof of Proposition 3.5.2:

Let $\alpha>0$ and $M_1>0$ and assume that $s\in B_{\alpha,2,\infty}(M_1)$. By Theorem 3.4.1, we have

$$\mathbb{E}\left(\|\tilde{s}-s\|_2^2\right)\ \leq\ L_\theta\left(\inf_{m\in\tilde{\mathcal{M}}_n}\{\|s-s_m\|_2^2+\frac{D_m}{n}\}\right).$$

Inequality (3.19) leads to $\|s-s_m\|_2^2 \leq L_{\alpha,M_1} D_m^{-2\alpha}$, so that for a model m in $\tilde{M}_n$ with dimension D_m such that

$$c_0(\log n)^{\gamma_1} \leq L_1 n^{1/(2\alpha+1)} \leq D_m \leq L_2 n^{1/(2\alpha+1)},$$

we find

$$\mathbb{E}\left(\|\tilde{s}-s\|_2^2\right) \leq L_{\theta,\alpha,M_1} n^{-2\alpha/(2\alpha+1)}.\square$$

3.6.2 Proof of Theorem 3.3.1:

For all m_o in $\mathcal{M}_n$, we have, by definition of $\hat{m}$

$$\begin{aligned}
\gamma_n(\tilde{s}) + \mathrm{pen}(\hat{m}) &\leq \gamma_n(\hat{s}_{m_o}) + \mathrm{pen}(m_o) \\
P\gamma(\tilde{s}) + \nu_n\gamma(\tilde{s}) + \mathrm{pen}(\hat{m}) &\leq P\gamma(\hat{s}_{m_o}) + \nu_n\gamma(\hat{s}_{m_o}) + \mathrm{pen}(m_o) \\
P\gamma(\tilde{s}) - P\gamma(s) - 2\nu_n\tilde{s} + \mathrm{pen}(\hat{m}) &\leq P\gamma(\hat{s}_{m_o}) - P\gamma(s) - 2\nu_n\hat{s}_{m_o} + \mathrm{pen}(m_o)
\end{aligned}$$

Since for all $t \in L_2(\mu)$, $P\gamma(t) - P\gamma(s) = \|t-s\|_2^2$, we have

$$\|s-\tilde{s}\|_2^2 \leq \|s-\hat{s}_{m_o}\|_2^2 + \mathrm{pen}(m_o) - V(m_o) - (\mathrm{pen}(\hat{m}) - V(\hat{m})) - 2\nu_n(s_{m_o} - s_{\hat{m}}), \quad (3.21)$$

where, for all $m \in \mathcal{M}_n$

$$V(m) = 2\nu_n(\hat{s}_m - s_m) = 2\sum_{(j,k)\in m} \nu_n^2(\psi_{j,k}).$$

This decomposition is different from the one used in Birgé & Massart [15] and in Comte & Merlevède [23]. It allows to improve the constant in the oracle inequality in the β-mixing case. Moreover, we choose to prove an oracle inequality of the form (3.12) for β-mixing sequences, which allows to assume only $\theta > 2$ instead of $\theta > 3$. Let us now give a sketch of the proof:

1. we build an event Ω_C with $\mathbb{P}(\Omega_C^c) \leq p\beta_q$ such that, on Ω_C, $\nu_n = \nu_n^*$, where ν_n^* is built with independent data. A suitable choice of the integers p and q leads to $p\beta_q \leq C(\ln n)^r n^{-\theta/2}$.

2. We use the concentration's inequality (3.7.4) of Birgé & Massart [15] for χ^2-type statistics, derived from Talagrand's inequality. This allows us to find $p_1(m)$ such that on an event Ω_1 with $\mathbb{P}(\Omega_1^c \cap \Omega_C) \leq L_{1,s}c_n$

   $$\sup_{m\in\mathcal{M}_n} \{V(m) - p_1(m)\} \leq 0.$$

   $c_n < C(\ln n)^r n^{-\theta/2}$ and $L_{1,s}$ is some constant depending on s.

3. From Bernstein's inequality, we prove that, for all $m, m' \in \mathcal{M}_n$, there exists $p_2(m,m')$ such that, for all $\eta > 0$, on an event Ω_2 with $\mathbb{P}(\Omega_2^c \cap \Omega_C) \leq L_{2,s}c_n$,

   $$\sup_{m,m'\in\mathcal{M}_n} \left\{ \nu_n(s_m - s_{m'}) - \frac{\eta}{2}p_2(m,m') - \frac{\|s_m - s_{m'}\|_2^2}{2\eta} \right\} \leq 0.$$

   Moreover, for all $m, m' \in \mathcal{M}_n$, $p_2(m,m') \leq p_2(m,m) + p_2(m',m')$.

4. We have $\|s_{\hat{m}} - s_{m_o}\|_2^2 \leq \|s_{\hat{m}} - s\|_2^2 + \|s - s_{m_o}\|_2^2$ because $s_{\hat{m}} - s_{m_o}$ is either the projection of $s_{\hat{m}} - s$ onto S_{m_o} or the projection of $s - s_{m_o}$ onto $S_{\hat{m}}$. Take $\text{pen}(m) \geq p_1(m) + \eta p_2(m,m)$, we have, on $\Omega_1 \cap \Omega_2 \cap \Omega_C$

$$\begin{aligned} \|s-\tilde{s}\|_2^2 &\leq \|s-\hat{s}_{m_o}\|_2^2 - \frac{V_{m_o}}{2} + \text{pen}(m_o) - \frac{V_{m_o}}{2} && (3.22)\\ &\quad -(\text{pen}(\hat{m}) - p_1(\hat{m})) - (p_1(\hat{m}) - V(\hat{m})) - 2\nu_n(s_{m_o} - s_{\hat{m}}) \\ &\leq \|s - s_{m_o}\|_2^2 + \text{pen}(m_o) - \frac{V(m_o)}{2} - \eta p_2(\hat{m},\hat{m}) \\ &\quad + \eta p_2(\hat{m}, m_o) + \frac{\|s_{m_o} - s_{\hat{m}}\|_2^2}{\eta} && (3.23)\\ \left(1 - \frac{1}{\eta}\right)\|s-\tilde{s}\|_2^2 &\leq (1+\frac{1}{\eta})\|s - s_{m_o}\|_2^2 + \text{pen}(m_o) + \eta p_2(m_o, m_o). && (3.24) \end{aligned}$$

In (3.23), we used that $V(m_o) = 2\|s_{m_o} - \hat{s}_{m_o}\|_2^2 \geq 0$. In (3.24), we used that $V_{m_o} \geq 0$. Pythagoras Theorem gives

$$\|s - \hat{s}_{m_o}\|_2^2 - \frac{V(m_o)}{2} = \|s - s_{m_o}\|_2^2 \text{ and}; \|s - s_{\hat{m}}\|_2^2 \leq \|s - \tilde{s}\|_2^2 .$$

Finally, we prove that we can choose $\eta = (\log n)^\gamma$, with $\gamma > 0$ such that $\eta p_2(m_o, m_o) = o(\text{pen}(m_o))$ and we conclude the proof of (3.3.1) from the previous inequalities.
We decompose the proof in several claims corresponding to the previous steps.
Claim 1 : For all $l = 0, ..., p-1$, let us define $A_l = (X_{2lq+1}, \ ..., X_{(2l+1)q})$ and $B_l = (X_{(2l+1)q+1}, ..., X_{(2l+2)q})$. There exist random vectors $A_l^* = (X_{2lq+1}^*, ..., X_{(2l+1)q}^*)$ and $B_l^* = (X_{(2l+1)q+1}^*, ..., X_{(2l+2)q}^*)$ such that for all $l = 0, ..., p-1$:

1. A_l^* and A_l have the same law,
2. A_l^* is independent of $A_0, ..., A_{l-1}, A_0^* ..., A_{l-1}^*$
3. $\mathbb{P}(A_l \neq A_l^*) \leq \beta_q$

the same being true for the variables B_l.

Proof of Claim 1 :

The proof is derived from Berbee's lemma, we refer to Proposition 5.1 in Viennet [70] for further details about this construction.□

Hereafter, we assume that, for some $\kappa > 2$, $\sqrt{n}(\log n)^\kappa/2 \leq p \leq \sqrt{n}(\log n)^\kappa$ and for the sake of simplicity that $pq = n/2$, the modifications needed to handle the extra term when $q = [n/(2p)]$ being straightforward. Let $\Omega_C = \{\forall l = 0, ..., p-1 \ A_l = A_l^*, \ B_l = B_l^*\}$. We have

$$\mathbb{P}(\Omega_C^c) \leq 2p\beta_q \leq 2^{2+\theta}\frac{(\log n)^{(\theta+2)\kappa}}{n^{\theta/2}}.$$

Let us first deal with the quadratic term $V(m)$.
Claim 2 : *Under the assumptions of Theorem 3.3.1, let $\epsilon > 0$, $1 < \gamma < \kappa/2$. We define $L_1^2 = 2\Phi^2\kappa_1$, $L_2^2 = 8\Phi^{3/2}\sqrt{\kappa_2}$, $L_3 = 2\Phi\kappa(\epsilon)$ and*

$$L_{1,m} = 4\left((1+\epsilon)L_1 + L_2\sqrt{\frac{(\log n)^\gamma}{D_m^{1/4}}} + \frac{L_3}{(\log n)^{\kappa-\gamma}}\right)^2 . \tag{3.25}$$

Then, we have

$$\mathbb{P}\left(\sup_{m\in\mathcal{M}_n}\left\{V(m)-\frac{L_{1,m}D_m}{n}\right\}\geq 0\cap\Omega_C\right)\leq L_{s,\gamma}\exp\left(-\frac{(\log n)^\gamma}{\sqrt{\|s\|_2}}\right).$$

where $L_{s,\gamma}=2\sum_{D=1}^\infty\exp(-(\log D)^\gamma/\|s\|_2^{1/2})$*. In particular, for all* $r>0$*, there exists a constant* $L'_{s,r}$ *depending on* $\|s\|_2$*, such that*

$$\mathbb{P}\left(\sup_{m\in\mathcal{M}_n}\left\{V(m)-\frac{L_{1,m}D_m}{n}\right\}\geq 0\cap\Omega_C\right)\leq\frac{L'_{s,r}}{n^r}.$$

Remark : When $(L_2/L_1)^8(\log n)^{4(2\kappa-\gamma)}\leq D_m\leq n$, we have

$$L_{1,m}\leq\left[1+\epsilon+\left(1+\frac{\sqrt{2}\kappa(\epsilon)}{\sqrt{\kappa_1}}\right)(\log n)^{-(\kappa-\gamma)}\right]^2 4L_1^2.$$

Proof of Claim 2 :

Let $P_n^*(t)=\sum_{i=1}^n t(X_i^*)/n$ and $\nu_n^*(t)=(P_n^*-P)t$, we have

$$V(m)\mathbf{1}_{\Omega_C}=2\sum_{(j,k)\in m}(\nu_n^*)^2(\psi_{j,k})\mathbf{1}_{\Omega_C}.$$

Let $B_1(S_m)=\{t\in S_m;\ \|t\|_2\leq 1\}$. $\forall t\in B_1(S_m)$, let $\bar{t}(x_1,...,x_q)=\sum_{i=1}^q t(x_i)/2q$ and for all functions $g:\mathbb{R}^q\to\mathbb{R}$ let

$$P_{A,p}^*g=\frac{1}{p}\sum_{j=0}^{p-1}g(A_j^*),\ P_{B,p}^*g=\frac{1}{p}\sum_{j=0}^{p-1}g(B_j^*),\ \bar{P}g=\int g\mathbb{P}_A(d\mu),$$

$$\text{and } \bar{\nu}_{A,p}g=(P_{A,p}^*-\bar{P})g,\ \bar{\nu}_{B,p}g=(P_{B,p}^*-\bar{P})g.$$

Now we have

$$\sum_{(j,k)\in m}(\nu_n^*)^2(\psi_{j,k})\leq 2\sum_{(j,k)\in m}\bar{\nu}_{A,p}^2\bar{\psi}_{j,k}+2\sum_{(j,k)\in m}\bar{\nu}_{B,p}^2\bar{\psi}_{j,k}.$$

In order to handle these terms, we use Proposition 3.7.4 which is stated in Section 7. Taking

$$B_m^2=\sum_{(j,k)\in m}\mathrm{Var}(\bar{\psi}_{j,k}(A_1)),\ V_m^2=\sup_{t\in B_1(S_m)}\mathrm{Var}(\bar{t}(A_1)),\ \text{and } H_m^2=\left\|\sum_{(j,k)\in m}(\bar{\psi}_{j,k})^2\right\|_\infty,$$

we have

$$\forall x>0,\ \mathbb{P}\left(\sqrt{\sum_{(j,k)\in m}\bar{\nu}_{A,p}^2\bar{\psi}_{j,k}}\geq\frac{(1+\epsilon)}{\sqrt{p}}B_m+V_m\sqrt{\frac{2x}{p}}+\kappa(\epsilon)\frac{H_m x}{p}\right)\leq e^{-x}. \quad (3.26)$$

In order to evaluate B_m, V_m and H_m, we use Viennet's inequality (3.54). There exists a function b such that, for all $p=1,2$, $P|b|^p\leq\kappa_p$ where κ_p is defined in (3.2) and for all functions $t\in L_2(\bar{P})$,

$$\mathrm{Var}(\bar{t}(A_1))\leq\frac{1}{q}Pbt^2.$$

Thus

$$B_m^2 = \sum_{(j,k)\in m} \mathrm{Var}(\bar{\psi}_{j,k}(A_1)) \leq \frac{1}{q}\sum_{(j,k)\in m} Pb\psi_{j,k}^2 \leq \left\|\sum_{(j,k)\in m}\psi_{j,k}^2\right\|_\infty \frac{\kappa_1}{q}.$$

From Assumption $[M_2]$, $\left\|\sum_{(j,k)\in m}\psi_{j,k}^2\right\|_\infty \leq \Phi^2 D_m$, thus,

$$B_m^2 \leq \frac{\Phi^2\kappa_1 D_m}{q}. \tag{3.27}$$

From Viennet's and Cauchy-Schwarz inequalities

$$V_m^2 = \sup_{t\in B_1(S_m)} \mathrm{Var}(\bar{t}(A_1)) \leq \sup_{t\in B_1(S_m)} \frac{Pbt^2}{q} \leq \sup_{t\in B_1(S_m)} \|t\|_\infty \frac{(Pt^2)^{1/2}(Pb^2)^{1/2}}{q}.$$

Since $t \in B_1(S_m)$, we have by Cauchy-Schwarz inequality

$$(Pt^2)^{1/2} \leq (\|t\|_\infty \|t\|_2 \|s\|_2)^{1/2} \leq (\|t\|_\infty \|s\|_2)^{1/2}.$$

From Assumption $[M_2]$, we have $\|t\|_\infty \leq \Phi\sqrt{D_m}$, and from Viennet's inequality $Pb^2 \leq \kappa_2 < \infty$, thus we obtain

$$V_m^2 \leq \Phi^{3/2}(\|s\|_2\,\kappa_2)^{1/2}\frac{D_m^{3/4}}{q}. \tag{3.28}$$

Finally, from Assumption $[M_2]$, we have, using Cauchy-Schwarz inequality

$$H_m^2 = \left\|\sum_{(j,k)\in m}\bar{\psi}_{j,k}^2\right\|_\infty \leq \frac{1}{4}\left\|\sum_{(j,k)\in m}\psi_{j,k}^2\right\|_\infty \leq \frac{\Phi^2 D_m}{4}. \tag{3.29}$$

Let $y_n > 0$. We define

$$L_m = \left((1+\epsilon)L_1 + L_2\sqrt{\frac{(\log D_m)^\gamma + y_n}{2D_m^{1/4}}} + L_3\frac{(\log D_m)^\gamma + y_n}{2(\log n)^\kappa}\right)^2.$$

We apply Inequality (3.26) with $x = ((\log D_m)^\gamma + y_n)/\|s\|_2^{1/2}$ and the evaluations (3.27), (3.28) and (3.29). Recalling that $1/p \leq 2/(\sqrt{n}(\log n)^\kappa)$, this leads to

$$\mathbb{P}\left(\sum_{(j,k)\in m}\bar{\nu}_{A,p}^2\bar{\psi}_{j,k} \geq \frac{L_m D_m}{n}\right) \leq \exp\left(-\frac{(\log D_m)^\gamma}{\sqrt{\|s\|_2}}\right)\exp(-\frac{y_n}{\sqrt{\|s\|_2}}).$$

In order to give an upper bound on $H_m x$, we used that the support of s in included in $[0,1]$, thus

$$1 = \|s\|_1 \leq \|s\|_2\,.$$

The result follows by taking $y_n = (\log n)^\gamma \geq (\log D_m)^\gamma$.□

Claim 3. *We keep the notations $\kappa/2 > \gamma > 1$, L_2 of the proof of Claim 2. For all $m, m' \in \mathcal{M}_n$ we take*

$$L_{m,m'} = 4\left(L_2\sqrt{\frac{(\log n)^\gamma}{(D_m \vee D_{m'})^{1/4}}} + \frac{4\Phi}{3(\log n)^{\kappa-\gamma}}\right)^2, \tag{3.30}$$

we have, for all $\eta > 0$,

$$\mathbb{P}\left(\sup_{m,m'\in\mathcal{M}_n} \nu_n^*(s_m - s_{m'}) - \frac{\|s_m - s_{m'}\|_2^2}{2\eta} - \frac{\eta}{2}\frac{L_{m,m'}(D_m \vee D_{m'})}{n} > 0\right) \le L_{s,\gamma} e^{-\frac{(\log n)^\gamma}{\|s\|_2^{1/2}}}$$

with $L_{s,\gamma} = 2\sum_{m,m'\in\mathcal{M}_n} e^{-\frac{(\log(D_m \vee D_{m'}))^\gamma}{\|s\|_2^{1/2}}}$.

Remark : The constant $L_{s,\gamma}$ is finite since for all $x, y > 0$, $(\log(x\vee y))^\gamma \ge ((\log x)^\gamma + (\log y)^\gamma)/2$.

As in Claim 2, when $(L_2/L_1)^8(\log n)^{4(2\kappa-\gamma)} \le D_m \le n$, we have

$$L_{m,m'} \le \left(1 + \frac{2^{3/2}}{3\sqrt{\kappa_1}}\right)^2 (\log n)^{-2(\kappa-2\gamma)} 4L_1^2.$$

Proof of Claim 3.

We keep the notations of the proof of Claim 2 and for $m, m' \in \mathcal{M}_n$, let $t_{m,m'} = (s_m - s_{m'})/\|s_m - s_{m'}\|_2$. We use the inequality $2ab \le a^2\eta^{-1} + b^2\eta$, which holds for all $a, b \in \mathbb{R}$, $\eta > 0$. This leads to

$$\begin{aligned}
\nu_n^*(s_m - s_{m'}) &= \|s_m - s_{m'}\|_2 \nu_n^*(t_{m,m'}) \le \frac{\|s_m - s_{m'}\|_2^2}{2\eta} + \frac{\eta}{2}(\nu_n^*(t_{m,m'}))^2 \\
&= \frac{\|s_m - s_{m'}\|_2^2}{2\eta} + \frac{\eta}{2}(\bar{\nu}_{A,p}(\bar{t}_{m,m'}) + \bar{\nu}_{B,p}(\bar{t}_{m,m'}))^2 \\
&\le \frac{\|s_m - s_{m'}\|_2^2}{2\eta} + \eta(\bar{\nu}_{A,p}(\bar{t}_{m,m'}))^2 + \eta(\bar{\nu}_{B,p}(\bar{t}_{m,m'}))^2.
\end{aligned}$$

Now from Bernstein's inequality (see Section 7), we have

$$\forall x > 0,\ \mathbb{P}\left(\bar{\nu}_{A,p}(\bar{t}_{m,m'}) > \sqrt{\frac{2\mathrm{Var}(\bar{t}_{m,m'}(A_1))x}{p}} + \frac{\|\bar{t}_{m,m'}\|_\infty x}{3p}\right) \le e^{-x}. \tag{3.31}$$

From Viennet's and Cauchy-Schwarz inequalities, we have

$$\mathrm{Var}(\bar{t}_{m,m'}(A_1)) \le \frac{Pbt_{m,m'}^2}{q} \le \frac{\|t_{m,m'}\|_\infty\sqrt{Pb^2Pt_{m,m'}^2}}{q}.$$

Moreover

$$Pb^2 \le \kappa_2,\ Pt_{m,m'}^2 \le \|t_{m,m'}\|_\infty\|t_{m,m'}\|_2\|s\|_2.$$

Since $t_{m,m'} \in S_m \cup S_{m'}$ and $\|t_{m,m'}\|_2 = 1$, we have, from Assumption $[M_2]$ $\|t_{m,m'}\|_\infty \le \Phi\sqrt{D_m \vee D_{m'}}$. Let $y_n > 0$. We apply Inequality (3.31) with $x = [(\log(D_m \vee D_{m'}))^\gamma + y_n]/\|s\|_2^{1/2}$. We define

$$\frac{L'_{m,m'}}{4} = \left(L_2\sqrt{\frac{(\log(D_m \vee D_{m'}))^\gamma + y_n}{2(D_m \vee D_{m'})^{1/4}}} + \frac{4\Phi\left[(\log(D_m \vee D_{m'}))^\gamma + y_n\right]}{6(\log n)^\kappa}\right)^2,$$

we have

$$\mathbb{P}\left(\bar{\nu}_{A,p}(\bar{t}_{m,m'}) > \sqrt{\frac{L'_{m,m'}(D_m \vee D_{m'})}{4n}}\right) \leq \exp\left(-\frac{(\log(D_m \vee D_{m'}))^{\gamma}}{\|s\|_2^{1/2}}\right) e^{-y_n/\|s\|_2^{1/2}}.$$

The result follows by taking $y_n = (\log n)^{\gamma}$ and using $2 \leq D_m \leq n$.
Conclusion of the proof:
Let $\eta > 0$ and $\mathrm{pen}'(m) \geq (L_{1,m} + \eta L_{m,m})D_m/n$ where $L_{1,m}$ and $L_{m,m}$ are defined respectively by (3.25) and (3.30). From Claims 1, 2 and 3 and (3.24), we obtain that, for all m_o and with probability larger than $L_{s,\theta}(\log n)^{(\theta+2)\kappa}n^{-\theta/2}$

$$(1 - \frac{1}{\eta})\,\|s - \tilde{s}\|_2^2 \leq (1 + \frac{1}{\eta})\,\|s - s_{m_o}\|_2^2 + \mathrm{pen}'(m_o) + \eta L(m_o, m_o)\frac{D_{m_o}}{n}. \qquad (3.32)$$

Assume that $D_m \geq (L_2/L_1)^8(\log n)^{4(2\kappa-\gamma)}$, then we have from remarks 3.6.2 and 3.6.2

$$\begin{aligned} L_{1,m} &\leq \left[1 + \epsilon + \left(1 + \frac{2\kappa(\epsilon)}{\sqrt{\kappa_1}}\right)(\log n)^{-(\kappa-2\gamma)}\right]^2 4L_1^2 \text{ and} \\ L_{m,m} &\leq \left(1 + \frac{2^{3/2}}{3\sqrt{\kappa_1}}\right)^2 (\log n)^{-2(\kappa-\gamma)} 4L_1^2. \end{aligned}$$

Take $\eta = (\log n)^{\kappa-\gamma}$, we have $(L_{1,m_o} + \eta L_{m_o,m_o})D_{m_o}/n \leq C\mathrm{pen}(m_o)$. Fix $\epsilon > 0$ such that $[1+\epsilon]^2 < K/4$. Since $\kappa > \gamma$, for $n \geq n_o$, we have $L_{1,m} + \eta L_{m,m} \leq KL_1^2$, thus, inequality (3.13) follows follows from (3.32) as soon as $n > n_o$. We remove the condition $n > n_o$ by improving the constant L_s in (3.13) if necessary.□

3.6.3 Proof of Theorem 3.4.1.

The proof follows the previous one, the main difference is that the coupling lemma (Claim 1) as well as the covariance inequalities are much harder to handle in the τ-mixing case. This leads to more technical computations to recover the results obtained in the β-mixing case (see Claims 2, 3 and the proof of inequality (3.45)). We start with the decomposition (3.21). As in the previous proof, the decomposition of the risk given in Birgé & Massart [15] or in Comte & Merlevède [23] could be used. This leads to a loss in the constant in front of the main term in (3.18) without avoiding any of the main difficulties. We divide the proof in four claims.
Claim 1 : *For all $l = 0, ..., p-1$, let us denote by $A_l = (X_{2lq+1}, ..., X_{(2l+1)q})$ and $B_l = (X_{(2l+1)q+1}, ..., X_{(2l+2)q})$. There exist random vectors $A_l^* = (X^*_{2lq+1}, ..., X^*_{(2l+1)q})$ and $B_l^* = (X^*_{(2l+1)q+1}, ..., X^*_{(2l+2)q})$ such that for all $l = 0, ..., p-1$:*

- *A_l^* and A_l have the same law,*
- *A_l^* is independent of $A_0, ..., A_{l-1}, A_0^* ..., A_{l-1}^*$*
- *$\mathbb{E}(|A_l - A_l^*|_q) \leq q\tau_q$*

the same being true for the variables B_l.

Proof of Claim 1 :

We use the same recursive construction as Viennet [70].
Let $(\delta_j)_{0\leq j\leq p-1}$ be a sequence of independent random variables uniformly distributed over $[0,1]$ and independent of the sequence $(A_j)_{0\leq j\leq p-1}$. Let $A_0^*=(X_1^*,...,X_q^*)$ be the random variable given by equality (3.4) for $\mathcal{M}=\sigma(X_i,\ i\leq -q)$, A_0 and δ_0. Now suppose that we have built the variables A_l^* for $l<l'$. From equality (3.4) applied to the σ-algebra $\sigma(A_l,A_l^*,\ l<l')$, $A_{l'}$ and $\delta_{l'}$, there exists a random variable $A_{l'}^*$ satisfying the hypotheses of Claim 1.
We build in the same way the variables B_l^* for all $l=0,...,p-1$. □

We keep the notations $\nu_n^*,\bar{\nu}_{A,p},\bar{\nu}_{B,p}$, $\bar{t}$ and $B_1(S_m)$ that we introduced in the proof of Theorem 3.3.1. As in the proof of Theorem 3.3.1, we assume that, for some $\kappa>2$, $\sqrt{n}(\log n)^\kappa/2\leq p\leq\sqrt{n}(\log n)^\kappa$ and for the sake of simplicity that $pq=n/2$, the modifications needed to handle the extra term when $q=[n/(2p)]$ being straightforward. We have

$$V(\hat{m}) = \sum_{(j,k)\in\hat{m}}\nu_n^2(\psi_{j,k})\leq 2\sum_{(j,k)\in\hat{m}}(P_n-P_n^*)^2(\psi_{j,k})+2\sum_{(j,k)\in\hat{m}}(\nu_n^*)^2(\psi_{j,k})\tag{3.33}$$

Claim 2 : *There exists a constant $L=L_{A,K_L,K_\infty,\kappa,\theta}$ such that*

$$\mathbb{E}\left(\sum_{j,k\in\hat{m}}((P_n-P_n^*)(\psi_{j,k}))^2\right)\leq L\frac{(\log n)^{\kappa(\theta+1)}}{n^{(\theta-3)/2}}.\tag{3.34}$$

Proof of Claim 2 :

$$\begin{aligned}\mathbb{E}\left(\sum_{(j,k)\in\hat{m}}(P_n-P_n^*)^2(\psi_{j,k})\right) &\leq \mathbb{E}\left(\sup_{m\in\mathcal{M}_n}\sum_{(j,k)\in m}(P_n-P_n^*)^2(\psi_{j,k})\right)\\ &\leq \sum_{m\in\mathcal{M}_n}\sum_{(j,k)\in m}\mathbb{E}\left((P_n-P_n^*)^2(\psi_{j,k})\right)\\ &\leq \frac{2}{p^2}\sum_{m\in\mathcal{M}_n}\sum_{l,l'=1}^{p}(g_{A,m}(j,k,l,l')+g_{B,m}(j,k,l,l'))\end{aligned}$$

with

$$g_{m,A}(j,k,l,l')=\mathbb{E}\left(\sum_{(j,k)\in m}\left(\bar{\psi}_{j,k}(A_l)-\bar{\psi}_{j,k}(A_l^*)\right)\left(\bar{\psi}_{j,k}(A_{l'})-\bar{\psi}_{j,k}(A_{l'}^*)\right)\right).$$

We develop this last term and we get, since

$$\left|\bar{\psi}_{j,k}(x)-\bar{\psi}_{j,k}(y)\right|\leq\frac{K_L2^{3j/2}\left|x-y\right|_q}{2q}$$

$$\begin{aligned}
g_{A,m}(j,k,l,l') &\leq \mathbb{E}\left(\sum_{(j,k)\in m}\left|\bar{\psi}_{j,k}(A_l)-\bar{\psi}_{j,k}(A_l^*)\right|\left|\bar{\psi}_{j,k}(A_{l'})-\bar{\psi}_{j,k}(A_{l'}^*)\right|\right)\\
&\leq \mathbb{E}\left(\sum_{(j,k)\in m}\left|\bar{\psi}_{j,k}(A_l)-\bar{\psi}_{j,k}(A_l^*)\right| K_L 2^{3j/2}\frac{\left|A_{l'}-A_{l'}^*\right|_q}{2q}\right)\\
&\leq \frac{K_L\tau_q}{2}\sup_{x,y\in\mathbb{R}^q}\left\{\sum_{(j,k)\in m}2^{3j/2}\left|\bar{\psi}_{j,k}(x)-\bar{\psi}_{j,k}(y)\right|\right\}\\
&\leq \frac{K_L\tau_q}{4}\sum_{j=0}^{J_m}2^{3j/2}\sup_{x,y\in\mathbb{R}}\left\{\sum_{k\in\mathbb{Z}}\left|\psi_{j,k}(x)-\psi_{j,k}(y)\right|\right\}\\
&\leq \frac{2}{3}AK_LK_\infty 2^{2J_m}\tau_q \text{ since } \left\|\sum_{k\in\mathbb{Z}}|\psi_{j,k}|\right\|_\infty \leq AK_\infty 2^{j/2}
\end{aligned}$$

We can do the same computations for the term $g_{B,m}(j,k,l,l')$ and we obtain

$$\mathbb{E}\left(\sum_{j,k\in\hat{m}}\left((P_n-P_n^*)(\psi_{j,k})\right)^2\right)\leq L\tau_q\sum_{m\in\mathcal{M}_n}2^{2J_m}\leq L\tau_q 2^{2J_n}\leq L\frac{(\log n)^{\kappa(\theta+1)}}{n^{(\theta-3)/2}}.$$

The last inequality comes from $q\geq\sqrt{n}/(2(\log n)^\kappa)$ and Assumption [**AR**], the one before comes from Assumption [**W**]. □

Claim 3. *Let us keep the notations of Theorem 3.4.1, let $u=6/(7+\theta)<1/2$ and recall that $\kappa>2$. Let γ be a real number in $(1,\kappa/2)$. Let*

$$L_1^2=AK_\infty K_{BV}\sum_{l=0}^{\infty}\tilde{\beta}_l,\ L_2^2=2\Phi K_{BV}^u\sum_{k=0}^{\infty}\tilde{\beta}_k^u,\ L_3=\kappa(\epsilon)\Phi$$

$$\textit{and } L_{1,m}=4(1+\epsilon)\left((1+\epsilon)L_1+L_2\sqrt{\frac{(\log D_m)^\gamma}{D_m^{1/2-u}}}+L_3\frac{(\log D_m)^\gamma}{(\log n)^\kappa}\right)^2, \tag{3.35}$$

There exists a constant L_s such that

$$\mathbb{E}\left(\sup_{m\in\mathcal{M}_n}\left\{\sum_{(j,k)\in m}(\nu_n^*)^2(\psi_{j,k})-\frac{L_{1,m}D_m}{n}\right\}\right)\leq\frac{L_s}{n}.$$

Remark : The series $\sum_{l=0}^\infty\tilde{\beta}_l$ and $\sum_{k=0}^\infty\tilde{\beta}_k^u$ are convergent under our hypotheses on the coefficients τ. Since $s\in L^2([0,1])$, we have from Inequality (3.6), $\tilde{\beta}_l\leq 2\|s\|_2^{2/3}\tau_l^{1/3}$ and thus $\tilde{\beta}_l\leq 2\|s\|_2^{2/3}(1+l)^{-(1+\theta)/3}$. The series $\sum_{k=0}^\infty\tilde{\beta}_k^u$ converge since $\theta>5$ and

$$\frac{u(1+\theta)}{3}=\frac{2(1+\theta)}{7+\theta}=1+\frac{\theta-5}{\theta+7}>1.$$

We use here $\tilde{\beta}$ instead of τ which allows to take L_1 not depending on $\|s\|_2$.

Proof of Claim 3 :

As in the previous section we use the following decomposition

$$\begin{aligned}\sum_{(j,k)\in m}(\nu_n^*)^2(\psi_{j,k}) &= \sum_{(j,k)\in m}\left(\bar{\nu}_{A,p}(\bar{\psi}_{j,k})+\bar{\nu}_{B,p}(\bar{\psi}_{j,k})\right)^2\\ &\leq 2\sum_{(j,k)\in m}\left(\bar{\nu}_{A,p}(\bar{\psi}_{j,k})\right)^2+2\sum_{(j,k)\in m}\left(\bar{\nu}_{B,p}(\bar{\psi}_{j,k})\right)^2\end{aligned}$$

We treat both terms with Proposition 3.7.4 applied to the random variables $(A_l^*)_{0=1,..,p-1}$ and $(B_l^*)_{l=0,..,p-1}$ and to the class of functions $\left\{(\bar{\psi}_{j,k})_{(j,k)\in m}\right\}$. Let

$$B_m^2=\sum_{(j,k)\in m}\mathrm{Var}\left(\bar{\psi}_{j,k}(A_1)\right),\ V_m^2=\sup_{t\in B_1(S_m)}\mathrm{Var}(\bar{t}(A_1)),\ H_m^2=\|\sum_{(j,k)\in m}\bar{\psi}_{j,k}^2\|_\infty.$$

We have, from Proposition 3.7.4

$$\forall x>0,\ \mathbb{P}\left[\sqrt{\sum_{(j,k)\in m}(\bar{\nu}_{A,p})^2\bar{\psi}_{j,k}}\geq\frac{(1+\epsilon)}{\sqrt{p}}B_m+V_m\sqrt{\frac{2x}{p}}+\kappa(\epsilon)\frac{H_mx}{p}\right]\leq e^{-x}. \tag{3.36}$$

Let us now evaluate B_m, V_m and H_m, we have

$$B_m^2=\frac{1}{(2q)^2}\sum_{(j,k)\in m}\mathrm{Var}\left(\sum_{i=1}^q\psi_{j,k}(X_i)\right).$$

From (3.17) and (3.15) we have $\forall j,k\ \|\psi_{j,k}\|_{BV}\leq K_{BV}2^{j/2}$ and $\forall j\ \|\sum_{k\in\mathbb{Z}}|\psi_{j,k}|\|_\infty\leq AK_\infty 2^{j/2}$. Thus, from Inequality (3.5)

$$\begin{aligned}\sum_{(j,k)\in m}\mathrm{Var}\left(\sum_{i=1}^q\psi_{j,k}(X_i)\right) &\leq 2\sum_{(j,k)\in m}\sum_{l=1}^q(q+1-l)|\mathrm{Cov}(\psi_{j,k}(X_1),\psi_{j,k}(X_l))|\\ &\leq 2q\sum_{j=0}^{J_m}\sum_{k\in\mathbb{Z}}\sum_{l=1}^q\|\psi_{j,k}\|_{BV}\,\mathbb{E}\left(|\psi_{j,k}(X_1)|b(\sigma(X_1),X_l)\right)\\ &\leq 2K_{BV}q\sum_{j=0}^{J_m}2^{j/2}\left\|\sum_{k\in\mathbb{Z}}|\psi_{j,k}(X_0)|\right\|_\infty\sum_{l=1}^q\tilde{\beta}_{l-1}\\ &\leq 2q\left(AK_\infty K_{BV}\sum_{l=0}^\infty\tilde{\beta}_l\right)D_m.\end{aligned}$$

The last inequality comes from Assumption [**W**].
Since $L_1^2=AK_\infty K_{BV}\sum_{l=0}^\infty\tilde{\beta}_l$ we have

$$B_m^2\leq\frac{L_1^2D_m}{2q}. \tag{3.37}$$

Let us deal with the term V_m^2. We have

$$V_m^2\leq\sup_{t\in B_1(S_m)}\mathrm{Var}(\bar{t}(A_1))\leq\frac{2}{(2q)^2}\sum_{k=1}^q(q+1-k)\sup_{t\in B_1(S_m)}|\mathrm{Cov}(t(X_1),t(X_k))| \tag{3.38}$$

From Inequality (3.5), we have

$$|\mathrm{Cov}(t(X_1),t(X_k))| \leq \|t\|_{BV} \, \|t\|_\infty \, \tilde{\beta}_{k-1}.$$

Since t belongs to $B_1(S_m)$, we have $t = \sum_{(j,k)\in m} a_{j,k}\psi_{j,k}$, with $\sum_{(j,k)\in m} a_{j,k}^2 \leq 1$. Thus, by Cauchy-Schwarz inequality

$$\begin{aligned}\sum_{i=1}^{l} |t(x_{i+1})-t(x_i)| &\leq \sum_{(j,k)\in m} |a_{j,k}| \sum_{i=1}^{l} |\psi_{j,k}(x_{i+1})-\psi_{j,k}(x_i)| \\ &\leq \left(\sum_{(j,k)\in m} a_{j,k}^2\right)^{1/2} \left(\sum_{(j,k)\in m}\left(\sum_{i} |\psi_{j,k}(x_{i+1})-\psi_{j,k}(x_i)|\right)^2\right)^{1/2} \\ &\leq \left(\sum_{(j,k)\in m} \|\psi_{j,k}\|_{BV}^2\right)^{1/2} \leq K_{BV} D_m.\end{aligned}$$

Thus $\|t\|_{BV} \leq D_m K_{BV}$. From Assumption $[M_2]$, we have $\|t\|_\infty \leq \Phi\sqrt{D_m}$. Thus

$$|\mathrm{Cov}(t(X_1),t(X_k))| \leq \Phi K_{BV} \tilde{\beta}_{k-1} D_m^{3/2}. \tag{3.39}$$

Moreover, we have by Cauchy-Schwarz inequality and $[M_2]$

$$|\mathrm{Cov}(t(X_1),t(X_k))| \leq \|t\|_\infty \, \|t\|_2 \, \|s\|_2 \leq \Phi \, \|s\|_2 \sqrt{D_m}. \tag{3.40}$$

We use the inequality $a \wedge b \leq a^u b^{1-u}$ with

$$a = \Phi K_{BV}\tilde{\beta}_{k-1} D_m^{3/2}, \; b = \Phi \, \|s\|_2 \sqrt{D_m}, \; u = \frac{6}{7+\theta} < \frac{1}{2}.$$

From (3.39) and (3.40), we derive that

$$|\mathrm{Cov}(t(X_1),t(X_k))| \leq L_k' D_m^{1/2+u} \text{ where } L_k' = \Phi\left(K_{BV}\tilde{\beta}_{k-1}\right)^u \|s\|_2^{1-u}.$$

Pluging this inequality in (3.38), we obtain

$$V_m^2 \leq \frac{L_2^2 \, \|s\|_2^{1-u} \, D_m^{1/2+u}}{4q} \text{ since } L_2^2 = 2\Phi K_{BV}^u \sum_{k=0}^{\infty} \tilde{\beta}_k^u. \tag{3.41}$$

Finally, we have from hypothesis $[M_2]$

$$H_m^2 \leq \frac{1}{4}\left\|\sum_{(j,k)\in m} \psi_{j,k}^2\right\|_\infty \leq \frac{\Phi^2 D_m}{4}. \tag{3.42}$$

Let $y > 0$ and let us apply Inequality (3.36) with $x = ((\log D_m)^\gamma / \|s\|_2^{1-u}) + (y/D_m^{1/2+u})$. We have, from (3.37), (3.41) and (3.42)

$$\begin{aligned}\mathbb{P}\Bigg[\sum_{(j,k)\in m} (\bar{\nu}_{A,p})^2(\bar{\psi}_{j,k}) > \Bigg((1+\epsilon)\sqrt{\frac{L_1^2 D_m}{2pq}} + \frac{L_3\sqrt{D_m}}{2p}\left(\frac{(\log D_m)^\gamma}{\|s\|_2^{1-u}} + \frac{y}{D_m^{1/2+u}}\right) \\ + \sqrt{\frac{L_2^2\|s\|_2^{1-u} D_m^{1/2+u}}{2pq}\left(\frac{(\log D_m)^\gamma}{\|s\|_2^{1-u}} + \frac{y}{D_m^{1/2+u}}\right)}\Bigg)^2\Bigg] \leq e^{-\frac{(\log D_m)^\gamma}{\|s\|_2^{1-u}}} e^{-D_m^{-(1/2+u)} y}.\end{aligned}$$

Then, we use the inequality $\sqrt{\alpha+\beta} \leq \sqrt{\alpha}+\sqrt{\beta}$ with

$$\alpha = \frac{(\log D_m)^\gamma}{\|s\|_2^{1-u}} \text{ and } \beta = \frac{y}{D_m^{1/2+u}}$$

and the inequality $(a+b)^2 \leq (1+\epsilon)a^2 + (1+\epsilon^{-1})b^2$ with

$$a = \left((1+\epsilon)L_1 + L_2\sqrt{\frac{(\log D_m)^\gamma}{D_m^{1/2-u}}} + \frac{L_3(\log D_m)^\gamma}{\|s\|_2^{1-u}(\log n)^\kappa}\right)\sqrt{\frac{D_m}{n}}$$

$$\text{and } b = \frac{1}{\sqrt{n}}\left(L_2\sqrt{\|s\|_2^{1-u}y} + \frac{L_3 y}{(\log n)^\kappa D_m^u}\right).$$

Setting $L_m = (1+\epsilon)a^2 n/D_m$, we obtain

$$\mathbb{P}\left(\sum_{(j,k)\in m}(\bar{\nu}_{A,p})^2(\bar{\psi}_{j,k}) - \frac{L_m D_m}{n} > \frac{(1+\epsilon^{-1})}{n}\left(L_2\sqrt{\|s\|_2^{1-u}y} + \frac{L_3 y}{(\log n)^\kappa D_m^u}\right)^2\right)$$
$$\leq e^{-\frac{(\log D_m)^\gamma}{\|s\|_2^{1-u}}} e^{-D_m^{-(1/2+u)}y}.$$

Thus, for all $y > 0$,

$$\mathbb{P}\left(\sup_{m\in\mathcal{M}_n}\left\{\sum_{(j,k)\in m}(\bar{\nu}_{A,p})^2(\bar{\psi}_{j,k}) - \frac{L_m D_m}{n}\right\} > \frac{L_s}{n}(y+y^2)\right) \leq \sum_{m\in\mathcal{M}_n} e^{-\frac{(\log D_m)^\gamma}{\|s\|_2^{1-u}} - D_m^{-(1/2+u)}y}$$

where $L_s = 2(1+\epsilon^{-1})\left[(L_2\sqrt{\|s\|_2^{1-u}}) \vee L_3/((\log 2)^\kappa 2^u)\right]^2$. We can integrate this last inequality to prove Claim 3.□

Claim 4 : *We keep the notations of the previous Claims. Let*

$$L_2(m,m') = 4\left(L_2\sqrt{\frac{(\log(D_m \vee D_{m'}))^\gamma}{(D_m \vee D_{m'})^{1/2-u}}} + \frac{\Phi}{3(\log n)^{\kappa-\gamma}}\right)^2. \tag{3.43}$$

Then there exists a constant $L_{s,\theta}$ depending on $\|s\|_2$ and θ such that, for all $\eta > 0$

$$\mathbb{E}\left(\sup_{m,m'\in\mathcal{M}_n}\left\{\nu_n(s_m - s_{m'}) - \frac{\|s_m - s_{m'}\|_2^2}{2\eta} - \eta\frac{L_2(m,m')(D_m \vee D_{m'})}{n}\right\}\right) \leq \frac{\eta L_{s,\theta}}{n}.$$

Proof of Claim 4 :

$$\mathbb{E}\left(\sup_{m,m'\in\mathcal{M}_n}\left\{\nu_n(s_m - s_{m'}) - \frac{\|s_m - s_{m'}\|_2^2}{2\eta} - \eta\frac{L_2(m,m')(D_m \vee D_{m'})}{n}\right\}\right)$$
$$\leq \mathbb{E}\left(\sup_{m,m'}(P_n - P_n^*)(s_m - s_{m'})\right)$$
$$+\mathbb{E}\left(\sup_{m,m'}\left\{\nu_n^*(s_m - s_{m'}) - \frac{\|s_m - s_{m'}\|_2^2}{2\eta} - \eta\frac{L_2(m,m')(D_m \vee D_{m'})}{n}\right\}\right) \tag{3.44}$$

Since $\forall l = 0, ..., p-1$, $\mathbb{E}\left(|A_l - A_l^*|_q\right) \leq q\tau_q$, we have

$$\begin{aligned}\mathbb{E}\left(\sup_{m,m'}(P_n - P_n^*)(s_m - s_{m'})\right) &\leq 2\sum_{m,m'}\mathbb{E}\left(|(\bar{s}_m - \bar{s}_{m'})(A_1) - (\bar{s}_m - \bar{s}_{m'})(A_1^*)|\right)\\ &\leq \tau_q \sum_{m,m'} \mathrm{Lip}(s_m - s_{m'}).\end{aligned}$$

When $m \subset m'$, we have, for all $x, y \in \mathbb{R}$, using Assumption **[W]**,

$$\frac{|(s_m - s_{m'})(x-y)|}{|x-y|} \leq \sum_{j=J_m+1}^{J_{m'}} \sum_{k=-A_2}^{2^j - A_1} |P\psi_{j,k}| \frac{|\psi_{j,k}(x) - \psi_{j,k}(y)|}{|x-y|}$$

Let us fix $j \in [J_m+1, J_{m'}]$, from Assumption **[W]**, there is less than A indexes $k \in \mathbb{Z}$ such that $\psi_{j,k}(x) \neq 0$, thus there is less than $2A$ indexes such that $|\psi_{j,k}(x) - \psi_{j,k}(y)| \neq 0$. Hence

$$\begin{aligned}\sum_{k\in\mathbb{Z}} |P\psi_{j,k}| \frac{|\psi_{j,k}(x) - \psi_{j,k}(y)|}{|x-y|} &\leq 2A \sup_{k\in\mathbb{Z}} |P\psi_{j,k}| \mathrm{Lip}(\psi_{j,k})\\ &\leq 2A \left\|s\right\|_2 K_L 2^{3j/2}.\end{aligned}$$

Thus, $\mathrm{Lip}(s_m - s_{m'}) \leq A \left\|s\right\|_2 K_L \sqrt{8} 2^{3J_{m'}/2}/(\sqrt{8}-1)$ and by Assumptions **[W]**, **[AR]** and the value of q,

$$\mathbb{E}\left(\sup_{m,m'}(P_n - P_n^*)(s_m - s_{m'})\right) \leq L_s n^{3/2}(\log n)\tau_q \leq L_s \frac{(\log n)^{\kappa(\theta+1)+1}}{n^{(\theta-2)/2}}. \tag{3.45}$$

Let us deal with the other term in (3.44). We have, $\forall \eta > 0$

$$\begin{aligned}\nu_n^*(s_m - s_{m'}) &\leq \frac{\|s_m - s_{m'}\|_2^2}{2\eta} + \frac{\eta}{2}\left(\bar{\nu}_{A,p}(\bar{t}_{m,m'}) + \bar{\nu}_{B,p}(\bar{t}_{m,m'})\right)^2\\ &\leq \frac{\|s_m - s_{m'}\|_2^2}{2\eta} + \eta(\bar{\nu}_{A,p}(\bar{t}_{m,m'}))^2 + \eta(\bar{\nu}_{B,p}(\bar{t}_{m,m'}))^2 \qquad (3.46)\end{aligned}$$

where, as in the proof of Theorem 3.3.1, $t_{m,m'} = (s_m - s_{m'})/\|s_m - s_{m'}\|_2$. We apply Bernstein's inequality to the function $\bar{t}_{m,m'}$ and the variables A_l^*, we have

$$\forall x > 0,\ \mathbb{P}\left(\bar{\nu}_{A,p}(\bar{t}_{m,m'}) > \sqrt{\frac{2\mathrm{Var}(\bar{t}_{m,m'}(A_0))x}{p}} + \frac{\|\bar{t}_{m,m'}\|_\infty x}{3p}\right) \leq e^{-x}. \tag{3.47}$$

We proceed as in the proof of Claim 3 to control this variance. We have, by stationarity of the process $(X_n)_{n\in\mathbb{Z}}$,

$$\mathrm{Var}(\bar{t}_{m,m'}(A_0)) = \frac{1}{2q^2}\sum_{k=0}^{q-1}(q-k)\mathrm{Cov}(t_{m,m'}(X_1), t_{m,m'}(X_{k+1})).$$

From Inequality (3.5), we have

$$|\mathrm{Cov}(t_{m,m'}(X_1), t_{m,m'}(X_{k+1}))| \leq \|t_{m,m'}\|_{BV} \|t_{m,m'}\|_\infty \tilde{\beta}_k.$$

Let $m \triangle m'$ be the set of indexes that belong to $m \cup m'$ but do not belong to $m \cap m'$. We use the same computations as in the proof of Claim 3 to get

$$\|t_{m,m'}\|_{BV} \leq \frac{\left\|\sum_{(j,k)\in m'\triangle m}(P\psi_{j,k})\psi_{j,k}\right\|_{BV}}{\|s_m - s_{m'}\|_2} \leq \sqrt{\sum_{(j,k)\in m'\triangle m} \|\psi_{j,k}\|_{BV}^2} \leq K_{BV}(D_m \vee D_{m'}).$$

Since $\|t_{m,m'}\|_\infty = \Phi\sqrt{D_m \vee D_{m'}}$, we have

$$|\mathrm{Cov}(t_{m,m'}(X_1), t_{m,m'}(X_{k+1}))| \leq \Phi K_{BV}\tilde{\beta}_k(D_m \vee D_{m'})^{3/2}. \tag{3.48}$$

Moreover, we have

$$\mathrm{Cov}(t_{m,m'}(X_1), t_{m,m'}(X_{k+1})) \leq \|t_{m,m'}\|_\infty \|t_{m,m'}\|_2 \|s\|_2 \leq \Phi \|s\|_2 \sqrt{(D_m \vee D_m')}. \tag{3.49}$$

Thus, using $a \wedge b \leq a^u b^{1-u}$ with

$$a = \Phi K_{BV}\tilde{\beta}_k(D_m \vee D_{m'})^{3/2},\ b = \Phi \|s\|_2 \sqrt{(D_m \vee D_{m'})},\ \text{and } u = \frac{6}{7+\theta} < \frac{1}{2},$$

we have

$$|\mathrm{Cov}(t_{m,m'}(X_1), t_{m,m'}(X_{k+1}))| \leq \Phi K_{BV}^u \tilde{\beta}_k^u \|s\|_2^{1-u} (D_m \vee D_{m'})^{1/2+u}.$$

Thus

$$\mathrm{Var}(\bar{t}_{m,m'}(A_0)) \leq \Phi K_{BV}^u \left(\sum_{k=0}^{\infty} \tilde{\beta}_k^u\right) \|s\|_2^{1-u} \frac{(D_m \vee D_{m'})^{1/2+u}}{2q}. \tag{3.50}$$

Moreover

$$\|\bar{t}_{m,m'}\|_\infty \leq \frac{1}{2}\|t_{m,m'}\|_\infty \leq \frac{1}{2}\Phi\sqrt{D_m \vee D_m'}. \tag{3.51}$$

Now, we use (3.47) with $x = (\log(D_m \vee D_{m'}))^\gamma / \|s\|_2^{1-u} + y/(D_m \vee D_{m'})^{1/2+u}$. From (3.50) and (3.51), we have for all $y > 0$,

$$\mathbb{P}\left(\bar{\nu}_{A,p}(\bar{t}_{m,m'}) > L_2\sqrt{\frac{(D_m \vee D_{m'})^{1/2+u}}{2pq}\left((\log(D_m \vee D_{m'}))^\gamma + \frac{\|s\|_2^{1-u} y}{(D_m \vee D_{m'})^{1/2+u}}\right)} \right.$$
$$\left. + \frac{\Phi\sqrt{D_m \vee D_m'}}{6p}\left(\frac{(\log(D_m \vee D_{m'}))^\gamma}{\|s\|_2^{1-u}} + \frac{y}{(D_m \vee D_{m'})^{1/2+u}}\right)\right)$$
$$\leq e^{-\frac{(\log(D_m \vee D_{m'}))^\gamma}{\|s\|_2^{1-u}}} e^{-\frac{y}{(D_m \vee D_{m'})^{1/2+u}}}.$$

Now we use the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ with

$$a = (\log(D_m \vee D_{m'}))^\gamma \text{ and } b = \frac{\|s\|_2^{1-u} y}{(D_m \vee D_{m'})^{1/2+u}}$$

and we obtain, using Assumption $[M_1]$

$$\mathbb{P}\left(\bar{\nu}_{A,p}\bar{t}_{m,m'} - \sqrt{\frac{L_2(m,m')(D_m \vee D_m')}{n}} > \frac{L_s}{\sqrt{n}}(\sqrt{y} + y)\right)$$
$$\leq e^{-\frac{(\log(D_m \vee D_{m'}))^\gamma}{\|s\|_2^{1-u}}} e^{-(D_m \vee D_{m'})^{-(1/2+u)} y},$$

with

$$L_2(m,m') = \left(L_2 \sqrt{\frac{(\log(D_m \vee D_{m'}))^\gamma}{(D_m \vee D_{m'})^{1/2-u}}} + \frac{\Phi(\log(D_m \vee D_{m'}))^\gamma}{3(\log n)^\kappa} \right)^2,$$

$$\text{and } L_s = L_2 \sqrt{\|s\|_2^{1-u}} \vee \frac{\Phi}{3(\log 2)^\kappa 2^u}.$$

Thus, we obain

$$\mathbb{P}\left((\bar{\nu}_{A,p} \bar{t}_{m,m'})^2 > 2\frac{L_2(m,m')(D_m \vee D'_m)}{n} + 4\frac{L_s^2}{n}(y+y^2) \right)$$
$$\leq e^{-\frac{(\log(D_m \vee D_{m'}))^\gamma}{\|s\|_2^{1-u}} - \frac{y}{(D_m \vee D_{m'})^{1/2+u}}}.$$

The same result holds for $\bar{\nu}_{B,p}\bar{t}_{m,m'}$. Thus we obtain from (3.46)

$$\mathbb{P}\left(\nu_n^*(s_m - s_{m'}) \geq \frac{\|s_m - s_{m'}\|_2^2}{2\eta} + 4\eta \frac{L_2(m,m')(D_m \vee D'_m)}{n} + 8\eta \frac{L_s^2}{n}(y+y^2) \right)$$
$$\leq 2e^{-\frac{(\log(D_m \vee D_{m'}))^\gamma}{\|s\|_2^{1-u}} - \frac{y}{(D_m \vee D_{m'})^{1/2+u}}}.$$

We deduce that

$$\mathbb{P}\left(\exists m, m' \in \mathcal{M}_n,\ \nu_n^*(s_m - s_{m'}) - \frac{\|s_m - s_{m'}\|_2^2}{2\eta} - 4\eta \frac{L_2(m,m')(D_m \vee D'_m)}{n} \right.$$
$$\left. \geq 8\eta \frac{L_s^2}{n}(y+y^2) \right) \leq 2 \sum_{m,m' \in \mathcal{M}_n} \left(e^{-\frac{(\log(D_m \vee D_{m'}))^\gamma}{\|s\|_2^{1-u}}} \right) e^{-\frac{y}{(D_m \vee D_{m'})^{1/2+u}}}.$$

We integrate this last inequality to get Claim 4.□

Conclusion of the proof:
Take

$$\mathrm{pen}'(m) \geq (2L_{1,m} + \eta L_2(m,m))\frac{D_m}{n},$$

where $L_{1,m}$ and $L_2(m,m)$ are defined by (3.35) and (3.43) respectively. From Claims 2, 3 and 4, if we take the expectation in (3.21), we have, for some constant L_s,

$$\mathbb{E}\left(\|s - \tilde{s}\|_2^2\right) \leq \mathbb{E}\left(\|s - \hat{s}_{m_o}\|_2^2 + \mathrm{pen}'(m_o) - V(m_o) + 2\eta L_2(m_o, m_o)\frac{D_{m_o}}{n} \right) + \frac{\eta L_s}{n}. \tag{3.52}$$

Moreover, if $D_m \geq \left((L_2/L_1)(\log n)^{\kappa - \gamma/2}\right)^{2(7+\theta)/(\theta-5)}$, we have

$$\begin{aligned} \frac{L_{1,m}}{4L_1^2} &\leq (1+\epsilon)\left((1+\epsilon) + \left(1 + \frac{L_3}{2L_1}\right)(\log n)^{-(\kappa-\gamma)} \right)^2 \\ &\leq (1+\epsilon)^3 + (1+\epsilon^{-1})(1+\epsilon)\left(1 + \frac{L_3}{2L_1}\right)^2 (\log n)^{-2(\kappa-\gamma)}. \end{aligned} \tag{3.53}$$

We use the inequality $(a+b)^2 \leq (1+\epsilon)a^2 + (1+\epsilon^{-1})b^2$ to obtain (3.53). Moreover, we have

$$L_2(m,m) \leq 4L_1^2 \left(\left(1 + \frac{\Phi}{6L_1}\right)(\log n)^{-(\kappa-\gamma)} \right)^2.$$

As in the proof of Theorem 3.3.1, we take $\eta = (\log n)^{\kappa-\gamma}$ and we fix ϵ sufficiently small. For $n \geq n_o$, we have $2L_{1,m} + \eta L_2(m,m) < KL_1^2$. Thus inequality (3.18) follows from (3.52).□

3.7 Appendix

This section is devoted to technical lemmas that are needed in the proofs.

3.7.1 Covariance inequality

Lemma 3.7.1 *Viennet's inequality Let $(X_n)_{n\in\mathbb{Z}}$ be a stationary and β-mixing process. There exists a positive function b such that $P(b) \leq \sum_{l=0}^{\infty} \beta_l$, $P(b^p) \leq p\sum_{l=1}^{\infty} l^{p-1}\beta_l$, and for all function $h \in L_2(P)$*

$$Var\left(\sum_{l=1}^{q} h(X_l)\right) \leq 4qP(bh^2). \tag{3.54}$$

3.7.2 Concentration inequalities

We sum up in this section the concentration inequalities we used in the proofs. We begin with Bernstein's inequality

Proposition 3.7.2 *Bernstein's inequality*
Let $X_1, ..., X_n$ be iid random variables valued in a measurable space $(X, \mathcal{X})$ and let t be a measurable real valued function. Let $v = Var(t(X_1))$ and $b = \|t\|_\infty$, then, for all $x > 0$, we have

$$\mathbb{P}\left((P_n - P)t > v\sqrt{\frac{2x}{n}} + \frac{bx}{3n}\right) \leq e^{-x}.$$

Now we give the most important tool of our proof, it is a concentration's inequality for the supremum of the empirical process over a class of function. We give here the version of Bousquet [18].

Theorem 3.7.3 *Talagrand's Theorem*
Let $X_1, ..., X_n$ be i.i.d random variables valued in some measurable space $[X, \mathcal{X}]$. Let $\mathcal{F}$ be a separable class of bounded functions from X to $\mathbb{R}$ and assume that all functions t in $\mathcal{F}$ are P-measurable, and satisfy $Var(t(X_1)) \leq \sigma^2$, $\|t\|_\infty \leq b$. Then

$$\mathbb{P}\left(\sup_{t\in\mathcal{F}} \nu_n(t) > \mathbb{E}\left(\sup_{t\in\mathcal{F}} \nu_n(t)\right) + \sqrt{\frac{2x\left(\sigma^2 + 2b\mathbb{E}\left(\sup_{t\in\mathcal{F}} \nu_n(t)\right)\right)}{n}} + \frac{bx}{3n}\right) \leq e^{-x}.$$

In particular, for all $\epsilon > 0$, if $\kappa(\epsilon) = 1/3 + \epsilon^{-1}$, we have

$$\mathbb{P}\left(\sup_{t\in\mathcal{F}} \nu_n(t) > (1+\epsilon)\mathbb{E}\left(\sup_{t\in\mathcal{F}} \nu_n(t)\right) + \sigma\sqrt{\frac{2x}{n}} + \kappa(\epsilon)\frac{bx}{n}\right) \leq e^{-x}.$$

We can deduce from this Theorem a concentration's inequality for χ-square type statistics. This is Proposition (7.3) of Massart [55].

Proposition 3.7.4 *Let $X_1, ..., X_n$ be independent and identically distributed random variables valued in some measurable space $(X, \mathcal{X})$. Let P denote their common distribution. Let ϕ_λ be a finite family of measurable and bounded functions on $(X, \mathcal{X})$. Let*

$$H_\Lambda^2 = \|\sum_{\lambda\in\Lambda} \phi_\lambda^2\|_\infty \text{ and } B_\Lambda^2 = \sum_{\lambda\in\Lambda} Var(\phi_\lambda(X_1)).$$

Moreover, let $\mathcal{S}_\Lambda = \left\{a \in \mathbb{R}^\Lambda : \sum_{\lambda\in\Lambda} a_\lambda^2 = 1\right\}$ and

$$V_\Lambda^2 = \sup_{a\in\mathcal{S}_\Lambda} \left\{ Var\left(\sum_{\lambda\in\Lambda} a_\lambda \phi_\lambda(X_1)\right)\right\}.$$

Then the following inequality holds, for all positive x and ϵ

$$\mathbb{P}\left[\left(\sum_{\lambda\in\Lambda}(P_n - P)^2\phi_\lambda\right)^{1/2} \geq \frac{1+\epsilon}{\sqrt{n}}B_\Lambda + V_\Lambda\sqrt{\frac{2x}{n}} + \kappa(\epsilon)\frac{H_\Lambda x}{n}\right] \leq e^{-x}, \qquad (3.55)$$

where $\kappa(\epsilon) = \epsilon^{-1} + 1/3$.

Proof :

Following Massart [55] Proposition 7.3, we remark that, by Cauchy-Schwarz's inequality

$$\left(\sum_{\lambda\in\Lambda} \nu_n^2\phi_\lambda\right)^{1/2} = \sup_{a\in\mathcal{S}_\Lambda} \sum_{\lambda\in\Lambda} a_\lambda \nu_n \phi_\lambda = \sup_{a\in\mathcal{S}_\Lambda} \nu_n\left(\sum_{\lambda\in\Lambda} a_\lambda\phi_\lambda\right).$$

Thus the result follows by applying Talagrand's Theorem to the class of functions

$$\mathcal{F} = \left\{t = \sum_{\lambda\in\Lambda} a_\lambda\phi_\lambda;\ a \in \mathcal{S}_\Lambda\right\}.$$

Chapter 4

Optimal model selection for stationary, β and τ-mixing data.

Abstract:

We build penalized least-squares estimators of the marginal density of a stationary process, using the slope algorithm and resampling penalties. When the data are β or τ-mixing, these estimators satisfy oracle inequalities with leading constant asymptotically equal to 1.

Key words: Density estimation, optimal model selection, resampling methods, slope heuristic, weak dependence.
2000 Mathematics Subject Classification: 62G07, 62G09, 62M99.

4.1 Introduction

The history of statistical model selection goes back at least to Akaike [1], [2] and Mallows [53]. They proposed to select among a collection of parametric models the one which minimizes an empirical loss plus some penalty term proportional to the dimension of the models. Birgé & Massart [15] and Barron, Birgé & Massart [10] generalize this approach, making in particular the link between model selection and adaptive estimation. They proved that several estimation procedures as cross validation (Rudemo [62]) or hard thresholding (Donoho *et.al.* [27]) can be interpreted in terms of model selection.
More recently, Birgé & Massart [17], Arlot & Massart [7] and Arlot [4], [5] arised the problem of optimal model selection. Basically, the aim is to select an estimator satisfying an oracle inequality with leading constant asymptotically equal to 1.
Two totally data driven procedures are known to achieve this goal: the slope algorithm, introduced by Birgé & Massart [17] and the resampling penalties defined by Arlot [5]. Arlot & Massart [7] and Arlot [5] proved that these estimators are efficient to select the best histogram in a general regression framework. In Chapter 2, we proved that these procedures are also optimal in density estimation, when the data are independent.
There exists a lot of statistical frameworks where the data are not independent. The previous results may therefore not hold. Baraud *et.al.* [9] proved that penalties proportional to the dimension can also be used when the data are β-mixing (for a

definition of the coefficient β, see Rozanov & Volkonskii [71] or Section 4.2). They worked in a regression framework and Comte & Merlevède [23] extended the result to density estimation. In Chapter 3, we proved that the same penalties can also be used with τ-mixing data (the coefficient τ has been introduced by Dedecker & Prieur [25], see Section 4.2). The main problem of the algorithm proposed by Comte & Merlevède [23] is that the penalty term involves a constant depending on the mixing coefficients (both in the β and τ-mixing cases) which is typically unknown in practice.

As in the independent case, we prove that a resampling estimator catches the shape of the ideal penalty with great generality as it "learns" part of the mixing structure of the data (Künsch [44], Liu & Singh [51]). We will also prove that the slope algorithm can be used to calibrate in an optimal way the constant in front of the penalty term. The new penalization procedure is totally data driven.

Let us now explain more precisely the problem that we will consider.

4.1.1 Least-squares estimators

We study efficient penalized least-squares estimators in density estimation when the error is measured with the L^2-loss. We observe n identically distributed random variables $X_1, ..., X_n$, defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, valued in a measurable space $(\mathbb{X}, \mathcal{X})$, with common law P. We assume that a probability measure μ on $(\mathbb{X}, \mathcal{X})$ is given. We denote by $L^2(\mu)$ the Hilbert space of square integrable real valued functions defined on $\mathbb{X}$ and by $\|.\|$ the associated L^2-norm. The parameter of interest is the density s of P with respect to μ, we assume that it belongs to $L^2(\mu)$. For all function g in $L^1(P)$, we define

$$Pg = \int_{\mathbb{X}} gsd\mu = \mathbb{E}\left(g(X)\right),$$

where X is a copy of X_1, independent of $(X_1, ..., X_n)$. s minimizes the integrated contrast $PQ(t)$, where the contrast function $Q : L^2(\mu) \to L^1(P)$, is defined for all t in $L^2(\mu)$ by $Q(t) = \|t\|^2 - 2t$. The risk of an estimator $\hat{s}$ of s is measured with the L^2-loss, that is $\|s - \hat{s}\|^2$, which is random when $\hat{s}$ is.

The problem of density estimation is a problem of M-estimation. These problems are classically solved in two steps when the data are independent. First, we choose a "model" S_m close to the parameter s, which means that $\inf_{t\in S_m} \|s-t\|^2$ is "small". Define the empirical process P_n for all functions g in $L^1(P)$ by

$$P_n g = \frac{1}{n}\sum_{i=1}^{n} g(X_i).$$

We minimize over S_m the empirical version of the integrated contrast, that is, we choose

$$\hat{s}_m \in \arg\min_{t\in S_m} P_n Q(t).$$

When the data are mixing, the coupling method is a very powerful tool to extend the methods developed in the independent case. It can be summarized as follows.

Coupling method: Let $I_0, J_0, ..., I_{p-1}, J_{p-1}$ be a partition of $\{1, ..., n\}$ satisfying $q = \min_{k=0,...,p-1} \min(I_{k+1}) - \max(I_k) > 0$ (for a proper definition of this partition see Section 4.2). For all $k = 0, ..., p-1$, let $A_k = (X_l)_{l \in I_k}$. A coupling lemma associates to the sequence $(A_k)_{k=0,...,p-1}$ independent random variables (A_k^*) such that $\mathbb{E}\left(d(A_k, A_k^*)\right) \leq \gamma(q)$, where γ is the mixing coefficient of the data, d is a distance on $\mathbb{X}^{l_k}$. Let $I = \cup_{k=0}^{p-1} I_k$ and let P_A be the empirical process based on the data $(X_i,\ i \in I)$, that is $P_A = \sum_{i \in I} \delta_{X_i}/|I|$. To bound quantities of the form $F(P_n)$, built with the empirical process, we first use algebraic inequalities to obtain

$$F(P_n) \leq CF(P_A). \tag{4.1}$$

Then we have

$$F(P_A) \leq F(P_{A^*}) + |F(P_A) - F(P_{A^*})|.$$

We can now use the results available for independent random variables to bound $F(P_{A^*})$ and the mixing properties to bound $|F(P_A) - F(P_{A^*})|$.

Up to our knowledge, all the model selection procedures proposed for mixing data used the coupling methods. In this scheme, the bounds given on $F(P_n)$ are the same as those given for $F(P_A)$ and the only essentially suboptimal bound is the first one: $F(P_n) \leq CF(P_A)$. We want to extend the procedures developed in the independent case in Chapter 2 through the coupling method. As we are looking for optimal results, we will work with the process P_A instead of P_n, avoiding the lost (4.1). The counterpart of this choice is that we do not use all the data to build our estimator. In particular, the variance of an oracle built only with the variables $(X_i)_{i \in I}$ is bigger than the one of an oracle built with all the sample when the data are independent. However, we will see in Section 4.4 that our final estimator improves the previous procedures proposed in a mixing setting. Let us now define the least-squares estimators by $\hat{s}_{A,m} \in \arg\min_{t \in S_m} P_A Q(t)$. The minimization problem defining $\hat{s}_{A,m}$ can be computationaly untractable for general sets S_m, leading to untractable procedures in practice. However, in density estimation, it can be easily solved when S_m is a linear subspace of $L^2(\mu)$ since, for any orthonormal basis $(\psi_\lambda)_{\lambda \in m}$ of S_m, we have

$$\hat{s}_{A,m} = \sum_{\lambda \in m} (P_A \psi_\lambda) \psi_\lambda.$$

The risk of $\hat{s}_{A,m}$ is decomposed in the classical bias and variance terms thanks to Pythagoras relation. Let s_m be the orthogonal projection of s onto S_m, then

$$\|s - \hat{s}_{A,m}\|^2 = \|s - s_m\|^2 + \|s_m - \hat{s}_{A,m}\|^2. \tag{4.2}$$

The space S_m should be chosen in order to realize a trade-off between those quantities. Actually, when the complexity of S_m increases, the bias term $\|s - s_m\|^2$ decreases whereas the variance term $\|\hat{s}_{A,m} - s_m\|^2$ increases. In Chapter 2, we proved a concentration inequality for $\|s_m - \hat{s}_{A,m}\|^2$ around its expectation when the data are independent. It proves that $D_{A,m}^* = n\mathbb{E}(\|s_m - \hat{s}_{A,m}\|^2)$ is a natural complexity measure of S_m and, when the models S_m are sufficiently regular, we recovered that the dimension d_m of S_m has the same order as $D_{A,m}^*$. However, this is not true in general, because there exist simple models (histograms with a small d_m) where $D_{A,m}^* >> d_m$.

4.1.2 Model selection

The choice of a "good" model S_m is impossible without strong assumptions on s, for example that we have precise information on its regularity. However, if we only assume that s is regular, it is possible to choose a collection of models $(S_m)_{m\in\mathcal{M}_n}$ such that that one of them realizes an optimal trade-off (see for example Birgé & Massart [15] or Barron, Birgé & Massart [10]). Given the projection estimators $(\hat{s}_{A,m})_{m\in\mathcal{M}_n}$ associated to this collection, the aim is then to build an estimator $\hat{m}$ such that the final estimator, $\tilde{s}=\hat{s}_{A,\hat{m}}$ behaves almost as well as any model m_o in the set of oracles

$$\mathcal{M}_n^*=\{m_o\in\mathcal{M}_n,\ \|\hat{s}_{A,m_o}-s\|^2=\inf_{m\in\mathcal{M}_n}\|\hat{s}_{A,m}-s\|^2\}.$$

This is the problem of model selection. More precisely, we want the final estimator $\tilde{s}=\hat{s}_{A,\hat{m}}$ to satisfy one of the following type of oracle inequalities

$$\exists K>0, C_n>0, \gamma>1,\ \mathbb{P}\left(\|\tilde{s}-s\|^2>C_n\inf_{m\in\mathcal{M}_n}\left\{\|s-\hat{s}_{A,m}\|^2\right\}\right)\leq\frac{K}{n^\gamma}. \tag{4.3}$$

$$\exists K>0, C_n>0,\ \mathbb{E}\left(\|\tilde{s}-s\|^2\right)\leq C_n\mathbb{E}\left(\inf_{m\in\mathcal{M}_n}\left\{\|s-\hat{s}_{A,m}\|^2\right\}\right)+\frac{K}{n}. \tag{4.4}$$

In both cases, we want the leading constant C_n being as close as possible to 1. In order to build $\hat{m}$, we remark that, for all m in $\mathcal{M}_n$, we have

$$\|s-\hat{s}_{A,m}\|^2=\|\hat{s}_{A,m}\|^2-2P\hat{s}_{A,m}+\|s\|^2=P_AQ(\hat{s}_{A,m})+2\nu_A(\hat{s}_{A,m})+\|s\|^2,$$

where $\nu_A=P_A-P$. An oracle minimizes $\|s-\hat{s}_{A,m}\|^2$ and thus $P_AQ(\hat{s}_{A,m})+2\nu_A(\hat{s}_{A,m})$. As we want to imitate the oracle, we will design a map pen : $\mathcal{M}_n\to\mathbb{R}^+$ and choose

$$\hat{m}\in\arg\min_{m\in\mathcal{M}_n}P_AQ(\hat{s}_{A,m})+\mathrm{pen}(m),\ \tilde{s}=\hat{s}_{A,\hat{m}}. \tag{4.5}$$

It is clear that the ideal penalty is $\mathrm{pen}_{id}(m)=2\nu_A(\hat{s}_{A,m})$ and our goal is to design sharp estimators of this quantity as penalty functions.
The key point to obtain oracle inequalities is the following decomposition of the risk of $\tilde{s}$. For all m in $\mathcal{M}_n$, let

$$p(m)=\nu_A(\hat{s}_{A,m}-s_m)=\|\hat{s}_{A,m}-s_m\|^2.$$

For all m in $\mathcal{M}_n$,

$$\begin{aligned}\|s-\tilde{s}\|^2 &= \|\tilde{s}\|^2-2P\tilde{s}+\|s\|^2=\|\tilde{s}\|^2-2P_A\tilde{s}+2\nu_A\tilde{s}+\|s\|^2\\ &\leq P_AQ(\hat{s}_{A,m})+\mathrm{pen}(m)+(2\nu_A(\tilde{s})-\mathrm{pen}(\hat{m}))+\|s\|^2\\ &= \|s-\hat{s}_{A,m}\|^2+(\mathrm{pen}(m)-2\nu_A(\hat{s}_{A,m}))+(2\nu_A(\tilde{s})-\mathrm{pen}(\hat{m}))\end{aligned}$$

Thus, for all m in $\mathcal{M}_n$,

$$\|s-\tilde{s}\|^2\leq\|s-\hat{s}_{A,m}\|^2+(\mathrm{pen}(m)-2p(m))+(2p(\hat{m})-\mathrm{pen}(\hat{m}))+2\nu_A(s_{\hat{m}}-s_m). \tag{4.6}$$

4.1.3 Optimal model selection

Let us now precise the definition of the methods that we will use to calibrate the penalty.

The slope algorithm

The "slope heuristic" was introduced by Birgé & Massart [17] in the Gaussian regression framework. It states that there exists a complexity measure Δ_m of S_m and a constant $K_{\min}$ such that

1. if $\mathrm{pen}(m) < K_{\min}\Delta_m$, $\Delta_{\hat{m}}$ is too large, typically $\Delta_{\hat{m}} \geq C \sup_{m \in \mathcal{M}_n} \Delta_m$,
2. if $\mathrm{pen}(m) \simeq K\Delta_m$ for some $K > K_{\min}$, then $\Delta_{\hat{m}}$ is "much smaller",
3. if $\mathrm{pen}(m) \simeq 2K_{\min}\Delta_m$, then the risk of the selected estimator satisfies
$$\|\tilde{s} - s\|^2 \leq C_n \inf_{m \in \mathcal{M}_n} \left\{ \|s - \hat{s}_{A,m}\|^2 \right\}, \text{ with } C_n \to 1, \text{ when } n \to \infty$$
in expectation and with large probability.

When both Δ_m and the associated $K_{\min}$ are known, point 3 in this heuristic says that $\mathrm{pen}(m) \simeq 2K_{\min}\Delta_m$ is an optimal penalty. This heuristic is classically used when Δ_m is known and $K_{\min}$ is unknown. Arlot & Massart [7] introduced the following algorithm to calibrate the penalty term in this situation.

Slope algorithm

- For all $K > 0$, compute the selected model $\hat{m}(K)$ given by (4.5) with the penalty $\mathrm{pen}(m) = K\Delta_m$ and the associated complexity $\Delta_{\hat{m}(K)}$.
- Find a constant K_o such that $\Delta_{\hat{m}(K)}$ is large when $K < K_o$, and "much smaller" when $K > K_o$.
- Take the final $\hat{m} = \hat{m}(2K_o)$.

In Chapter 2, we justified the slope heuristic in density estimation with independent data for $\Delta_m = \mathbb{E}(\|s_m - \hat{s}_{A,m}\|^2)$, $K_{\min} = 1$. This complexity is unknown in practice and has to be estimated. We proposed a resampling estimator and proved that it works without extra assumptions on our collection of models. Then, we remarked that, when the models are very regular, we can also use the linear dimension d_m of S_m as a complexity measure and calibrate K_o with the slope algorithm. In this paper, we will extend all these results to mixing processes.

Resampling penalties

Data-driven penalties have been studied in density estimation, in particular, cross validation methods as in Stone [64], Rudemo [62] or Celisse [21]. We extend the approach of Chapter 2 based on resampling penalties introduced by Arlot [5]. We prove that it provides optimal model selection procedures. The main ingredient in the proofs is a concentration inequality for the supremum of the resampling-based empirical process proved in Chapter 2. This inequality states that the resampling penalty defined in Section 4.2 is essentially equal to the ideal penalty when the data are independent. Another important ingredient is the coupling properties of mixing processes. The coupling result proved in Viennet [70] for β-mixing processes allows a straightforward extension of the results of Chapter 2. The coupling result available for τ-mixing sequences is not so powerful and the extension of the results of Chapter 2 to that case requires new methods of proofs.

The chapter is organized as follows. In Section 4.2, we introduce our new estimation procedure and describe our main assumptions. In Section 4.3, we state our main results, we prove the efficiency of the penalized least-squares estimators based on the slope heuristic and on resampling methods. In Section 4.4, we compare our new estimators with those given in Chapter 3. The proofs of the main theorems are postponed to Section 4.5. Section 4.6 is an Appendix where we recall some probabilistic lemmas proved in Chapter 2.

4.2 New estimation procedures

4.2.1 Blockwise decomposition of the data

Assume that n is even and let p and q be two integers such that $2pq = n$. For all $k = 0, ..., p-1$, let $I_k = (2kq+1, ..., (2k+1)q)$, $A_k = (X_l)_{l \in I_k}$ and $I = \cup_{k=0}^{p-1} I_k$. For all functions t in $L^2(\mu)$ and all $x_1, ..., x_q$ in $\mathbb{X}$, let

$$L_q(t)(x_1, ..., x_q) = \frac{1}{q}\sum_{i=1}^{q} t(x_i),\ P_A t = \frac{1}{p}\sum_{k=0}^{p-1} L_q(t)(A_k) = \frac{2}{n}\sum_{i \in I} t(X_i),$$

$$\nu_A(t) = (P_A - P)(t).$$

Let S_m be linear subspace of $L^2(\mu)$ and let $Q : L^2(\mu) \to L^1(P),\ t \mapsto \|t\|^2 - 2t$. The estimator $\hat{s}_{A,m}$ associated to S_m, is defined by

$$\hat{s}_{A,m} \in \arg\min_{t \in S_m} P_A Q(t). \tag{4.7}$$

Given an orthonormal basis $(\psi_\lambda)_{\lambda \in m}$ of S_m, classical computations prove that

$$\hat{s}_{A,m} = \sum_{\lambda \in m} (P_A \psi_\lambda)\psi_\lambda,\ \|s_m - \hat{s}_{A,m}\|^2 = \sum_{\lambda \in m} (\nu_A(\psi_\lambda))^2 = \sup_{t \in B_m} (\nu_A(t))^2.$$

4.2.2 Resampling penalties

The first penalization procedure is based on the resampling penalties introduced by Arlot [5]. The resampling algorithm is slightly modified in order to keep the dependence structure inside the blocks (see Künsh [44], Liu & Singh [51] or Radulovic [59]).

Let $W_0, ..., W_{p-1}$ be a resampling scheme, that is, a vector of random variables, independent of $X_1, ..., X_n$ and exchangeable, i.e., for all permutation ξ of $\{0, ...p-1\}$,

$$(W_{\xi(0)}, ..., W_{\xi(p-1)}) \text{ has the same law as } (W_0, ..., W_{p-1}).$$

Let P_A^W and ν_A^W be the associated resampling empirical processes defined, for all t in $L^2(\mu)$, by

$$P_A^W(t) = \frac{1}{p}\sum_{k=0}^{p-1} W_k L_q(t)(A_k),$$

$$\nu_A^W(t) = (P_A^W - \bar{W}_p P_A)(t) = \frac{1}{p}\sum_{k=0}^{p-1}(W_k - \bar{W}_p)L_q(t)(A_k), \text{ where } \bar{W}_p = \frac{1}{p}\sum_{k=0}^{p-1} W_k.$$

For all m in $\mathcal{M}_n$, let

$$\hat{s}_{A,m}^W = \arg\min_{t\in S_m} P_A^W(t) = \sum_{\lambda\in m}(P_A^W\psi_\lambda)\psi_\lambda.$$

Setting $v_W^2 = \mathrm{Var}(W_1 - \bar{W}_p)$, the resampling penalty is defined by

$$\mathrm{pen}(m) = \frac{2}{v_W^2}\mathbb{E}^W\left(\sup_{t\in B_m}(\nu_A^W(t))^2\right) = \frac{2}{v_W^2}\sum_{\lambda\in m}\mathbb{E}^W\left((\nu_A^W(\psi_\lambda))^2\right). \qquad (4.8)$$

Hereafter, for all m in $\mathcal{M}_n$ and for all function pen, the final estimator is always denoted by

$$\tilde{s} = \hat{s}_{A,\hat{m}}, \text{ where } \hat{m} = \arg\min_{m\in\mathcal{M}_n} P_A Q(\hat{s}_{A,m}) + \mathrm{pen}(m). \qquad (4.9)$$

4.2.3 Some measures of dependence

β-mixing data

The coefficient β was introduced by Rozanov & Volkonskii [71]. For a random variable Y defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a σ-algebra $\mathcal{M}$ in $\mathcal{A}$, let

$$\beta(\mathcal{M}, \sigma(Y)) = \mathbb{E}\left(\sup_{A\in\mathcal{B}}|\mathbb{P}_{Y|\mathcal{M}}(A) - \mathbb{P}_Y(A)|\right).$$

For all stationary sequence of random variables $(X_n)_{n\in\mathbb{Z}}$ defined on $(\Omega, \mathcal{A}, \mathbb{P})$, let

$$\beta_k = \beta(\sigma(X_i, i\le 0), \sigma(X_i, i\ge k)).$$

The process $(X_n)_{n\in\mathbb{Z}}$ is said to be β-mixing when $\beta_k \to 0$ as $k\to\infty$. Examples of β-mixing processes can be found in the books of Doukhan [28] and Bradley [19]. One of the most important is the following: a stationary, irreducible, aperiodic and positively recurent Markov chain $(X_i)_{i\ge 1}$ is β-mixing.
Let us recall Lemma 5.1 in Viennet [70].

Lemma: *(Viennet 1997) Assume that the process $(X_1, ..., X_n)$ is β-mixing and let p, q and $A_0, ..., A_{p-1}$ be respectively the integers and the random variables defined in Section 4.2.1. There exist random variables $A_0^*, ..., A_{p-1}^*$ such that:*

1. *for all $k = 0, ..., p-1$, $A_k^* = (X_{2kq+1}^*, ..., X_{(2k+1)q}^*)$ has the same law as A_k,*

2. *for all $k = 0, ..., p-1$, A_k^* is independent of $A_0, ..., A_{k-1}, A_1^*, ..., A_{k-1}^*$,*

3. *for all $k = 0, ..., p-1$, $\mathbb{P}(A_k \neq A_k^*) \le \beta_q$.*

τ-mixing data

The coefficient τ was introduced by Dedecker & Prieur [25]. For all l in $\mathbb{N}^*$, for all x, y in $\mathbb{R}^l$, let $d_l(x,y) = \sum_{i=1}^{l} |x_i - y_i|$. For all l in $\mathbb{N}^*$, for all function t defined on $\mathbb{R}^l$, the Lipschitz semi-norm of t is defined by

$$\mathrm{Lip}_{d_l}(t) = \sup_{x \neq y \in \mathbb{R}^l} \frac{|t(x) - t(y)|}{d_r(x,y)}.$$

Let λ_1 be the set of all functions $t : \mathbb{R}^l \to \mathbb{R}$ such that $\mathrm{Lip}_{d_l}(t) \leq 1$. For all integrable, $\mathbb{R}^l$-valued, random variables Y defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and all σ-algebra $\mathcal{M}$ in $\mathcal{A}$, let

$$\tau(\mathcal{M}, Y) = \mathbb{E}\left(\sup_{t \in \lambda_1} |\mathbb{P}_{Y|\mathcal{M}}(t) - \mathbb{P}_Y(t)| \right).$$

For all stationary sequences of integrable random variables $(X_n)_{n \in \mathbb{Z}}$ defined on $(\Omega, \mathcal{A}, \mathbb{P})$, for all integers k, r, let

$$\tau_{k,r} = \max_{1 \leq l \leq r} \frac{1}{l} \sup_{k \leq i_1 < .. < i_l} \{\tau(\sigma(X_p, p \leq 0), (X_{i_1}, ..., X_{i_l}))\}, \ \tau_k = \sup_{r \in \mathbb{N}^*} \tau_{k,r}.$$

The process $(X_n)_{n \in \mathbb{Z}}$ is said to be τ-mixing when $\tau_k \to 0$ as $k \to \infty$. Examples of τ-mixing processes can be found in the book of Dedecker *et. al* [24] or the articles of Dedecker & Prieur [25] and Comte *et. al* [22].
The following result has been obtain in Claim 1 in the proof of Theorem 3.4.1 of Chapter 3. This is a consequence of a coupling lemma proved by Dedecker & Prieur [25].

Lemma: [τ-coupling, Claim 1 p17 in [49]] *Assume that the process $(X_1, ..., X_n)$ is τ-mixing and let p, q and $A_0, ..., A_{p-1}$ be respectively the integers and the random variables defined in Section 4.2.1. There exist random variables $A_0^*, ..., A_{p-1}^*$ such that:*

1. *for all $k = 0, ..., p-1$, $A_k^* = (X_{2kq+1}^*, ..., X_{(2k+1)q}^*)$ has the same law as A_k,*
2. *for all $k = 0, ..., p-1$, A_k^* is independent of $A_0, ..., A_{k-1}, A_1^*, ..., A_{k-1}^*$,*
3. *for all $k = 0, ..., p-1$, $\mathbb{E}(d_q(A_k, A_k^*)) \leq q\tau_q$.*

4.2.4 Main assumptions

Let p, q and $A_0, ..., A_{p-1}$ be respectively the integers and the random variables defined in Section 4.2.1. For all m, m' in $\mathcal{M}_n$, let

$$v_{A,m,m'}^2 = \sup_{t \in S_m + S_{m'}, \|t\| \leq 1} q\mathrm{Var}(L_q(t)(A_0)), \ D_{A,m} = q \sum_{\lambda \in m} \mathrm{Var}(L_q(\psi_\lambda)(A_0)),$$

$$b_{m,m'} = \sup_{t \in S_m + S_{m'}, \|t\| \leq 1} \|t\|_\infty .$$

For all m in $\mathcal{M}_n$, let

$$R_{A,m} = n\|s - s_m\|^2 + 2D_{A,m}, \ e_{A,m,m'} = \frac{q}{p} b_{m,m'}^2.$$

We denote by $e_{A,m} = e_{A,m,m}$, $v_{A,m} = v_{A,m,m}$. For all $k \in \mathbb{N}$, let $\mathcal{M}_n^k = \{m \in \mathcal{M}_n,\ R_{A,m} \in [k, k+1)\}$ and for all n in $\mathbb{N}^*$ and, for all $k > 0$, $k' > 0$ and $\gamma \geq 0$, let $[k]$ denote the integer part of k and let

$$l_{n,\gamma}(k,k') = \ln\left((1+\mathrm{Card}(\mathcal{M}_n^{[k]}))(1+\mathrm{Card}(\mathcal{M}_n^{[k']}))(k+1)(k'+1)\right) + (\ln n)^{\gamma} \tag{4.10}$$

The following assumptions generalize Assumptions **[V]** and **[BR]** made in Chapter 2.

[V']: *There exist $\gamma > 1$ and a sequence $(\epsilon_n)_{n\in\mathbb{N}}$, with $\epsilon_n \to 0$ such that, for all n in $\mathbb{N}$,*

$$\sup_{(m,m')\in(\mathcal{M}_n)^2}\left\{\left(\left(\frac{v^2_{A,m,m'}}{R_{A,m}\vee R_{A,m'}}\right)^2 \vee \frac{e_{A,m,m'}}{R_{A,m}\vee R_{A,m'}}\right) l_{m,m'}\right\} \leq \epsilon_n^4,$$

where, for all m, m' in $\mathcal{M}_n$, $l_{m,m'} = l_{n,\gamma}(R_{A,m}, R_{A,m'})$.

[BR'] *There exist two sequences $(h_n^*)_{n\in\mathbb{N}^*}$ and $(h_n^o)_{n\in\mathbb{N}^*}$ with $(h_n^o \vee h_n^*) \to 0$ as $n \to \infty$ such that, for all n in $\mathbb{N}^*$, for all $m_o \in \arg\min_{m\in\mathcal{M}_n} R_{A,m}$ and all $m^* \in \arg\max_{m\in\mathcal{M}_n} D_{A,m}$, we have*

$$\frac{R_{A,m_o}}{D_{A,m^*}} \leq h_n^o,\quad \frac{n\|s - s_{m^*}\|^2}{D_{A,m^*}} \leq h_n^*.$$

4.3 Main results

4.3.1 Resampling penalties

The first theorem gives oracle inequalities satisfied by the estimator selected by the resampling penalty.

Theorem 4.3.1 *Let $X_1, ..., X_n$ be a strictly stationary sequence of random variables with common density s and let $(S_m)_{m\in\mathcal{M}_n}$ be a collection of linear subspaces of $L^2(\mu)$ satisfying Assumption* **[V']**. *Let $\tilde{s}$ be the estimator defined in (4.9) with pen(m) defined in (4.8).*
Assume that $X_1, ..., X_n$ are β-mixing, then, there exists a constant $C > 0$ such that

$$\mathbb{P}\left(\|s-\tilde{s}\|^2 > (1+110\epsilon_n)\inf_{m\in\mathcal{M}_n}\|s - \hat{s}_{A,m}\|^2\right) \leq Ce^{-\frac{1}{2}(\ln n)^{\gamma}} + p\beta_q. \tag{4.11}$$

Assume that $X_1, ..., X_n$ are real valued and τ-mixing, then, there exists an absolute constant $C > 0$ such that we have

$$\mathbb{E}\left(\|s-\tilde{s}\|^2\right) \leq (1+160\epsilon_n)\mathbb{E}\left(\inf_{m\in\mathcal{M}_n}\|s-\hat{s}_{A,m}\|^2\right) + C\left(e^{-\frac{1}{2}(\ln n)^{\gamma}} + \tau_q MC_n\right), \tag{4.12}$$

where the mixing complexity MC_n is defined by the following formula:

$$MC_n = \sum_{m\in\mathcal{M}_n}\left(\left\|\sum_{\lambda\in m}|\psi_\lambda|\right\|_\infty \sup_{\lambda\in m} Lip(\psi_\lambda) + \|s\||\mathcal{M}_n| \sup_{t\in B_m} Lip(t)\right).$$

Comments:

- Theorem 4.3.1 can be compared with Theorem 2.2.5 in Chapter 2. An extra term $p\beta_q$ appears in the control of the deviation probability when the data are β-mixing. We show in Section 4.4 that p and q can be chosen in order to have $p\beta_q \leq Cn^{-\alpha}$ for some $\alpha > 1$ under classical assumptions on the mixing coefficients.

- When the data are τ-mixing, we see that the mixing coefficient τ_q must control the mixing complexity MC_n. It is clear that $MC_n = \infty$ for many collections of linear spaces $(S_m)_{m\in\mathcal{M}_n}$ (as histogram spaces for example). Therefore, we have to choose carefully the collection $\mathcal{M}_n$ when we deal with τ-mixing data. In Section 4.4, we show that, on wavelet spaces, p and q can be chosen in order to have $\tau_q MC_n \leq Cn^{-1}$ under classical assumptions on the mixing coefficient.

- Up to our knowledge, inequalities (4.11) and (4.12) are the first oracle inequalities obtained for totally data driven PLSE of the density s when the data are mixing. Moreover, this is the first time that the risk of the selected estimator is compared with the risk of an oracle and not with an upper bound.

4.3.2 Slope heuristic

We will now justify the use of the slope heuristic when the data are mixing. Recall that two types of results are required. First, we need to prove that a small penalty leads to a too large complexity of the selected model and that we cannot obtain an oracle inequality in this case. This is the purpose of the following theorem, which generalizes Theorem 2.2.2 in Chapter 2.

Theorem 4.3.2 *Let $X_1, ..., X_n$ be a strictly stationary sequence of random variables, with common density s. Let $\mathcal{M}_n$ be a collection of models satisfying Assumptions* **[V']**, **[BR']** *and let $\epsilon_n^* = \epsilon_n \vee h_n^*$.*
Assume that there exists a constant $0 < \delta < 1$ such that, for all m in $\mathcal{M}_n$,

$$0 \leq pen(m) \leq \frac{(2-\delta)D_{A,m}}{n}.$$

Let $\hat{m}, \tilde{s}$ be the random variables defined in (4.9). Assume that $X_1, ..., X_n$ are β-mixing and let

$$c_n = \frac{\delta - 75\epsilon_n^*}{2(1+27\epsilon_n)}.$$

There exists a constant $C > 0$, such that, with probability larger than $1 - Ce^{-\frac{1}{2}(\ln n)^\gamma} - p\beta_q$,

$$D_{A,\hat{m}} \geq c_n D_{A,m^*},\ \|s - \tilde{s}\|^2 \geq \frac{c_n}{h_n^o} \inf_{m\in\mathcal{M}_n} \|s - \hat{s}_{A,m}\|^2. \tag{4.13}$$

Assume now that $X_1, ..., X_n$ are τ-mixing, let MC_n be the mixing complexity defined in Theorem 4.3.1 and let

$$c_n' = \frac{\delta - h_n^*}{2(1+35\epsilon_n)}.$$

There exists an absolute constant $C > 0$ such that

$$\mathbb{E}(D_{A,\hat{m}}) \geq c_n' D_{A,m^*} - Cn\left(e^{-\frac{1}{2}(\ln n)^\gamma} + \tau_q MC_n\right). \tag{4.14}$$

$$\mathbb{E}\left(\|s-\tilde{s}\|^2\right) \geq 2\frac{c_n'}{h_n^o}\mathbb{E}\left(\inf_{m\in\mathcal{M}_n}\|s-\hat{s}_{A,m}\|^2\right) - C\left(e^{-\frac{1}{2}(\ln n)^\gamma}+\tau_q MC_n\right). \tag{4.15}$$

Remark: When n is sufficiently large, $c_n \geq \delta/4$, $c_n' \geq \delta/4$. Hence, when $\mathrm{pen}(m)$ is not larger than $2D_{A,m}/n$, inequalities (4.13) and (4.14) ensure that with high probability or in expectation $D_{A,\hat{m}} \geq cD_{A,m^*}$, which is as large as possible and inequalities (4.13) and (4.15) show that no optimal oracle inequality can hold. This proves the first point of the slope heuristic.
The following theorem justifies the remaining points.

Theorem 4.3.3 *Let $X_1,...,X_n$ be a stationary sequence of random variables with common density s. Let $(S_m)_{m\in\mathcal{M}_n}$ be a collection of models satisfying* **[V']**. *For all m in $\mathcal{M}_n$, let pen(m) be a penalty function and let $\tilde{s}$ be the estimator defined in (4.9).*
Assume that $X_1,...,X_n$ are β-mixing and that there exist constants $\bar{\delta} \geq \underline{\delta} > -1$ and $0 \leq p' < 1$ such that, with probability at least $1-p'$, for all m in $\mathcal{M}_n$,

$$\frac{4D_{A,m}}{n} + \underline{\delta}\frac{R_{A,m}}{n} \leq pen(m) \leq \frac{4D_{A,m}}{n} + \bar{\delta}\frac{R_{A,m}}{n}.$$

Let

$$c_n = \left(\begin{array}{ll} \frac{1+\bar{\delta}+37\epsilon_n}{2(1+\underline{\delta}-27\epsilon_n)} & \text{if } 1+\underline{\delta}-27\epsilon_n > 0 \\ +\infty & \text{if } 1+\underline{\delta}-27\epsilon_n \leq 0 \end{array}\right..$$

There exists a constant $C > 0$, such that, with probability at least $1 - Ce^{-\frac{1}{2}(\ln n)^\gamma} - p\beta_q - p'$,

$$D_{A,\hat{m}} \leq c_n R_{A,m_o},\ \|s-\tilde{s}\|^2 \leq 2c_n \inf_{m\in\mathcal{M}_n}\|s-\hat{s}_{A,m}\|^2, \tag{4.16}$$

Assume now that $X_1,...,X_n$ are τ-mixing and that there exist constants $\underline{\delta} > -1$, $\bar{\delta} > -1$ and a sequence $(e_n)_{n\in\mathbb{N}}$, with $\sum_{n\in\mathbb{N}} e_n < \infty$ such that

$$\mathbb{E}\left(\sup_{m\in\mathcal{M}_n}\left(\frac{4D_{A,m}}{n} + \underline{\delta}\frac{R_{A,m}}{n} - pen(m)\right)_+\right) \leq e_n,$$

$$\mathbb{E}\left(\sup_{m\in\mathcal{M}_n}\left(pen(m) - \frac{4D_{A,m}}{n} - \bar{\delta}\frac{R_{A,m}}{n}\right)_+\right) \leq e_n.$$

Let MC_n be the mixing complexity defined in Theorem 4.3.1 and let

$$c_n' = \left(\begin{array}{ll} \frac{1+\bar{\delta}+55\epsilon_n}{2(1+\underline{\delta}-125\epsilon_n)} & \text{if } 1+\underline{\delta}-125\epsilon_n > 0 \\ +\infty & \text{if } 1+\underline{\delta}-125\epsilon_n \leq 0 \end{array}\right..$$

There exists a constant $C > 0$, such that,

$$\mathbb{E}\left(D_{A,\hat{m}}\right) \leq c_n\left(R_{A,m_o} + n(C\tau_q MC_n + e_n)\right). \tag{4.17}$$

$$\mathbb{E}\left(\|s-\tilde{s}\|^2\right) \leq c_n\left(\mathbb{E}\left(\inf_{m\in\mathcal{M}_n}\|s-\hat{s}_{A,m}\|^2\right) + C\left(\tau_q MC_n + e_n\right)\right) \tag{4.18}$$

Comments:

- $D_{A,\hat{m}}$ jumps from D_{A,m^*} (Theorem 4.3.2) to R_{A,m_o} when $\mathrm{pen}(m)$ is around $2D_{A,m}/n$, this justifies the second point of the slope heursitic and clarifies what we meant by "much smaller". Point 3 of the slope heuristic comes from inequalities (4.16) and (4.18) applied with $\underline{\delta} = \bar{\delta} = 0$.

- We see in this theorem why it may be usefull to overpenalize a little from a non asymptotic point of view. Imagine that $1 - 67\epsilon_n$ is very close to 0, then c_n will be much smaller if we choose $\underline{\delta} > 0$ than if we take its asymptotic optimal value 0.

- Theorem 4.3.3 can be compared with Theorem 2.2.2 in Chapter 2. We observe the same differences as those between Theorem 4.3.1 and Theorem 2.2.5 in Chapter 2. In the following Section we will discuss more precisely these differences under classical Assumptions on the mixing coefficients.

- The practical implementation of these algorithms is discussed in general in Arlot & Massart [7], see also the discussion for density estimation in Chapter 2. The slope heuristic is very quick to compute and shall be prefered when we know a shape of the ideal penalty. The resampling-based estimators give this shape for more general collections. The resampling penalty given in Theorem 4.3.1 do not depend on the observation of a jump of $D_{\hat{m}}$ that may be hard to detect in practice. However, the constant C_W given in this theorem is asymptotically optimal and the slope algorithm clearly improves the selected estimator in the simulation study of Chapter 2.

4.4 Comparison with previous results

In this section, we compare the estimator given by the resampling penalty with those given in Chapter 3. Let us first insist here on the fact that the resampling procedure is totally data driven and entirely computable. In Chapter 3, recall that the estimator was chosen among the collection of least-squares estimators $(\hat{s}_m)_{m\in\mathcal{M}_n}$, where $\hat{s}_m = \arg\min_{t\in S_m} P_nQ(t)$, by a penalization procedure

$$\tilde{s} = \hat{s}_{\hat{m}},\text{ where } \hat{m} = \arg\min_{m\in\mathcal{M}_n} P_nQ(\hat{s}_m) + \mathrm{pen}(m). \tag{4.19}$$

Mixing assumptions In Chapter 3, we considered two kinds of rates of convergence to 0 of the mixing coefficients. Let $\gamma = \beta$ or τ.
[**AR**(θ)] arithmetical γ-mixing with rate θ: there exists $C > 0$ such that, for all k in $\mathbb{N}$, $\gamma_k \leq C(1+k)^{-(1+\theta)}$,
[**GEO**(θ)] geometrical γ-mixing with rate θ: there exists $C > 0$ such that, for all k in $\mathbb{N}$, $\gamma_k \leq Ce^{-\theta k}$.

4.4.1 β-mixing processes

In Chapter 3, in the β-mixing framework, the collection of models was assumed to satisfy the following assumptions.

$[M_1]$ $(\psi_\lambda)_{\lambda\in\Lambda}$ *is an orthonormal system of* $L^2(\mu)$ *and, for all* $m \in \mathcal{M}_n$*,* S_m *is the linear span of* $\{\psi_\lambda\}_{\lambda\in m}$ *with finite dimension* $d_m = |m| \geq 2$ *and* $N_n = \max_{m\in\mathcal{M}_n} d_m$

satisfies $N_n \leq n$*;*
$[M_2]$ *there exists a constant* Φ *such that*

$$\forall m, m' \in \mathcal{M}_n, \forall t \in S_m, \forall t' \in S_{m'}, \|t + t'\|_\infty \leq \Phi\sqrt{\dim(S_m + S_{m'})}\|t + t'\|_2;$$

$[M_3]$ $d_m \leq d_{m'}$ *implies that* $m \subset m'$ *and so* $S_m \subset S_{m'}$*.*

We can easily deduce from $[M_1]$ that for all $k > n$, $\mathcal{M}_n^k = \emptyset$ and, from $[M_3]$, for all $k \leq n$, $\mathrm{Card}(\mathcal{M}_n^k) \leq 1$. Hence, there exists a constant c_V such that, for all $\gamma > 1$,

$$l_{m,m'} \leq c_V(\ln n)^{2\gamma}$$

In Chapter 3, we derived the following inequalities. If there exists $\theta > 1$ such that $X_1, ..., X_n$ are arithmetically $[\mathbf{AR}(\theta)]$, β-mixing, there exists constants c_v, c_e, c_D, c_M such that, for all m, m' in $\mathcal{M}_n$,

$$v^2_{A,m,m'} \leq c_v(d_m \vee d_{m'})^{3/4},\ b^2_{A,m,m'} \leq c_e(d_m \vee d_{m'}), D_{A,m} \leq c_D d_m.$$

The constants c_v and c_D depend on the mixing coefficients and are unknown in practice. In order to verify **[V']**, we need two other assumptions.

$[M_4]$ *There exists* $c'_D > 0$ *such that, for all* n *in* $\mathbb{N}^*$*, for all* m *in* $\mathcal{M}_n$*,* $D_{A,m} \geq c'_D d_m$*.*
$[M_5]$ *There exist* $\gamma > 1$ *and a sequence* $r_n \to \infty$ *such that* $R_n(\ln n)^{-4\gamma} \geq r_n$*, where* $R_n = \inf_{m\in\mathcal{M}_n} R_{A,m}$*.*

Without loss of generality, assume that $\gamma \leq 3/2$ in $[M_5]$. Now, choose $p \geq \sqrt{n}(\ln n)^2/2$, $q \geq \sqrt{n}(\ln n)^{-2}/2$ such that $2pq = n$. Hence, there exists a constant c_M such that $p\beta_q \leq c_M(\log n)^{2(\theta+2)}n^{-\theta/2}$. For all m, m' in $\mathcal{M}_n$,

$$e_{A,m,m'} \leq 2c_e\frac{d_m \vee d_{m'}}{(\ln n)^4} \leq \frac{2c_e}{c'_D}\frac{D_{A,m} \vee D_{A,m'}}{(\ln n)^4} \leq \frac{2c_e}{c'_D}(\ln n)^{-1}\frac{R_{A,m} \vee R_{A,m'}}{(\ln n)^{2\gamma}}.$$

When $d_m \vee d_{m'} \leq r_n(\ln n)^{4\gamma}$, then

$$v^2_{A,m,m'} \leq c_v(d_m \vee d_{m'})^{3/4} \leq c_v(r_n)^{-1/4}\frac{R_n}{(\ln n)^\gamma} \leq c_v(r_n)^{-1/4}\frac{R_{A,m} \vee R_{A,m'}}{(\ln n)^\gamma}.$$

When $d_m \vee d_{m'} \geq r_n(\ln n)^{4\gamma}$, then

$$v^2_{A,m,m'} \leq c_v(d_m \vee d_{m'})^{3/4} \leq \frac{c_v}{c'_D}\frac{D_{A,m} \vee D_{A,m'}}{(d_m \vee d_{m'})^{1/4}} \leq \frac{c_v}{c'_D}\frac{R_{A,m} \vee R_{A,m'}}{r_n^{1/4}(\ln n)^\gamma}.$$

Therefore, $[M_1]$-$[M_5]$ and $[\mathbf{AR}(\theta)]$ with $\theta > 1$ imply **[V']** with

$$\epsilon_n = C\left((\ln n)^{-1/4} \wedge r_n^{-1/8}\right).$$

We have obtained the following corollary of Theorem 4.3.1.

Corollary 4.4.1 *Let* $\mathcal{M}_n$ *be a collection of models satisfying* $[M_1]$*-* $[M_5]$*. Assume that the process* $(X_n)_{n\in\mathbb{Z}}$ *is strictly stationary and arithmetically* $[\mathbf{AR}(\theta)]$ *β-mixing with mixing rate* $\theta > 1$*. Let* $\tilde{s}$ *be the estimator defined in (4.9) with a resampling penalty (4.8). Let* $\epsilon_n^* = (\ln n)^{-1/4} \wedge r_n^{-1/8}$*.*
There exist constants $C > 0$ *and* $\kappa > 0$ *such that*

$$\mathbb{P}\left(\|\tilde{s} - s\|_2^2 > (1 + \kappa\epsilon_n^*)\inf_{m\in\mathcal{M}_n}\|s - \hat{s}_{A,m}\|_2^2\right) \leq C\frac{(\log n)^{2(\theta+2)}}{n^{\theta/2}}.$$

Comments: This result can be compared with Corollary 2.3.1 in Chapter 2. In the independent case, these rates of convergence of the leading constant is always given by $(r_n)^{-1/4}$ and his rate is often polynomial in n. It is not faster than $(\ln n)^{-1/4}$ in the β-mixing case. The deviation probability was upper bounded by $Ce^{-\frac{1}{2}(\ln n)^\gamma}$ for some constants $C > 0$ and $\gamma > 1$, it is now polynomial in n.
$[M_5]$ is hard to check in general. Let $c_d^{-1} = 2\sup_{x\geq 1}(\ln x)^8 x^{-1}$. $[M_5]$ is satisfied for example, if there is no model in $\mathcal{M}_n$ that have dimension $d_m \leq c_d(\ln n)^8$ and if $[M_4]$ is satisfied. In this case, $[M_5]$ holds with $r_n = (\ln n)^2$ and we deduce from our previous computations the following result.

Corollary 4.4.2 *Let $\mathcal{M}_n$ be a collection of models satisfying $[M_1]$- $[M_4]$. Assume that the process $(X_n)_{n\in\mathbb{Z}}$ is strictly stationary and arithmetically $[\mathbf{AR}(\theta)]$ β-mixing with mixing rate $\theta > 1$. Let $\tilde{s}$ be the estimator defined in (4.9) with a resampling penalty (4.8). Then, there exist constants $\kappa > 0$, $C > 0$ such that, with probability larger than $1 - Cn^{-\theta/2}(\log n)^{2(\theta+2)}$,*

$$\|\tilde{s} - s\|_2^2 \leq \left(1 + \frac{\kappa}{(\ln n)^{1/4}}\right) \inf_{m\in\mathcal{M}_n, d_m\geq c_d(\ln n)^8} \|s - \hat{s}_{A,m}\|_2^2 .$$

Recall the theorem we proved in Chapter 3 for β-mixing sequences.

Theorem 4.4.3 *Let $\mathcal{M}_n$ be a collection of models satisfying $[M_1]$, $[M_2]$ and $[M_3]$. Assume that the process $(X_n)_{n\in\mathbb{Z}}$ is strictly stationary and arithmetically $[\mathbf{AR}(\theta)]$ β-mixing with mixing rate $\theta > 2$. Let $\tilde{s}$ be the estimator defined in (4.19) with*

$$pen(m) = Kc_D \frac{d_m}{n}, \text{ where } K > 4.$$

Then, for all $\kappa > 2$, there exist $c_0 > 0, L_s > 0$, $\gamma_1 > 0$ and a sequence $\epsilon_n \to 0$, such that, with probability larger than $1 - Cn^{-\theta/2}(\log n)^{\kappa(\theta+2)}$

$$\|\tilde{s} - s\|_2^2 \leq (1 + \epsilon_n) \inf_{m\in\mathcal{M}_n, d_m\geq c_0(\log n)^{\gamma_1}} \left(\|s - s_m\|_2^2 + pen(m)\right) .$$

Comments: Both procedures lead to trajectorial oracle inequalities (4.3). The procedure of Theorem 4.4.3 depends on the constant c_D, which is in general unknown and which may be be very pessimistic. On the other hand, in Corollary 4.4.2, the selection algorithm $X_1, ..., X_n \mapsto \tilde{s}$ is totally computable in practice. Moreover, the risk of $\tilde{s}$ in Corollary 4.4.2 is compared with the best of the risks in the collection $\mathcal{M}_n$. It is compared with an upper bound on $\|s - s_m\|^2 + 2\mathbb{E}\left(\|s_m - \hat{s}_{A,m}\|^2\right)$ in Theorem 4.4.3. Finally, Theorem 4.4.3 does not require $[M_4]$ to work. However, our new estimator improves the estimator given in Theorem 4.4.3 every time that $[M_4]$ (or any other assumption ensuring $[\mathbf{V'}]$) holds.

Let us now assume that there exists $\theta > 0$ such that the data $X_1, ..., X_n$ are geometrically $[\mathbf{GEO}(\theta)]$ β-mixing. We still assume $[M_1]$- $[M_5]$ on the models. Let $p \geq n(\ln n)^{-2}/2$, $q \geq (\ln n)^2/2$ such that $2pq = n$. Then there exist constants c_e, c_M such that, for all m, m' in $\mathcal{M}_n$,

$$p\beta_q \leq c_M \frac{n}{(\ln n)^2} e^{-\frac{\theta}{2}(\ln n)^2},$$

$$e_{A,m,m'} \leq c_e \frac{(\ln n)^4}{n}(d_m \vee d_{m'}) \leq \frac{c_e}{c'_D}\frac{(\ln n)^{4+2\gamma}}{n}\frac{R_{A,m} \vee R_{A,m'}}{(\ln n)^{2\gamma}}.$$

Thus, we deduce from the previous computations the following corollary.

Corollary 4.4.4 *Let $\mathcal{M}_n$ be a collection of models satisfying $[M_1]$- $[M_5]$. Assume that the process $(X_n)_{n\in\mathbb{Z}}$ is strictly stationary and geometrically* $[\mathbf{GEO}(\theta)]$ *β-mixing with mixing rate $\theta > 0$. Let $\tilde{s}$ be the estimator defined in (4.9) with a resampling penalty (4.8). Let $\epsilon_n^* = \left(r_n^{-1/8} \vee n^{-1/4}(\ln n)^{1+\gamma/2}\right)$ and $\theta_1 = \theta \wedge 1$.*
There exist constants $C > 0$ and $\kappa > 0$ such that

$$\mathbb{P}\left(\|\tilde{s} - s\|_2^2 > (1 + \kappa\epsilon_n^*) \inf_{m\in\mathcal{M}_n} \|s - \hat{s}_{A,m}\|_2^2\right) \leq C\frac{n}{(\ln n)^2}e^{-\frac{\theta_1}{2}(\ln n)^2}$$

Comments : Under the stronger assumption that the process is geometrically β-mixing, we recover the same results as in the independent case. The rate of convergence is essentially given by $r_n^{-1/8}$, it was $r_n^{-1/4}$ in the independent case and the deviation probability is upper bounded by $Cn(\ln n)^{-2}e^{-\frac{\theta_1}{2}(\ln n)^2}$ instead of $Ce^{-\frac{1}{2}(\ln n)^2}$.

4.4.2 τ-mixing processes

Our results for τ-mixing processes do not apply to general collections of models as mentioned before. We give in this section a classical collection where they might be used.

Dyadic Wavelet spaces:

This collection was the one of Chapter 3. Wavelet spaces are classically considered because the oracle is adaptive over Besov spaces (see for example Birgé & Massart [15] or Chapter 3). Hereafter, r is a real number, $r \geq 1$ and we work with an r-regular orthonormal multiresolution analysis of $L^2(\mu)$, associated with a compactly supported scaling function ϕ and a compactly supported mother wavelet ψ. Without loss of generality, we suppose that the support of the functions ϕ and ψ is included in an interval $[A_1, A_2)$ where A_1 and A_2 are integers such that $A_2 - A_1 = A \geq 1$.

For all functions t in $L^2(\mu)$, we denote by $\|t\|_{BV}$ its bounded variation semi-norm, that is

$$\|t\|_{BV} = \sup_{l\in\mathbb{N}^*} \sup_{-\infty<a_1<\ldots<a_l<+\infty} \sum_{j=1}^{l-1} |t(a_{j+1}) - t(a_j)|.$$

For all k in $\mathbb{Z}$ and j in $\mathbb{N}^*$, let $\psi_{0,k} : \ x \to \sqrt{2}\phi(2x - k)$ and $\psi_{j,k} : \ x \to 2^{j/2}\psi(2^j x - k)$. The family $\{(\psi_{j,k})_{j\geq 0,k\in\mathbb{Z}}\}$ is an orthonormal basis of $L^2(\mu)$. Let us recall the following inequalities: let $K_\infty = (\sqrt{2}\|\phi\|_\infty) \vee \|\psi\|_\infty$, $K_L = (2\sqrt{2}\mathrm{Lip}(\phi)) \vee \mathrm{Lip}(\psi)$, $K_{BV} = AK_L$.

Then for all $j \geq 0$, we have $\|\psi_{j,k}\|_\infty \leq K_\infty 2^{j/2}$,

$$\left\|\sum_{k\in\mathbb{Z}} |\psi_{j,k}|\right\|_\infty \leq AK_\infty 2^{j/2} \tag{4.20}$$

$$\mathrm{Lip}(\psi_{j,k}) \leq K_L 2^{3j/2}, \tag{4.21}$$

$$\|\psi_{j,k}\|_{BV} \leq K_{BV}2^{j/2},. \tag{4.22}$$

We assume that $\mathcal{M}_n$ is the following collection.

[W] *dyadic wavelet generated spaces: let* $J_n = [\ln(n)/\ln(2)]$, *for all* $J_m = 1, ..., J_n$, *let*

$$m = \{(j,k),\ 0 \leq j \leq J_m,\ k \in \mathbb{Z}\}$$

and let S_m *be the linear span of* $\{\psi_{j,k}\}_{(j,k)\in m}$.

It is a classical result (see for example Birgé & Massart [15]) that this collection satisfies the following Assumptions

[T1] *for all* $m \in \mathcal{M}_n$, $2^{J_m} \leq n$*;*
[T2] *there exists a constant* Φ *such that*

$$\forall m, m' \in \mathcal{M}_n, \forall t \in S_m, \forall t' \in S_{m'}, \|t+t'\|_\infty \leq \Phi 2^{(J_m \vee J_{m'})/2}\|t+t'\|_2;$$

[T3] $|\mathcal{M}_n| \leq \ln n/\ln 2$.

Under these Assumptions, the following Lemma hold.

Lemma 4.4.5 *Let* $\theta > 2$ *and assume that* $X_1, ..., X_n$ *are arithmetically* $[\mathbf{AR}(\theta)]$ τ*-mixing and let* $u = 3/(1+\theta) \wedge 1$. *Let* $\mathcal{M}_n$ *be a collection of regular wavelet spaces* **[W]**. *There exist constants* c_D, c_v, c_b *such that, for all* m, m' *in* $\mathcal{M}_n$,

$$D_{A,m} \leq c_D 2^{J_m},\ v^2_{A,m,m'} \leq c_v \left(2^{J_m \vee J_{m'}}\right)^{\frac{1}{2}(1+u)},\ b^2_{A,m,m'} \leq c_b 2^{J_m \vee J_{m'}}.$$

Moreover, $MC_n \leq c_T n^2$.

Hereafter, u denotes the real number defined in Lemma 4.4.5, that is

$$u = \frac{3}{1+\theta} \wedge 1.$$

As in the previous section, we add extra assumptions to prove **[V']**.

[T4] *There exists a constant* $c'_D > 0$ *such that, for all* $n \in \mathbb{N}^*$, *for all* m *in* $\mathcal{M}_n$,

$$D_{A,m} \geq c'_D 2^{J_m}.$$

[T5] *There exist a sequence* $r_n \to \infty$ *and a constant* $\gamma > 1$ *such that,*

$$R_n (\ln n)^{-\frac{2\gamma}{1-u}} \geq r_n.$$

Without loss of generality, assume that $\gamma \leq 3/2$ in **[T5]** and that there exists $\theta > 2$ such that $X_1, ..., X_n$ are arithmetically $[\mathbf{AR}(\theta)]$ τ-mixing. Choose $p \geq \sqrt{n}(\ln n)^2/2$, $q \geq \sqrt{n}(\ln n)^{-2}/2$ such that $2pq = n$. Then, $u < 1$ and there exists constants $c_T^{(2)}$, c_e such that

$$\tau_q MC_n \leq c_T^{(2)} \frac{(\ln n)^{2(1+\theta)}}{n^{(\theta-3)/2}}, e_{A,m,m'} \leq \frac{c_e}{\ln n} \frac{R_{A,m} \vee R_{A,m'}}{(\ln n)^{2\gamma}}.$$

When $2^{J_m \vee J_{m'}} \leq r_n (\ln n)^{\frac{2\gamma}{1-u}}$,

$$v^2_{A,m,m'} \leq c_v \left(r_n (\ln n)^{\frac{2\gamma}{1-u}}\right)^{\frac{1}{2}(1+u)} \leq c_v r_n^{-\frac{1-u}{2}} \frac{R_n}{(\ln n)^\gamma} \leq c_v \frac{R_{A,m} \vee R_{A,m'}}{r_n^{\frac{1-u}{2}} (\ln n)^\gamma}.$$

When $2^{J_m \vee J_{m'}} \geq r_n (\ln n)^{\frac{2\gamma}{1-u}}$,

$$v^2_{A,m,m'} \leq \frac{c_v}{c'_D} \frac{D_{A,m} \vee D_{A,m'}}{\left(r_n (\ln n)^{\frac{2\gamma}{1-u}}\right)^{\frac{1-u}{2}}} \leq \frac{c_v}{c'_D} r_n^{-\frac{1-u}{2}} \frac{R_{A,m} \vee R_{A,m'}}{(\ln n)^\gamma}.$$

As in the β-mixing case, we deduce the following corollary.

Corollary 4.4.6 *Assume that the process $(X_n)_{n\in\mathbb{Z}}$ is strictly stationary and arithmetically* [**AR**(θ)] *τ-mixing with mixing rate $\theta > 2$. Let $\mathcal{M}_n$ be a collection of regular wavelet spaces* [**W**] *and assume moreover that* [**T4**], [**T5**] *hold. Let $\tilde{s}$ be the estimator defined in (4.9) with a resampling penalty (4.8). Let $\epsilon^*_n = (\ln n \wedge r_n^{1-u})^{-1/4}$. There exist constants $C > 0$ and $\kappa > 0$ such that*

$$\mathbb{E}\left(\|\tilde{s} - s\|_2^2\right) \leq (1 + \kappa\epsilon^*_n)\mathbb{E}\left(\inf_{m\in\mathcal{M}_n} \|s - \hat{s}_{A,m}\|_2^2\right) + C\frac{(\ln n)^{2(1+\theta)}}{n^{(\theta-3)/2}}.$$

Comments:

- With a mixing rate $\theta > 5$, the estimator selected by a resampling penalty satisfies an oracle inequality (4.4). This result can be compared with Corollary 4.4.1. When the data are τ-mixing, we do not obtain a trajectorial oracle inequality (4.3) and the condition on the mixing rate is stronger than in the β-mixing case, but, as mentioned in the introduction, this result is very interesting because there is a lot of examples of processes that are τ-mixing and not β-mixing.

- Assumption [**T5**] is hard to check in practice but it can be removed as in the β-mixing, provided that we only consider models with dimension larger than $c_M(\ln n)^\eta$ for some well chosen constants c_M and η.

- We can get better rates of convergence if we assume that the process is geometrically τ-mixing and if we choose p and q as in Corollary 4.4.4.

This result can also be compared with our result on τ-mixing processes proved in Chapter 3. Let us recall here this result.

Theorem 4.4.7 *Let $(X_n)_{n\in\mathbb{Z}}$ be a strictly stationary process and assume that there exists $\theta > 5$ such that it is arithmetically* [**AR**](θ) *τ-mixing. Let $\mathcal{M}_n$ be a collection of wavelet models satisfying* [**W**]. *Let $\tilde{s}$ be the estimator defined in (4.9) with*

$$pen(m) = Kc_D\frac{D_m}{n}, \text{ where } K > 4.$$

Then there exist constants $c_0 > 0, \gamma_1 > 0$ and a sequence $\epsilon_n \to 0$ such that

$$\mathbb{E}\left(\|\tilde{s} - s\|_2^2\right) \leq (1 + \epsilon_n)\left(\inf_{m\in\mathcal{M}_n,\, D_m \geq c_0(\log n)^{\gamma_1}} \|s - s_m\|_2^2 + pen(m)\right).$$

Comments: As in the β-mixing case, the main improvement of Corollary 4.4.6 is that the new procedure is totally data driven. Moreover the risk of the selected estimator is compared with the oracle in Corollary 4.4.6 whereas it is compared with an upper bound on $\inf_{m\in\mathcal{M}_n} \left\{\|s - s_m\|^2 + 2\mathbb{E}\left(\|s_m - \hat{s}_{A,m}\|^2\right)\right\}$ in Theorem 4.4.7. Therefore, our new procedure improve the one given in Chapter 3 every time that Assumption [**T4**] or any other Assumption ensuring [**V'**] holds. [**T4**] is not necessary in Theorem 4.4.7.

4.5 Proofs

This section is devoted to the proof of the main results. Let us give some notations that we will use repeatedly all along the proofs.
For all $k = 0, ..., p-1$, let $I_k = (2kq+1, ..., (2k+1)q)$, $A_k = (X_i)_{i\in I_k}$ and $I = \cup_{k=0}^{p-1} I_k$.
For all functions t in $L^2(\mu)$ and all $x_1, ..., x_q$ in $\mathbb{X}$, let

$$L_q(t)(x_1, ..., x_q) = \frac{1}{q}\sum_{i=1}^{q} t(x_i),\ P_A t = \frac{1}{p}\sum_{k=0}^{p-1} L_q(t)(A_k) = \frac{2}{n}\sum_{i\in I} t(X_i),$$

$$\nu_A(t) = (P_A - P)(t).$$

The estimator $\hat{s}_{A,m}$ associated to the model S_m, is then defined as

$$\hat{s}_{A,m} \in \arg\min_{t\in S_m} P_A Q(t).$$

For all m, m' in $\mathcal{M}_n$, let

$$T_m = \sum_{\lambda\in m}(L_q(\psi_\lambda) - P\psi_\lambda)^2,$$

$$U_m = \frac{1}{p(p-1)}\sum_{i\neq j=0}^{p-1}\sum_{\lambda\in m}(L_q(\psi_\lambda)(A_i) - P\psi_\lambda)(L_q(\psi_\lambda)(A_j) - P\psi_\lambda),$$

$$p(m) = \|s_m - \hat{s}_{A,m}\|^2 = \sup_{t\in B_m}(\nu_A(t))^2 = \sum_{\lambda\in m}(\nu_A(\psi_\lambda))^2.$$

$$p_W(m) = \frac{1}{v_W^2}\sum_{\lambda\in m}\mathbb{E}^W\left((\nu_A^W(\psi_\lambda))^2\right),\ \delta(m, m') = 2\nu_A(s_m - s_{m'}).$$

Lemma 4.6.2 applied with $n = p$, $\Lambda = m$, $t_\lambda = L_q(\psi_\lambda)$, $X_i = A_{i-1}$, gives

$$p_W(m) = \tfrac{1}{p}(P_A(T_m) - U_m) \tag{4.23}$$

$$p(m) - p_W(m) = U_m, \tag{4.24}$$

where $P_A(T_m) = \sum_{k=0}^{p-1} T_m(A_k)/p$.
For all functional $T = F(A_0, ..., A_{p-1})$, let $T^* = F(A_0^*, ..., A_{p-1}^*)$, where the random variables (A_k^*) are given by the coupling Lemmas given in Section 4.2.3. In particular, we will use repeatedly the notations P_A^*, ν_A^*, U_m^*, $p^*(m)$, $p_W^*(m)$, $\delta^*(m, m')$.
For all function t of $L^2(\mu)$, for all r in $\mathbb{N}$ and all $x_1, ..., x_r, y_1, ...y_r$ in $\mathbb{X}$,

$$\begin{aligned} |L_r(t)(x_1, ..., x_r) - L_r(t)(y_1, ..., y_r)| &\leq \frac{1}{r}\sum_{i=1}^{r}|t(x_i) - t(y_i)| \\ &\leq \frac{1}{r}\mathrm{Lip}_d(t) d_r((x_1, ..., x_r), (y_1, ..., y_r)). \end{aligned}$$

Thus $\mathrm{Lip}_{d_r}(L_r(t)) \leq \mathrm{Lip}_d(t)/r$.
For all $k \in \mathbb{N}$, $\mathcal{M}_n^k = \{m \in \mathcal{M}_n,\ R_{A,m} \in [k, k+1)\}$ and for all n in $\mathbb{N}$ and, for all $k > 0$, $k' > 0$ and $\gamma \geq 0$, let

$$l_{n,\gamma}(k, k') = \ln\Big((1 + \mathrm{Card}(\mathcal{M}_n^{[k]}))(1 + \mathrm{Card}(\mathcal{M}_n^{[k']}))(k+1)(k'+1)\Big) + (\ln n)^\gamma.$$

For all m, m' in $\mathcal{M}_n$, let $l_{m,m'} = l_{n,\gamma}(R_{A,m}, R_{A,m'})$. From Lemma 4.6.1, for all $K > 1$, there exists a constant $C > 0$ such that

$$\sum_{(m,m')\in(\mathcal{M}_n)^2} e^{-Kl_{m,m'}} = Ce^{-K(\ln n)^\gamma}.$$

Under **[V']**,

$$\sup_{(m,m')\in(\mathcal{M}_n)^2} \left\{ \left(\left(\frac{v^2_{A,m,m'}}{R_{A,m} \vee R_{A,m'}} \right)^2 \vee \frac{e_{A,m,m'}}{R_{A,m} \vee R_{A,m'}} \right) l^2_{m,m'} \right\} \leq \epsilon^4_n.$$

We will now state and prove some technical lemmas. Lemmas 4.5.1 and 4.5.2 are coupling lemmas. They allow to work with $p^*(m)$, $p^*_W(m)$, $\delta^*(m,m')$ instead of $p(m)$, $p_W(m)$, $\delta(m,m')$. Lemma 4.5.3 is a consequence of our study of the independent case. It allows to extend the proofs of Chapter 2 to the mixing case. It is the main tool of this paper.

Lemma 4.5.1 *Let $X_1, ..., X_n$ be stationary random variables, real valued and β-mixing. Let p and q be two integers such that $2pq = n$ and let $A^*_0, ..., A^*_{p-1}$ be the random variables given by Viennet's Lemma in Section 4.2.3. There exists an event Ω_C such that $\mathbb{P}(\Omega^c_C) \leq p\beta_q$ and such that, on Ω_C, for all m, m' in $\mathcal{M}_n$, we have*

$$p(m) = p^*(m),\ p_W(m) = p^*_W(m),\ \delta(m,m') = \delta^*(m,m'). \tag{4.25}$$

Proof :
Let $\Omega_C = \{\forall k = 0, ..., p-1,\ A_k = A^*_k\}$. It comes from Viennet's Lemma that $\mathbb{P}(\Omega^c_C) \leq p\beta_q$ and it is clear that, on Ω_C, (4.25) holds.

Lemma 4.5.2 *Let $X_1, ..., X_n$ be stationary random variables, real valued, τ-mixing and with common density s. Let p and q be two integers such that $2pq = n$ and let $A^*_0, ..., A^*_{p-1}$ be the random variables given by the coupling's Lemma in Section 4.2.3. Let $\mathcal{M}_n$ be a collection of models and let MC_n be the associated mixing complexity defined by*

$$MC_n = \sum_{m\in\mathcal{M}_n} \left(\left\| \sum_{\lambda\in m} |\psi_\lambda| \right\|_\infty \sup_{\lambda\in m} Lip_d(\psi_\lambda) + \|s\| |\mathcal{M}_n| \sup_{t\in B_m} Lip(t) \right).$$

For all m, m' in $\mathcal{M}_n$,

$$\mathbb{E}\left(\sup_{m\in\mathcal{M}_n} |p(m) - p^*(m)| \right) \leq 4\tau_q MC_n \tag{4.26}$$

$$\mathbb{E}\left(\sup_{m\in\mathcal{M}_n} |p_W(m) - p^*_W(m)| \right) \leq \frac{8\tau_q}{p} MC_n \tag{4.27}$$

$$\mathbb{E}\left(\sup_{m,m'\in\mathcal{M}_n} \delta(m,m') - \delta^*(m,m') \right) \leq 4\tau_q MC_n. \tag{4.28}$$

Proof :
For all m in $\mathcal{M}_n$, we have

$$\mathbb{E}\left(\sup_{m\in\mathcal{M}_n} |p(m) - p^*(m)| \right) \leq \sum_{m\in\mathcal{M}_n} \mathbb{E}\left(|p(m) - p^*(m)| \right).$$

Moreover, for all m in $\mathcal{M}_n$,

$$\begin{aligned}
|p(m)-p^*(m)| &= \left|\sum_{\lambda\in m}((P_A-P)\psi_\lambda)^2-((P_A^*-P)\psi_\lambda)^2\right|\\
&= \left|\sum_{\lambda\in m}((\nu_A+\nu_A^*)\psi_\lambda)\left((P_A-P_A^*)\psi_\lambda\right)\right|\\
&\leq \sum_{\lambda\in m}|(\nu_A+\nu_A^*)\psi_\lambda|\frac{1}{p}\sum_{k=0}^{p-1}|L_q(\psi_\lambda)(A_k)-L_q(\psi_\lambda)(A_k^*)|\\
&\leq 4\left\|\sum_{\lambda\in m}|\psi_\lambda|\right\|_\infty \sup_{\lambda\in m}\mathrm{Lip}_{d_q}(L_q(\psi_\lambda))\frac{1}{p}\sum_{k=0}^{p-1}d_q(A_k,A_k^*)\\
&\leq \frac{4}{q}\left\|\sum_{\lambda\in m}|\psi_\lambda|\right\|_\infty \sup_{\lambda\in m}\mathrm{Lip}_{d}(\psi_\lambda)\frac{1}{p}\sum_{k=0}^{p-1}d_q(A_k,A_k^*).
\end{aligned}$$

We take the expectation in this last inequality and we use the τ-coupling Lemma of Section 4.2.3 to obtain (4.26).
From (4.23), we have

$$|p_W(m)-p_W^*(m)|=\frac{1}{p}\left|(P_A-P_A^*)(T_m)-(U_m-U_m^*)\right|.$$

We have

$$\begin{aligned}
&(P_A-P_A^*)T_m=\\
&\sum_{\lambda\in m}\frac{1}{p}\sum_{k=0}^{p-1}\left(L_q(\psi_\lambda)(A_k)-L_q(\psi_\lambda)(A_k^*)\right)\left(L_q(\psi_\lambda)(A_k)+L_q(\psi_\lambda)(A_k^*)-2P\psi_\lambda\right),
\end{aligned}$$

thus

$$\begin{aligned}
|(P_A-P_A^*)T_m| &= 4\left\|\sum_{\lambda\in m}|\psi_\lambda|\right\|_\infty \sup_{\lambda\in m}\mathrm{Lip}_{d_q}(L_q(\psi_\lambda))\frac{1}{p}\sum_{k=0}^{p-1}d_q(A_k,A_k^*)\\
&\leq \frac{4}{q}\left\|\sum_{\lambda\in m}|\psi_\lambda|\right\|_\infty \sup_{\lambda\in m}\mathrm{Lip}_{d}(\psi_\lambda)\frac{1}{p}\sum_{k=0}^{p-1}d_q(A_k,A_k^*).
\end{aligned}$$

Moreover

$$\begin{aligned}
U_m-U_m^* &= \frac{1}{p(p-1)}\sum_{i\neq j=0}^{p-1}\sum_{\lambda\in m}(L_q(\psi_\lambda(A_j))-P\psi_\lambda)(L_q(\psi_\lambda(A_i))-L_q(\psi_\lambda(A_i^*)))\\
&+\frac{1}{p(p-1)}\sum_{i\neq j=0}^{p-1}\sum_{\lambda\in m}(L_q(\psi_\lambda(A_i^*))-P\psi_\lambda)(L_q(\psi_\lambda(A_j))-L_q(\psi_\lambda(A_j^*))),
\end{aligned}$$

thus

$$|U_m-U_m^*|\leq\frac{4}{q}\left\|\sum_{\lambda\in m}|\psi_\lambda|\right\|_\infty \sup_{\lambda\in m}\mathrm{Lip}_{d}(\psi_\lambda)\frac{1}{p}\sum_{k=0}^{p-1}d_q(A_k,A_k^*).$$

Therefore,

$$\begin{aligned}\mathbb{E}\left(\left|p_W(m)-p_W^*(m)\right|\right) &\leq \frac{8}{pq}\left\|\sum_{\lambda\in m}|\psi_\lambda|\right\|_\infty \sup_{\lambda\in m}\mathrm{Lip}_d(\psi_\lambda)\frac{1}{p}\sum_{k=0}^{p-1}\mathbb{E}(d_q(A_k,A_k^*))\\ &\leq \frac{8\tau_q}{p}\left\|\sum_{\lambda\in m}|\psi_\lambda|\right\|_\infty \sup_{\lambda\in m}\mathrm{Lip}_d(\psi_\lambda).\end{aligned}$$

Thus

$$\begin{aligned}\mathbb{E}\left(\sup_{m\in\mathcal{M}_n}\left|p_W(m)-p_W^*(m)\right|\right) &\leq \sum_{m\in\mathcal{M}_n}\mathbb{E}\left(\left|p_W(m)-p_W^*(m)\right|\right)\\ &\leq \frac{8\tau_q}{p}\sum_{m\in\mathcal{M}_n}\left\|\sum_{\lambda\in m}|\psi_\lambda|\right\|_\infty \sup_{\lambda\in m}\mathrm{Lip}_d(\psi_\lambda).\end{aligned}$$

Finally,

$$\mathbb{E}\left(\sup_{m,m'\in\mathcal{M}_n}\delta(m,m')-\delta^*(m,m')\right)\leq\sum_{m,m'\in\mathcal{M}_n}\mathbb{E}\left(\left|\delta(m,m')-\delta^*(m,m')\right|\right)$$

and, for all m, m' in $\mathcal{M}_n$,

$$\begin{aligned}\mathbb{E}\left(\left|\delta(m,m')-\delta^*(m,m')\right|\right) &= 2\mathbb{E}\left(\left|(P_A-P_A^*)(s_m-s_{m'})\right|\right)\\ &\leq \frac{2}{pq}\mathrm{Lip}_d(s_m-s_{m'})\sum_{k=0}^{p-1}\mathbb{E}\left(d_q(A_k,A_k^*)\right)\\ &\leq 2\tau_q\mathrm{Lip}_d(s_m-s_{m'}).\end{aligned}$$

For all x,y in $\mathbb{X}$ and all m,m' in $\mathcal{M}_n$,

$$(s_m-s_{m'})(x)-(s_m-s_{m'})(y)\leq\|s\|\left(\sup_{t\in B_m}\mathrm{Lip}(t)+\sup_{t\in B_{m'}}\mathrm{Lip}(t)\right)d(x,y)$$

Hence, $\mathrm{Lip}_d(s_m-s_{m'})\leq\|s\|\left(\sup_{t\in B_m}\mathrm{Lip}(t)+\sup_{t\in B_{m'}}\mathrm{Lip}(t)\right)$, thus

$$\mathbb{E}\left(\sup_{m,m'\in\mathcal{M}_n}\delta(m,m')-\delta^*(m,m')\right)\leq 4\tau_q\|s\||\mathcal{M}_n|\sum_{m\in\mathcal{M}_n}\sup_{t\in B_m}\mathrm{Lip}(t).$$

Let us now derive some consequences of the results of Chapter 2.

Lemma 4.5.3 *Let $\mathcal{M}_n$ be a collection of models satisfying* **[V']**. *Then there exists a constant $C>0$ such that*

$$\mathbb{P}\left(\bigcup_{m\in\mathcal{M}_n}\left\{p^*(m)-\frac{2D_{A,m}}{n}>15\epsilon_n\frac{R_{A,m}}{n}\right\}\right)\leq Ce^{-\frac{1}{2}(\ln n)^\gamma},\tag{4.29}$$

$$\mathbb{P}\left(\bigcup_{m\in\mathcal{M}_n}\left\{p^*(m)-\frac{2D_{A,m}}{n}<-25\epsilon_n\frac{R_{A,m}}{n}\right\}\right)\leq Ce^{-\frac{1}{2}(\ln n)^\gamma}.\tag{4.30}$$

$$\mathbb{P}\left(\bigcup_{m\in\mathcal{M}_n}\left\{p^*(m)-p^*_W(m)>15\epsilon_n\frac{R_{A,m}}{n}\right\}\right)\leq Ce^{-\frac{1}{2}(\ln n)^\gamma}, \tag{4.31}$$

$$\mathbb{P}\left(\bigcup_{m\in\mathcal{M}_n}\left\{p^*(m)-p^*_W(m)<-25\epsilon_n\frac{R_{A,m}}{n}\right\}\right)\leq Ce^{-\frac{1}{2}(\ln n)^\gamma}. \tag{4.32}$$

$$\mathbb{P}\left(\bigcup_{m,m'\in\mathcal{M}_n}\left\{\delta^*(m,m')>12\epsilon_n\left(\frac{R_{A,m}\vee R_{A,m'}}{n}\right)\right\}\right)\leq Ce^{-(\ln n)^\gamma}. \tag{4.33}$$

There exists an absolute constant $C>0$ such that

$$\mathbb{E}\left(\sup_{m\in\mathcal{M}_n}\left(p^*(m)-\frac{2D_{A,m}}{n}-15\epsilon_n\frac{R_{A,m}}{n}\right)_+\right)\leq Ce^{-\frac{1}{2}(\ln n)^\gamma}. \tag{4.34}$$

$$\mathbb{E}\left(\sup_{m\in\mathcal{M}_n}\left(-p^*(m)+\frac{2D_{A,m}}{n}-35\epsilon_n\frac{R_{A,m}}{n}\right)_+>\right)\leq Ce^{-\frac{1}{2}(\ln n)^\gamma}. \tag{4.35}$$

$$\mathbb{E}\left(\sup_{m\in\mathcal{M}_n}\left(p^*(m)-p^*_W(m)-20\epsilon_n\frac{R_{A,m}}{n}\right)_+\right)\leq Ce^{-\frac{1}{2}(\ln n)^\gamma}. \tag{4.36}$$

$$\mathbb{E}\left(\sup_{m\in\mathcal{M}_n}\left(-p^*(m)+p^*_W(m)-35\epsilon_n\frac{R_{A,m}}{n}\right)_+\right)\leq Ce^{-\frac{1}{2}(\ln n)^\gamma}. \tag{4.37}$$

$$\mathbb{E}\left(\sup_{m,m'\in\mathcal{M}_n}\left(\delta^*(m,m')-20\epsilon_n\left(\frac{R_{A,m}\vee R_{A,m'}}{n}\right)\right)_+\right)\leq Ce^{-\frac{1}{2}(\ln n)^\gamma}. \tag{4.38}$$

Proof of the concentration inequalities :
$p^*(m)=\sup_{t\in B_m}((\nu^*_A)(t))^2$ and $A^*_0,...,A^*_{p-1}$ are independent. Thus

$$\mathbb{E}(p^*(m))=\sum_{\lambda\in m}\frac{\mathrm{Var}\left(L_q(\psi_\lambda)(A_0)\right)}{p}=\frac{2D_{A,m}}{n},$$

$$\sup_{t\in B_m}\mathrm{Var}\left(L_q(t)(A_0)\right)=\frac{v^2_{A,m}}{q},\quad\frac{\sup_{t\in B_m}\left\|L_q(t)\right\|^2_\infty}{p}\leq\frac{e_{A,m}}{q}.$$

We apply Proposition 4.6.3 in the Appendix with $B=\{L_q(t),t\in B_m\}$, $D=D_{A,m}/q$, $v^2=v^2_{A,m}/q$, $\epsilon=e_{A,m}/q$ and $n=p$. For all $x>0$ and all m in $\mathcal{M}_n$, with probability larger than $1-e^{-x}$

$$p^*(m)-\frac{2D_{A,m}}{n}\leq\frac{2D^{3/4}_{A,m}(e_{A,m}(19x)^2)^{1/4}+6\sqrt{D_{A,m}v^2_{A,m}x}+6v^2_{A,m}x+2e_{A,m}(19x)^2}{n}$$

and, with probability larger than $1-2.8e^{-x}$

$$\frac{2D_{A,m}}{n}-p^*(m)\leq\frac{16D^{3/4}_{A,m}(e_{A,m}x^2)^{1/4}+15.22\sqrt{D_{A,m}v^2_{A,m}x}+2e_{A,m}(40.25x)^2}{n}. \tag{4.39}$$

Let $K>0$ be a constant to be chosen later, let $l_m=l_{n,\gamma}(R_m,R_m)$, and let $x=K^2l_m$. From **[V']** applied with $m=m'$, since $D_{A,m}\leq R_{A,m}$,

$$v_{A,m}^2x\leq (K\epsilon_n)^2R_{A,m},\ e_{A,m}x^2\leq (K\epsilon_n)^4R_{A,m},$$

$$D_{A,m}^{3/4}(e_{A,m}x^2)^{1/4}\leq K\epsilon_nR_{A,m},\ \sqrt{D_{A,m}v_{A,m}^2x}\leq K\epsilon_nR_m. \tag{4.40}$$

Let $e_n(K)=(2\sqrt{19}+6)K+6K^2\epsilon_n+2(19)^2K^4\epsilon_n^3$, from (4.40),

$$\frac{2D_{A,m}^{3/4}(e_{A,m}(19x)^2)^{1/4}+6\sqrt{D_{A,m}v_{A,m}^2x}+6v_{A,m}^2x+2e_{A,m}(19x)^2}{n}\leq e_n(K)\epsilon_n\frac{R_m}{n}.$$

Thus, from Lemma 4.6.1, for all $K>1/\sqrt{2}$, there exists a constant $C>0$ such that

$$\mathbb{P}\left(\bigcup_{m\in\mathcal{M}_n}\left\{\frac{2D_{A,m}}{n}-p^*(m)>e_n(K)\epsilon_n\frac{R_{A,m}}{n}\right\}\right)\leq$$
$$\sum_{m\in\mathcal{M}_n}\mathbb{P}\left(p^*(m)-\frac{2D_{A,m}}{n}>e_n(K)\epsilon_n\frac{R_{A,m}}{n}\right)\leq\sum_{m\in\mathcal{M}_n}e^{-K^2l_m}\leq Ce^{-K^2(\ln n)^\gamma}.$$

Let $K=11/(2\sqrt{19}+6)>1/\sqrt{2}$ and choose n sufficiently large such that $6K^2\epsilon_n+2(19)^2K^4\epsilon_n^3\leq 4$, then $e_n(K)\leq 15$ and (4.29) holds for all n sufficiently large. It holds for all n provided that we enlarge C if necessary.
Let $e_n^{(2)}(K)=31,22K+2(40,25)^2K^4\epsilon_n^3$, from (4.40),

$$\frac{16D_{A,m}^{3/4}(e_{A,m}x^2)^{1/4}+15.22\sqrt{D_{A,m}v_{A,m}^2x}+2e_{A,m}(40.25x)^2}{n}\leq e_n^{(2)}(K)\epsilon_n\frac{R_{A,m}}{n}.$$

We apply inequality (4.39) with $x=K^2l_m$. For all $K>1/\sqrt{2}$, there exists a constant $C>0$ such that

$$\mathbb{P}\left(\bigcup_{m\in\mathcal{M}_n}\left\{p^*(m)-\frac{2D_{A,m}}{n}<-e_n^{(2)}(K)\epsilon_n\frac{R_{A,m}}{n}\right\}\right)\leq$$
$$\sum_{m\in\mathcal{M}_n}\mathbb{P}\left(p^*(m)-\frac{2D_{A,m}}{n}<-e_n^{(2)}(K)\epsilon_n\frac{R_{A,m}}{n}\right)\leq 2.8\sum_{m\in\mathcal{M}_n}e^{-K^2l_m}\leq Ce^{-K^2(\ln n)^\gamma}.$$

Take $K=23/31.22>1/\sqrt{2}$ and n sufficiently large to have $2(40,25)^2K^4\epsilon_n^3\leq 2$, then $e_n^{(2)}(K)\leq 25$ and (4.30) holds for sufficiently large n. It holds then in general, provided that we enlarge the constant C if necessary.
From (4.24), $p^*(m)-p_W^*(m)=U_m^*$. Therefore, from Lemma 4.6.4 in the appendix, for all m in $\mathcal{M}_n$ and all $x>0$, with probability larger than $1-2e^{-x}$,

$$p^*(m)-p_W^*(m)\leq\frac{10.62D_{A,m}^{3/4}(e_{A,m}x^2)^{1/4}+6\sqrt{v_{A,m}^2D_{A,m}x}+6v_{A,m}^2x+2e_{A,m}(19.1x)^2}{n-1}, \tag{4.41}$$

and, with probability larger than $1-3.8e^{-x}$,

$$p_W^*(m)-p^*(m) > \frac{18D_{A,m}^{3/4}(e_{A,m}x^2)^{1/4}+15.22\sqrt{v_{A,m}^2D_{A,m}x}+2e_{A,m}(40.3x)^2}{n-1}. \quad (4.42)$$

Let $K>0$, $e_n^{(3)}(K)=(16.62K+6K^2\epsilon_n+2(19.1)^2K^4\epsilon_n^3)n/(n-1)$ and $x=K^2l_m$, from (4.40),

$$\frac{10.62D_{A,m}^{3/4}(e_{A,m}x^2)^{1/4}+6\sqrt{v_{A,m}^2D_{A,m}x}+6v_{A,m}^2x+2e_{A,m}(19.1x)^2}{n-1}\leq e_n^{(3)}(K)\epsilon_n\frac{R_{A,m}}{n}.$$

We apply (4.41) with $x=K^2l_m$. From Lemma 4.6.1, for all $K>1/\sqrt{2}$, there exists a constant C such that

$$\mathbb{P}\left(\bigcup_{m\in\mathcal{M}_n}\left\{p^*(m)-p_W^*(m)>e_n^{(3)}(K)\epsilon_n\frac{R_{A,m}}{n}\right\}\right) \leq$$
$$\sum_{m\in\mathcal{M}_n}\mathbb{P}\left(p^*(m)-p_W^*(m)>e_n^{(3)}(K)\epsilon_n\frac{R_{A,m}}{n}\right) \leq 2\sum_{m\in\mathcal{M}_n}e^{-K^2l_m}\leq Ce^{-K^2(\ln n)^\gamma}.$$

Take $K=12/16.62>1/\sqrt{2}$ and $n\geq 15$ such that $6K^2\epsilon_n+2(19.1)^2K^4\epsilon_n^3\leq 2$, then $e_n^{(3)}(K)\leq 15$ and (4.31) holds for sufficiently large n. It holds in general provided that we enlarge C if necessary.

Let $K>0$, $e_n^{(4)}(K)=(33.22K+2(40.3)^2K^4\epsilon_n^3)n/(n-1)$ and $x=K^2l_m$. From (4.40),

$$\frac{18D_{A,m}^{3/4}(e_{A,m}x^2)^{1/4}+15.22\sqrt{v_{A,m}^2D_{A,m}x}+2e_{A,m}(40.3x)^2}{n-1}\leq e_n^{(4)}(K)\epsilon_n\frac{R_{A,m}}{n}.$$

We apply (4.41) with $x=K^2l_m$. From Lemma 4.6.1, for all $K>1/\sqrt{2}$, there exists a constant C such that

$$\mathbb{P}\left(\bigcup_{m\in\mathcal{M}_n}\left\{p_W^*(m)-p^*(m)>e_n^{(4)}(K)\epsilon_n\frac{R_{A,m}}{n}\right\}\right) \leq$$
$$\sum_{m\in\mathcal{M}_n}\mathbb{P}\left(p_W^*(m)-p^*(m)>e_n^{(4)}(K)\epsilon_n\frac{R_{A,m}}{n}\right) \leq 3.8\sum_{m\in\mathcal{M}_n}e^{-K^2l_m}\leq Ce^{-K^2(\ln n)^\gamma}.$$

Take $K=23.5/33.22>1/\sqrt{2}$ and $n\geq 25$ such that $2(40.3)^2K^4\epsilon_n^3\leq 0.5$, then $e_n^{(4)}(K)\leq 25$ and (4.32) holds for sufficiently large n. It holds in general provided that we enlarge C if necessary.

Finally, we apply Lemma 4.6.5 in the appendix to the functions $s_m-s_{m'}$, with $L=L_q$ and $\nu_n=\nu_A$, we have $v^2\leq v_{A,m,m'}^2/q$ and $\epsilon\leq e_{A,m,m'}/q$. For all m,m' in $\mathcal{M}_n$,

$$\|s_m-s_{m'}\|^2\leq 2(\|s_m-s\|^2+\|s_{m'}-s\|^2)\leq 4\frac{R_{A,m}\vee R_{A,m'}}{n},$$

thus, for all $\eta>0$, for all $x>0$,

$$\mathbb{P}\left(\delta^*(m,m')>4\eta\left(\frac{R_{A,m}\vee R_{A,m'}}{n}\right)+\frac{8v_{A,m,m'}^2x+4e_{A,m,m'}x^2/9}{\eta n}\right)\leq e^{-x}. \quad (4.43)$$

Let $K>0$, $l_{m,m'}=l_{n,\gamma}(R_{A,m},R_{A,m'})$, $x=K^2l_{m,m'}$ and $e_n^{(5)}(K)=\sqrt{2K^2+K^4\epsilon_n^2/9}$. From (4.40),

$$8v_{A,m,m'}^2x+4e_{A,m,m'}x^2/9\leq 4(e_n^{(5)}(K))\epsilon_n)^2R_{A,m}\vee R_{A,m'},$$

thus, for $\eta=e_n^{(5)}(K))\epsilon_n$,

$$4\eta\left(\frac{R_{A,m}\vee R_{A,m'}}{n}\right)+\frac{8v_{A,m,m'}^2x+4e_{A,m,m'}x^2/9}{\eta n}\leq 8e_n^{(5)}(K))\epsilon_n\frac{R_{A,m}\vee R_{A,m'}}{n}.$$

Hence, for all $K>1$, there exists a constant $C>0$ such that

$$\begin{aligned}&\mathbb{P}\left(\bigcup_{m,m'\in\mathcal{M}_n}\left\{\delta^*(m,m')>8e_n^{(5)}(K))\epsilon_n\frac{R_{A,m}\vee R_{A,m'}}{n}\right\}\right)\\&\leq\sum_{m,m'\in\mathcal{M}_n}\mathbb{P}\left(\delta^*(m,m')>8e_n^{(5)}(K))\epsilon_n\frac{R_{A,m}\vee R_{A,m'}}{n}\right)\\&\leq\sum_{m,m'\in\mathcal{M}_n}e^{-K^2l_{m,m'}}\leq Ce^{-K^2(\ln n)^\gamma}.\end{aligned}$$

Take $K=11.4/(8\sqrt{2})>1$ and n sufficiently large to have $8\sqrt{K^4\epsilon_n^2/9}\leq 0.6$, then $8e_n^{(5)}(K))\leq 12$ and (4.33) holds for sufficiently large n. It holds in general provided that we increase C if necessary.

Proof of the results in expectation

Let $K>0$, $z>0$, $l_m=l_{n,\gamma}(R_m,R_m)$, $x=K^2l_m(1+z)$ and

$$e_n^{(6)}(K,z)=(2\sqrt{19}+6)K\sqrt{x}+6K^2\epsilon_nx+4(19)^2K^4\epsilon_n^3x^2.$$

From (4.40),

$$\begin{aligned}&\frac{2D_{A,m}^{3/4}(e_{A,m}(19x)^2)^{1/4}+6\sqrt{D_{A,m}v_{A,m}^2x}+6v_{A,m}^2x+2e_{A,m}(19x)^2}{n}\\&\leq(e_n^{(6)}(K,1)+e_n^{(6)}(K,z))\epsilon_n\frac{R_{A,m}}{n}.\end{aligned}$$

Thus, from Proposition 4.6.3 in the Appendix, for all $z>0$ and all m in $\mathcal{M}_n$,

$$\mathbb{P}\left(p^*(m)-\frac{2D_{A,m}}{n}-e_n^{(6)}(K,1)\epsilon_n\frac{R_{A,m}}{n}>e_n^{(6)}(K,z))\epsilon_n\frac{R_{A,m}}{n}\right)\leq e^{-K^2l_m(1+z)}.$$

Let us now briefly explain how to deduce from this concentration inequalities the results in expectation.

[MI]: Integration of the concentration inequality

Let $\epsilon_m=\epsilon_nR_{A,m}/n$ and $f(m)=p^*(m)-2D_{A,m}/n-e_n^{(6)}(K,1)\epsilon_m$, we have

$$\mathbb{E}\left(\sup_{m\in\mathcal{M}_n}(f(m))_+\right)\ \leq\ \sum_{m\in\mathcal{M}_n}\mathbb{E}\left((f(m))_+\right)=\sum_{m\in\mathcal{M}_n}\int_0^\infty\mathbb{P}\left(f(m)>y\right)dy.$$

Since $z \mapsto g(z) = e_n^{(6)}(K,z))$ is clearly a C^1-diffeomorphism of $\mathbf{R}_+^*$, this last integral is equal to

$$\int_0^\infty \mathbb{P}\left(f(m) > \epsilon_m g(y)\right) \epsilon_m g'(y) dy$$

For all $K > 0$, there exists a constant $C > 0$ such that $g'(z) \leq C(z^{-1/2} + 1 + z)$. From Lemma 4.6.1, for all $K > 1$, $n \geq 2$, there exists a constant $C > 0$ such that

$$\mathbb{E}\left(\sup_{m \in \mathcal{M}_n} \left(p^*(m) - \frac{2D_{A,m}}{n} - e_n^{(6)}(K,1)\epsilon_n \frac{R_{A,m}}{n}\right)_+\right) \leq$$
$$C \sum_{m \in \mathcal{M}_n} \epsilon_n R_{A,m} e^{-K^2 l_m} \left(\int_0^\infty (z^{-1/2} + 1 + z) e^{-K^2 l_m z} dz\right) \leq C e^{-K^2 (\ln n)^\gamma}.$$

The last inequality comes from the fact that ϵ_n is bounded and $K^2 l_m \geq c > 0$ for all $n \geq 2$, $K > 1$. Take $K = 14.75/(2\sqrt{19} + 6) > 1$ and choose n sufficiently large such that $6K^2\epsilon_n + 4(19)^2 K^4 \epsilon_n^3 \leq 0.25$, then $e_n^{(6)}(K) \leq 15$ and (4.34) holds for all n sufficiently large. It holds for all n provided that we enlarge C if necessary.
We obtain (4.35) with the same arguments.
Let us now turn to the result on the resampling estimator of $p(m)$. Let $K > 0$, $z > 0$, $l_m = l_{n,\gamma}(R_m, R_m)$, $x = K^2 l_m (1 + z)$,

$$e_n^{(7)}(K,z) = \frac{n}{n-1}\left(16,62K\sqrt{x} + 6K^2\epsilon_n x + 4(19.1)^2 K^4 \epsilon_n^2 x^2\right),$$

From inequalities (4.40), we have

$$\frac{10.62 D_{A,m}^{3/4}(e_{A,m}x^2)^{1/4} + 6\sqrt{v_{A,m}^2 D_{A,m} x} + 6v_{A,m}^2 x + 2e_{A,m}(19.1x)^2}{n-1}$$
$$\leq (e_n^{(7)}(K,1) + e_n^{(7)}(K,z))\epsilon_n \frac{R_{A,m}}{n}$$

From inequalities (4.41) with $x = K^2 l_m(1+z)$ and for all $z > 0$, for all m in $\mathcal{M}_n$ and all $z > 0$

$$\mathbb{P}\left(p^*(m) - p_W^*(m) - e_n^{(7)}(K,1)\epsilon_n \frac{R_{A,m}}{n} > e_n^{(7)}(K,z)\epsilon_n \frac{R_{A,m}}{n}\right) \leq 2e^{-K^2 l_m (1+z)}.$$

We use again the method of integration **[MI]** to prove that, for all $K > 1$, there exists a constant $C > 0$ such that

$$\mathbb{E}\left(\sup_{m \in \mathcal{M}_n} \left(p^*(m) - p_W^*(m) - e_n^{(7)}(K,1)\epsilon_n \frac{R_{A,m}}{n}\right)_+\right) \leq C e^{-K^2 (\ln n)^\gamma}.$$

Take $K = 17/16.62 > 1$ and $n \geq 20$ such that $6K^2\epsilon_n + 4(19.1)^2 K^4 \epsilon_n^3 \leq 2$, then $e_n^{(7)}(K,1) \leq 20$ and (4.36) holds for sufficiently large n. It holds in general provided that we enlarge C if necessary.
We obtain (4.37) with the same arguments.
Let $K > 0$, $l_{m,m'} = l_{n,\gamma}(R_m, R_{m'})$, $z > 0$, $x = K^2 l_{m,m'}(1+z)$,

$$e_n^{(8)}(K,z) = \sqrt{2K^2 z + 2K^4 \epsilon_n^2 x^2/9},$$

$e_n^{(8)}(K) = e_n^{(8)}(K,1)$, $g_K(z) = (e_n^{(8)}(K,z))^2/e_n^{(8)}(K)$ and $\eta = e_n^{(8)}(K,1)\epsilon_n$.

$$\begin{aligned}&4\eta\frac{R_{A,m}\vee R_{A,m'}}{n} + \frac{8v_{A,m,m'}^2 x + 4e_{A,m,m'}x^2/9}{\eta n}\\ &\leq 4\left(2e_n^{(8)}(K) + g_K(z)\right)\epsilon_n\frac{R_{A,m}\vee R_{A,m'}}{n}.\end{aligned}$$

Thus from (4.43), for all $z > 0$, for all m, m' in $\mathcal{M}_n$ and all $K > 0$,

$$\mathbb{P}\left(\delta(m,m') - 8e_n^{(8)}(K)\epsilon_n\frac{R_{A,m}\vee R_{A,m'}}{n} > \epsilon_n\frac{R_{A,m}\vee R_{A,m'}}{n}g_K(z)\right) \leq e^{-K^2 l_{m,m'}(1+z)}.$$

Thus

$$\begin{aligned}&\mathbb{E}\left(\sup_{(m,m')\in\mathcal{M}_n^2}\left(\delta^*(m,m') - 8e_n^{(8)}(K)\epsilon_n\frac{R_{A,m}\vee R_{A,m'}}{n}\right)_+\right)\\ &\leq \sum_{(m,m')\in\mathcal{M}_n^2}\mathbb{E}\left(\left(\delta^*(m,m') - 8e_n^{(8)}(K)\epsilon_n\frac{R_{A,m}\vee R_{A,m'}}{n}\right)_+\right)\\ &= \sum_{(m,m')\in\mathcal{M}_n^2}\int_0^\infty \mathbb{P}\left(\delta^*(m,m') - 8e_n^{(8)}(K)\epsilon_n\frac{R_{A,m}\vee R_{A,m'}}{n} > x\right)dx\end{aligned}$$

Let $x = \epsilon_n\frac{R_{A,m}\vee R_{A,m'}}{n}g_K(z)$. For all $K > 0$, for all $n \geq 2$, there exists a constant $C > 0$ such that $g_K(z)' \leq C(1+z)$. Thus, from Lemma 4.6.1, for all $K > \sqrt{2}$, there exists a constant $C > 0$ such that

$$\begin{aligned}&\mathbb{E}\left(\sup_{(m,m')\in\mathcal{M}_n^2}\left(\delta^*(m,m') - 8e_n^{(8)}(K)\epsilon_n\frac{R_{A,m}\vee R_{A,m'}}{n}\right)_+\right)\\ &\leq C\sum_{(m,m')\in\mathcal{M}_n^2}\epsilon_n\frac{R_{A,m}\vee R_{A,m'}}{n}e^{-K^2 l_{m,m'}}\int_0^\infty(1+z)e^{-K^2 l_{m,m'}z}dz \leq Ce^{-K^2(\ln n)^\gamma}.\end{aligned}$$

Take $K = 17/(8\sqrt{2}) > \sqrt{2}$ and n sufficiently large to have $8\sqrt{2K^4\epsilon_n^2/9} \leq 3$, then $8e_n^{(8)}(K,1)) \leq 20$ and (4.38) holds for sufficiently large n. It holds in general provided that we increase C if necessary. We can now turn to the proofs of the main results of this part.

4.5.1 Proof of Theorem 4.3.1

Let us first assume that $X_1, ..., X_n$ are β-mixing. Let us define the events

$$\Omega_p = \bigcap_{m\in\mathcal{M}_n}\left\{-25\epsilon(R_m)\frac{R_m}{n} \leq p^*(m) - \frac{2D_{A,m}}{n} \leq 15\epsilon(R_m)\frac{R_m}{n}\right\}, \qquad (4.44)$$

$$\tilde{\Omega}_p = \bigcap_{m\in\mathcal{M}_n}\left\{-25\epsilon(R_m)\frac{R_m}{n} \leq p^*(m) - p_W^*(m) \leq 15\epsilon(R_m)\frac{R_m}{n}\right\}$$

$$\Omega_d = \bigcap_{(m,m')\in\mathcal{M}_n}\left\{\delta(m,m') \leq 12\epsilon_n\left(\frac{R_{A,m}\vee R_{A,m'}}{n}\right)\right\}, \qquad (4.45)$$

$$\begin{aligned}\Omega_C &= \left(\bigcap_{m\in\mathcal{M}_n}\{p(m)=p^*(m)\}\right)\cap\left(\bigcap_{m\in\mathcal{M}_n}\{p_W(m)=p^*_W(m)\}\right)\\ &\cap\left(\bigcap_{(m,m')\in\mathcal{M}_n}\{\delta(m,m')=\delta^*(m,m')\}\right). \end{aligned} \tag{4.46}$$

From Lemmas 4.5.1 and 4.5.3, there exists a constant $C>0$ such that

$$\mathbb{P}(\Omega_p^c)\le Ce^{-\frac{1}{2}(\ln n)^\gamma},\ \mathbb{P}(\tilde{\Omega}_p^c)\le Ce^{-\frac{1}{2}(\ln n)^\gamma},\ \mathbb{P}(\Omega_d^c)\le Ce^{-(\ln n)^\gamma},\ \mathbb{P}(\Omega_C^c)\le p\beta_q.$$

Let $\Omega=\Omega_p\cap\tilde{\Omega}_p\cap\Omega_d\cap\Omega_C$. Recall that $\mathrm{pen}(m)=2p_W(m)$. On Ω, from inequality (4.6), for all m in $\mathcal{M}_n$,

$$\begin{aligned}\|s-\tilde{s}\|^2 &\le \|s-\hat{s}_{A,m}\|^2+2\left(p_W(m)-p(m)\right)-2\left(p_W(\hat{m})-p(\hat{m})\right)+\delta(m,\hat{m})\\ &= \|s-\hat{s}_{A,m}\|^2+2\left(p^*_W(m)-p^*(m)\right)-2\left(p^*_W(\hat{m})-p^*(\hat{m})\right)+\delta^*(m,\hat{m})\\ &\le \|s-\hat{s}_{A,m}\|^2+62\epsilon_n\frac{R_m}{n}+42\epsilon_n\frac{R_{\hat{m}}}{n}.\end{aligned}$$

On Ω,

$$\frac{R_m}{n} = \|s-\hat{s}_{A,m}\|^2+\frac{2D_{A,m}}{n}-p^*(m)\le\|s-\hat{s}_{A,m}\|^2+25\epsilon_n\frac{R_m}{n}.$$

If $25\epsilon_n<1$, on Ω,

$$\|s-\tilde{s}\|^2 \le \frac{1+37\epsilon_n}{1-25\epsilon_n}\|s-\hat{s}_{A,m}\|^2+\frac{42\epsilon_n}{1-25\epsilon_n}\|s-\tilde{s}\|^2.$$

Hence, if $67\epsilon_n<1$, on Ω,

$$\mathbb{P}\left(\|s-\tilde{s}\|^2>\frac{1+37\epsilon_n}{1-67\epsilon_n}\inf_{m\in\mathcal{M}_n}\|s-\hat{s}_{A,m}\|^2\right)\le Ce^{-\frac{1}{2}(\ln n)^\gamma}+p\beta_q.$$

Take n sufficiently large to have $67\epsilon_n<1$ and $104/(1-67\epsilon_n)\le110$. Then,

$$\frac{1+37\epsilon_n}{1-67\epsilon_n}=1+\frac{104}{1-67\epsilon_n}\epsilon_n\le1+110\epsilon_n$$

and (4.11) holds for sufficiently large n. It holds in general provided that we increase the constant C if necessary.
From inequality (4.6), for all m in $\mathcal{M}_n$,

$$\begin{aligned}\|s-\tilde{s}\|^2 &\le \|s-\hat{s}_{A,m}\|^2+2\left(p_W(m)-p(m)\right)-2\left(p_W(\hat{m})-p(\hat{m})\right)+\delta(m,\hat{m})\\ &= \|s-\hat{s}_{A,m}\|^2+2\left(p^*_W(m)-p^*(m)-35\epsilon_n\frac{R_{A,m}}{n}\right)+90\epsilon_n\frac{R_{A,m}}{n}\\ &\quad+2\left(p^*(\hat{m})-p^*_W(\hat{m})-20\epsilon_n\frac{R_{A,\hat{m}}}{n}\right)+60\epsilon_n\frac{R_{A,\hat{m}}}{n}\\ &\quad+\delta^*(m,\hat{m})-20\epsilon_n\frac{R_{A,m}\vee R_{A,\hat{m}}}{n}+2(p_W(m)-p^*_W(m))\\ &\quad+2(p^*(m)-p(m)+p^*_W(\hat{m})-p_W(\hat{m})+p(\hat{m})-p^*(\hat{m}))\\ &\quad+\delta(m,\hat{m})-\delta^*(m,\hat{m}).\end{aligned}$$

For all m in $\mathcal{M}_n$,

$$\begin{aligned}\frac{R_{A,m}}{n} &= \frac{\|s-\hat{s}_{A,m}\|^2}{1-35\epsilon_n}+\frac{(1-35\epsilon_n)R_{A,m}/n-\|s-\hat{s}_{A,m}\|^2}{1-35\epsilon_n}\\ &= \frac{\|s-\hat{s}_{A,m}\|^2}{1-35\epsilon_n}+\frac{2D_{A,m}/n-35\epsilon_n R_{A,m}/n-\|s_m-\hat{s}_{A,m}\|^2}{1-35\epsilon_n}. \qquad (4.47)\end{aligned}$$

In the control of $\|s-\tilde{s}\|^2$, we replace $R_{A,m}/n$ and $R_{A,\hat{m}}/n$ by the expressions obtained in (4.47) in the terms $90\epsilon_n R_{A,m}/n$ and $60\epsilon_n R_{A,\hat{m}}/n$. Assume that $35\epsilon_n<1$,

$$\begin{aligned}&\frac{1-95\epsilon_n}{1-35\epsilon_n}\|s-\tilde{s}\|^2 \le \frac{1+55\epsilon_n}{1-35\epsilon_n}\inf_{m\in\mathcal{M}_n}\|s-\hat{s}_{A,m}\|^2\\ &+\frac{150\epsilon_n}{1-35\epsilon_n}\sup_{m\in\mathcal{M}_n}\left(\frac{2D_{A,m}}{n}-p^*(m)-35\epsilon_n\frac{R_m}{n}\right)+\\ &\frac{4+10\epsilon_n}{1-35\epsilon_n}\sup_{m\in\mathcal{M}_n}|p^*(m)-p(m))|+\sup_{m,m'\in\mathcal{M}_n}\left(\delta^*(m,m')-20\epsilon_n\left(\frac{R_{A,m}\vee R_{A,m'}}{n}\right)\right)\\ &+2\sup_{m\in\mathcal{M}_n}\left(p^*_W(m)-p^*(m)-35\epsilon(R_m)\frac{R_m}{n}\right)+4\sup_{m\in\mathcal{M}_n}|p_W(m)-p^*_W(m)|\\ &+2\sup_{m\in\mathcal{M}_n}\left(p^*(m)-p^*_W(m)-15\epsilon(R_m)\frac{R_m}{n}\right)+\sup_{m,m'\in(\mathcal{M}_n)^2}\delta(m,m')-\delta^*(m,m').\end{aligned}$$

We take the expectation in this last inequality and we use inequalities (4.26), (4.27), (4.28), (4.35), (4.36), (4.37) and (4.38) to obtain that, when $95\epsilon_n<1$, there exists a constant $C>0$ such that

$$\|s-\tilde{s}\|^2\le\frac{1+55\epsilon_n}{1-95\epsilon_n}\inf_{m\in\mathcal{M}_n}\|s-\hat{s}_{A,m}\|^2+C\left(\tau_q MC_n+e^{-\frac{1}{2}(\ln n)^\gamma}\right)$$

Take n sufficiently large to have $95\epsilon_n<1$ and $150/(1-95\epsilon_n)\le 160$. Then,

$$\frac{1+55\epsilon_n}{1-95\epsilon_n}=1+\frac{150}{1-95\epsilon_n}\epsilon_n\le 1+160\epsilon_n$$

and (4.12) holds for sufficiently large n. It holds in general provided that we increase the constant C if necessary.

4.5.2 Proof of Theorem 4.3.2

Let us first assume that $X_1,...,X_n$ are β-mixing and let $A^*_0,...,A^*_{p-1}$ be the random variables built with Viennet's Lemma. Let $\Omega_T=\Omega_p\cap\Omega_d\cap\Omega_C$ where Ω_p, Ω_d and Ω_C are defined respectively in (4.44), (4.45) and (4.46). Recall that there exists a constant $C>0$ such that

$$\mathbb{P}(\Omega_p^c)\le Ce^{-\frac{1}{2}(\ln n)^\gamma},\ \mathbb{P}(\Omega_d^c)\le Ce^{-(\ln n)^\gamma},\ \mathbb{P}(\Omega_C^c)\le p\beta_q.$$

If $c_n\le 0$, there is nothing to prove, hence, we can assume that $c_n>0$ and thus that $75\epsilon^*_n<\delta<1$.

$\hat{m}$ minimizes by definition the following criterion

$$\begin{aligned}
\mathrm{Crit}(m) &= \|\hat{s}_{A,m}\|^2 - 2P_A(\hat{s}_{A,m}) + \mathrm{pen}(m) + \|s\|^2 + 2\nu_A(s_{m_o}) \\
&= \|\hat{s}_{A,m}\|^2 - 2P(\hat{s}_{A,m}) + \|s\|^2 - 2\nu_A(\hat{s}_{A,m}) + 2\nu_A(s_{m_o}) + \mathrm{pen}(m) \\
&= \|\hat{s}_{A,m} - s\|^2 - 2\nu_A(\hat{s}_{A,m} - s_m) + 2\nu_A(s_{m_o} - s_m) + \mathrm{pen}(m) \\
&= \|\hat{s}_{A,m} - s\|^2 - 2\|\hat{s}_{A,m} - s_m\|^2 + \delta(m_o, m) + \mathrm{pen}(m) \\
&= \|s - s_m\|^2 - p(m) + \delta(m_o, m) + \mathrm{pen}(m)
\end{aligned}$$

since $p(m) = \|s_m - \hat{s}_{A,m}\|^2 = \nu_A(\hat{s}_{A,m} - s_m)$. Thus, on Ω_T, $\hat{m}$ minimizes the following criterion

$$\mathrm{Crit}(m) = \|s - s_m\|^2 - p^*(m) + \delta^*(m, m_o) + \mathrm{pen}(m)$$

For all m in $\mathcal{M}_n$, we have $0 \leq \mathrm{pen}(m) < (2 - \delta)D_{A,m}/n$ and $R_{m_o} \leq R_m$. Thus, for all m in $\mathcal{M}_n$, on Ω_T

$$\begin{aligned}
\mathrm{Crit}(m) &\geq \|s - s_m\|^2 - \frac{2D_{A,m}}{n} + \left(\frac{2D_{A,m}}{n} - p^*(m)\right) + \delta^*(m, m_o) \\
&\geq (1 - 27\epsilon_n)\|s - s_m\|^2 - (1 + 27\epsilon_n)\frac{2D_{A,m}}{n} \geq -(1 + 27\epsilon_n)\frac{2D_{A,m}}{n} \\
\mathrm{Crit}(m) &\leq \|s - s_m\|^2 - (\delta - 74\epsilon_n)\frac{D_{A,m}}{n}.
\end{aligned}$$

If $D_{A,m} > c_n D_{A,m*}$, then

$$\begin{aligned}
\mathrm{Crit}(m) &\geq -(1 + 27\epsilon_n)\frac{2D_{A,m}}{n} > -(1 + 27\epsilon_n)\, c_n \frac{2D_{A,m^*}}{n} \\
&\geq -(\delta - 74\epsilon_n - h_n^*)\frac{D_{A,m^*}}{n} \geq \mathrm{Crit}(m^*).
\end{aligned}$$

This proves that $D_{A,m} \geq c_n D_{A,m*}$.
It follows that, on Ω_T,

$$\begin{aligned}
\|s - \tilde{s}\|^2 &= \frac{R_{A,\hat{m}}}{n} + \left(p(\hat{m}) - \frac{2D_{A,\hat{m}}}{n}\right) \geq (1 - 25\epsilon_n)\frac{R_{A,\hat{m}}}{n} \\
&\geq (1 - 25\epsilon_n)\frac{2D_{A,\hat{m}}}{n} \geq (1 - 25\epsilon_n)c_n\frac{2D_{A,m^*}}{n}.
\end{aligned}$$

Moreover, on Ω_T,

$$\inf_{m \in \mathcal{M}_n} \|s - \hat{s}_{A,m}\|^2 \leq \inf_{m \in \mathcal{M}_n} \frac{R_{A,m}}{n}(1 + 15\epsilon_n) \leq \frac{R_{A,m_o}}{n}(1 + 15\epsilon_n).$$

Thus

$$\|s - \tilde{s}\|^2 \geq (1 - 25\epsilon_n)c_n\frac{2D_{A,m^*}}{n} \geq 2c_n\left(\frac{1 - 25\epsilon_n}{1 + 15\epsilon_n}\right)\frac{D_{A,m^*}}{R_{m_o}} \inf_{m \in \mathcal{M}_n} \|s - \hat{s}_{A,m}\|^2.$$

Since $\epsilon_n < 1/75$, we have $2(1 - 25\epsilon_n)(1 + 15\epsilon_n) \leq 2(1 - 1/3)(1 + 1/5) \leq 1$. This conclude the proof of (4.13).

In order to prove inequality (4.14) observe that, for all m in $\mathcal{M}_n$, since $\mathrm{pen}(m) \geq 0$, and $\|s-s_m\|^2 - 35\epsilon_n R_{A,m}/n \geq -35\epsilon_n D_{A,m}/n$

$$\begin{aligned}
\mathrm{Crit}(m) &\geq \|s-s_m\|^2 + \left(-p^*(m) + 15\epsilon_n \frac{R_{A,m}}{n}\right) + (p^*(m) - p(m)) - 35\epsilon_n \frac{R_{A,m}}{n} \\
&\quad + \left(\delta^*(m,m_o) + 20\epsilon_n \frac{R_{A,m}}{n}\right) + (\delta(m,m_o) - \delta^*(m,m_o)) \\
&= -(1+35\epsilon_n)\frac{2D_{A,m}}{n} + \left(\frac{2D_{A,m}}{n} - p^*(m) + 15\epsilon_n \frac{R_{A,m}}{n}\right) + (p^*(m) - p(m)) \\
&\quad + \left(\delta^*(m,m_o) + 20\epsilon_n \frac{R_{A,m}}{n}\right) + (\delta(m,m_o) - \delta^*(m,m_o)).
\end{aligned}$$

Therefore,

$$\begin{aligned}
-\frac{2D_{A,\hat{m}}}{n}(1+35\epsilon_n) &\leq \mathrm{Crit}(\hat{m}) + \sup_{m\in\mathcal{M}_n}\left(p^*(m) - \frac{2D_{A,m}}{n} - 15\epsilon_n \frac{R_{A,m}}{n}\right) \\
&\quad + \sup_{(m,m')\in\mathcal{M}_n^2}\left(\delta^*(m,m') - 20\epsilon_n \frac{R_{A,m}\vee R_{A,m'}}{n}\right) \\
&\quad + \sup_{m\in\mathcal{M}_n}(p(m) - p^*(m)) + \sup_{m,m'\in\mathcal{M}_n}(\delta(m,m') - \delta^*(m,m'))
\end{aligned} \tag{4.48}$$

Since, for all m in $\mathcal{M}_n$, $\mathrm{pen}(m) \leq (2-\delta)D_{A,m}/n$,

$$\mathrm{Crit}(m) \leq \|s-s_m\|^2 + \left(\frac{2D_{A,m}}{n} - p^*(m)\right) - \delta\frac{D_{A,m}}{n} + (p^*(m) - p(m)) + \delta(m,m_o). \tag{4.49}$$

Since $\mathrm{Crit}(\hat{m}) \leq \mathrm{Crit}(m^*)$, from (4.49) and (4.26),

$$\begin{aligned}
\mathbb{E}\left(\mathrm{Crit}(\hat{m})\right) &\leq \mathbb{E}\left(\mathrm{Crit}(m^*)\right) \leq \|s-s_{m^*}\|^2 - \delta\frac{D_{A,m^*}}{n} + 4\tau_q MC_n \\
&\leq -(\delta - h_n^*)\frac{D_{A,m^*}}{n} + 4\tau_q MC_n \leq -c_n'(1+35\epsilon_n)\frac{2D_{A,m^*}}{n} + 4\tau_q MC_n.
\end{aligned}$$

Take the expectation in (4.48) and use inequalities (4.26), (4.28), (4.34) and (4.38) to obtain (4.14).

We deduce from (4.14) that there exists a constant $C>0$ such that

$$\begin{aligned}
\mathbb{E}\left(\|s-\tilde{s}\|^2\right) &\geq \frac{2}{n}\mathbb{E}\left(D_{A,\hat{m}}\right) \geq 2c_n'\frac{D_{A,m^*}}{n} - C(e^{-\frac{1}{2}(\ln n)^\gamma} + \tau_q MC_n) \\
&\geq 2\frac{c_n'}{h_n^o}\frac{R_{m_o}}{n} - C(e^{-\frac{1}{2}(\ln n)^\gamma} + \tau_q MC_n).
\end{aligned}$$

The proof of (4.15) is conclude since

$$\mathbb{E}\left(\inf_{m\in\mathcal{M}_n}\|s-\hat{s}_{A,m}\|^2\right) \leq \inf_{m\in\mathcal{M}_n}\mathbb{E}\left(\|s-\hat{s}_{A,m}\|^2\right) = \frac{R_{m_o}}{n} + C\tau_q MC_n.$$

4.5.3 Proof of Theorem 4.3.3

If $c_n = \infty$, there is nothing to prove. Thus we can assume that $c_n < \infty$ and thus that $1 + \underline{\delta} - 27\epsilon_n > 0$. Let us first assume that $X_1, ..., X_n$ are β-mixing and let $A_0^*, ...A_{p-1}^*$ be the random variables given by Viennet's Lemma. Recall that $\hat{m}$ minimizes over $\mathcal{M}_n$ the following criterion.

$$\mathrm{Crit}(m) \ = \ \|s - s_m\|^2 - p(m) + \delta(m, m_o) + \mathrm{pen}(m).$$

We keep the notations Ω_p, Ω_d and Ω_C defined by (4.44), (4.45), (4.46). We introduce the event

$$\Omega_{\mathrm{pen}} = \bigcap_{m \in \mathcal{M}_n} \left\{ \frac{4D_{A,m}}{n} + \underline{\delta}\frac{R_{A,m}}{n} \leq \mathrm{pen}(m) \leq \frac{4D_{A,m}}{n} + \bar{\delta}\frac{R_{A,m}}{n} \right\}$$

and let $\Omega = \Omega_p \cap \Omega_d \cap \Omega_C \cap \Omega_{\mathrm{pen}}$. Since $R_{m_o} \leq R_m$, on Ω,

$$\begin{aligned}
\mathrm{Crit}(m) \ &\geq \ (1 + \underline{\delta} - 12\epsilon_n)\frac{R_{A,m}}{n} + \left(\frac{2D_{A,m}}{n} - p^*(m)\right) \\
&\geq \ (1 + \underline{\delta} - 27\epsilon_n)\frac{R_{A,m}}{n} \geq (1 + \underline{\delta} - 27\epsilon_n)\frac{2D_{A,m}}{n}. \\
\mathrm{Crit}(m) \ &\leq \ (1 + \bar{\delta} + 37\epsilon_n)\frac{R_{A,m}}{n}.
\end{aligned}$$

If $D_{A,m} > c_n R_{A,m_o}$,

$$\begin{aligned}
\mathrm{Crit}(m) \ &\geq \ (1 + \underline{\delta} - 27\epsilon_n)\frac{2D_{A,m}}{n} \geq 2(1 + \underline{\delta} - 27\epsilon_n)c_n\frac{R_{A,m_o}}{n} \\
&\geq \ (1 + \bar{\delta} + 37\epsilon_n)\frac{R_{A,m_o}}{n} \geq \mathrm{Crit}(m_o)
\end{aligned}$$

This implies that $D_{\hat{m}} \leq c_n R_{A,m_o}$. Moreover, we have, from (4.6), for all m in $\mathcal{M}_n$

$$\begin{aligned}
\|s - \tilde{s}\|^2 \ &\leq \ \|s - \hat{s}_{A,m}\|^2 + (\mathrm{pen}(m) - 2p^*(m)) + (2p^*(\hat{m}) - \mathrm{pen}(\hat{m})) + \delta^*(m, \hat{m}) \\
&\leq \ \|s - \hat{s}_{A,m}\|^2 + 2\left(\frac{2D_{A,m}}{n} - p^*(m)\right) + (\bar{\delta} + 12\epsilon_n)\frac{R_{A,m}}{n} \\
&\quad +2\left(p^*(\hat{m}) - \frac{2D_{A,\hat{m}}}{n}\right) + (-\underline{\delta} + 12\epsilon_n)\frac{R_{A,\hat{m}}}{n} \\
&\leq \ \|s - \hat{s}_{A,m}\|^2 + (37\epsilon_n + \bar{\delta})\frac{R_{A,m}}{n} + (27\epsilon_n - \underline{\delta})\frac{R_{A,\hat{m}}}{n}.
\end{aligned}$$

For all m in $\mathcal{M}_n$, we have, on Ω,

$$\|s - \hat{s}_{A,m}\|^2 = \frac{R_{A,m}}{n} + \left(p^*(m) - \frac{2D_{A,m}}{n}\right) \geq (1 - 25\epsilon_n)\frac{R_{A,m}}{n}.$$

Assume that $25\epsilon_n < 1$, then, for all $m \in \mathcal{M}_n$,

$$\|s - \tilde{s}\|^2 \leq \|s - \hat{s}_{A,m}\|^2 \left(1 + \frac{37\epsilon_n + \bar{\delta}}{1 - 25\epsilon_n}\right) + \frac{27\epsilon_n - \underline{\delta}}{1 - 25\epsilon_n}\|s - \tilde{s}\|^2.$$

This proves (4.16) for sufficiently large n. (4.16) holds in general provided that we increase the constant C if necessary.
Let us now assume that $X_1,...,X_n$ τ-mixing and let $A_0^*,...,A_{p-1}^*$, be the random variables given by the τ-mixing Lemma. Recall that

$$\mathbb{E}\left(\sup_{m\in\mathcal{M}_n}\left(\frac{4D_{A,m}}{n}+\underline{\delta}\frac{R_{A,m}}{n}-\text{pen}(m)\right)_+\right)\leq e_n,$$

$$\mathbb{E}\left(\sup_{m\in\mathcal{M}_n}\left(\text{pen}(m)-\frac{4D_{A,m}}{n}-\bar{\delta}\frac{R_{A,m}}{n}\right)_+\right)\leq e_n.$$

For all m in $\mathcal{M}_n$, we have,

$$\begin{aligned}\frac{R_{A,m}}{n} &= \text{Crit}(m)+\left(p^*(m)-\frac{2D_{A,m}}{n}-15\epsilon_n\frac{R_{A,m}}{n}\right)+\left(\frac{4D_{A,m}}{n}-\text{pen}(m)+\underline{\delta}\frac{R_{A,m}}{n}\right)\\ &\quad-\left(\delta^*(m,m_o)+20\epsilon_n\frac{R_{A,m}}{n}\right)+(p(m)-p^*(m))\\ &\quad+(35\epsilon_n-\underline{\delta})\frac{R_{A,m}}{n}+(\delta^*(m,m_o)-\delta(m,m_o)).\end{aligned}$$

Therefore

$$\begin{aligned}(1+\underline{\delta}-35\epsilon_n)\frac{R_{A,\hat{m}}}{n} &\leq \text{Crit}(m_o)+\sup_{m\in\mathcal{M}_n}\left(p^*(m)-\frac{2D_{A,m}}{n}-15\epsilon_n\frac{R_{A,m}}{n}\right)\\ &\quad+\sup_{m\in\mathcal{M}_n}\left(\frac{4D_{A,m}}{n}-\text{pen}(m)+\underline{\delta}\frac{R_{A,m}}{n}\right)\\ &\quad+\sup_{(m,m')\in\mathcal{M}_n^2}\left(\delta(m,m')-20\epsilon_n\frac{R_{A,m}\vee R_{A,m'}}{n}\right)\\ &\quad+\sup_{(m,m')\in\mathcal{M}_n^2}(\delta^*(m,m')-\delta(m,m'))\\ &\quad+\sup_{m\in\mathcal{M}_n}|p(m)-p^*(m)|. \end{aligned}\tag{4.50}$$

On the other hand, for all m in $\mathcal{M}_n$, $\text{Crit}(m)=\|s-s_m\|^2-p(m)+\delta(m_o,m)+\text{pen}(m)$, thus

$$\begin{aligned}\text{Crit}(m_o) &\leq (1+\bar{\delta})R_{A,m_o}+\left(\frac{2D_{A,m_o}}{n}-p^*(m_o)\right)+(p^*(m_o)-p(m_o))\\ &\quad+\text{pen}(m_o)-\frac{4D_{A,m_o}}{n}-\bar{\delta}\frac{R_{A,m_o}}{n}.\end{aligned}$$

Since $\mathbb{E}\left(\text{pen}(m_o)-4D_{A,m_o}/n-\bar{\delta}R_{A,m_o}/n\right)\leq e_n$ and $2D_{A,m_o}/n=\mathbb{E}(p^*(m_o))$, from inequality (4.26), there exists a constant $C>0$ such that

$$\mathbb{E}\left(\text{Crit}(m_o)\right)\leq(1+\bar{\delta})\frac{R_{A,m_o}}{n}+C\tau_qMC_n+e_n.$$

For all m in $\mathcal{M}_n$, $2D_{A,m}\leq R_{A,m}$. Take the expectation in (4.50), from inequalities (4.26), (4.28), (4.34) and (4.38), there exists an absolut constant $C>0$ such that

$$\mathbb{E}\left(D_{\hat{m}}\right)\leq c_n\left(R_{m_o}+Cn\left[\tau_qMC_n+e^{-\frac{1}{2}(\ln n)^\gamma}+e_n\right]\right).$$

This proves inequality (4.17).

From (4.6), for all m in $\mathcal{M}_n$, we have

$$\begin{aligned}
\|s-\tilde{s}\|^2 \;\leq\;& \|s-\hat{s}_{A,m}\|^2+2\left(\frac{2D_{A,m}}{n}-p^*(m)-35\epsilon_n\frac{R_{A,m}}{n}\right)\\
&+\left(\delta^*(m,\hat{m})-20\epsilon_n\frac{R_{A,m}\vee R_{A,\hat{m}}}{n}\right)\\
&+2\left(-\frac{2D_{A,\hat{m}}}{n}+p^*(\hat{m})-15\epsilon(R_{\hat{m}})\frac{R_{\hat{m}}}{n})\right)\\
&+\left(-\mathrm{pen}(\hat{m})+2\frac{2D_{\hat{m}}}{n}+\underline{\delta}\frac{R_{A,\hat{m}}}{n}\right)+\left(\mathrm{pen}(m)-2\frac{2D_{A,m}}{n}-\bar{\delta}\frac{R_{A,m}}{n}\right)\\
&+(90\epsilon_n+\bar{\delta})\frac{R_{A,m}}{n}+(50\epsilon_n-\underline{\delta})\frac{R_{A,\hat{m}}}{n}\\
&+4\sup_{m\in\mathcal{M}_n}|p(m)-p^*(m)|+\sup_{(m,m')\in\mathcal{M}_n^2}(\delta(m,m')-\delta^*(m,m')). \qquad (4.51)
\end{aligned}$$

Assume that $35\epsilon_n<1$, for all m in $\mathcal{M}_n$, we have

$$\begin{aligned}
\frac{R_{A,m}}{n} \;=\;& \frac{(1-35\epsilon_n)R_{A,m}}{n}-\frac{\|s-\hat{s}_{A,m}\|^2}{1-35\epsilon_n}+\frac{\|s-\hat{s}_{A,m}\|^2}{1-35\epsilon_n}\\
\leq\;& \frac{1}{1-35\epsilon_n}\left(\|s-\hat{s}_{A,m}\|^2+\frac{2D_{A,m}}{n}-p(m)-35\epsilon_n\frac{R_{A,m}}{n}\right)\\
\leq\;& \frac{1}{1-35\epsilon_n}\left(\|s-\hat{s}_{A,m}\|^2+\frac{2D_{A,m}}{n}-p^*(m)-35\epsilon_n\frac{R_{A,m}}{n}+p(m)-p^*(m)\right)
\end{aligned}$$

We use this expression in the terms $(90\epsilon_n+\bar{\delta})R_{A,m}/n$ and $(50\epsilon_n-\underline{\delta})R_{A,\hat{m}}/n$ of inequality (4.51). We deduce that, for all m in $\mathcal{M}_n$,

$$\begin{aligned}
&\frac{1+\underline{\delta}-85\epsilon_n}{1-35\epsilon_n}\|s-\tilde{s}\|^2\leq\frac{1+\bar{\delta}+55\epsilon_n}{1-35\epsilon_n}\inf_{m\in\mathcal{M}_n}\|s-\hat{s}_{A,m}\|^2\\
&+\frac{2+70\epsilon_n+\bar{\delta}-\underline{\delta}}{1-35\epsilon_n}\sup_{m\in\mathcal{M}_n}\left(\frac{2D_{A,m}}{n}-p^*(m)-35\epsilon_n\frac{R_{A,m}}{n}\right)\\
&+\sup_{m\in\mathcal{M}_n}\left(\mathrm{pen}(m)-\frac{4D_{A,m}}{n}-\bar{\delta}\frac{R_{A,m}}{n}\right)+\sup_{m\in\mathcal{M}_n}\left(\frac{4D_{A,m}}{n}+\underline{\delta})\frac{R_{A,\hat{m}}}{n}-\mathrm{pen}(m)\right)\\
&+2\sup_{m\in\mathcal{M}_n}\left(p^*(m)-\frac{2D_{A,m}}{n}-15\epsilon_n\frac{R_{A,m}}{n})\right)\\
&+\sup_{(m,m')\in\mathcal{M}_n^2}\left(\delta^*(m,m')-20\epsilon_n\frac{R_{A,m}\vee R_{A,m'}}{n}\right)\\
&+\frac{4+\bar{\delta}-\underline{\delta}}{1-35\epsilon_n}\sup_{m\in\mathcal{M}_n}|p(m)-p^*(m)|+\sup_{(m,m')\in\mathcal{M}_n^2}(\delta(m,m')-\delta^*(m,m')).
\end{aligned}$$

We take the expectation in this last inequality and we deduce that, for sufficiently large n, (4.18) comes from (4.26), (4.28), (4.34), (4.35) and (4.38). It holds in general provided that we enlarge the constant C if necessary.

4.5.4 Proof of Lemma 4.4.5

Let us first recall the covariance inequality proved by Dedecker & Prieur [25] for τ-mixing sequences.

Lemma 4.5.4 *Let X, Y be two identically distributed real valued random variables, with common density s in $L^2(\mu)$. There exists a constant c_τ and a random variable $b(\sigma(X), Y)$ such that $\mathbb{E}(b(\sigma(X), Y)) = c_\tau \left(\tau(\sigma(X), Y)\right)^{1/3}$ such that, for all Lipschitz functions f and all h in BV*

$$|Cov(f(X), h(Y))| \leq \|h\|_{BV} \mathbb{E}\left(|f(X)|b(\sigma(X), Y)\right) \leq c_\tau \|h\|_{BV} \|f\|_\infty \left(\tau(\sigma(X), Y)\right)^{1/3}. \tag{4.52}$$

It comes from this Lemma and inequalities (4.20, 4.21, 4.22) that

$$\begin{aligned}
D_{A,m} &= \frac{1}{q} \sum_{(j,k)\in m} \operatorname{Var}\left(\sum_{i=1}^{q} \psi_{j,k}(X_i)\right) \leq 2 \sum_{(j,k)\in m} \sum_{l=1}^{q} (q+1-l)|\operatorname{Cov}(\psi_{j,k}(X_1), \psi_{j,k}(X_l))| \\
&\leq \frac{2}{q} \sum_{j=0}^{J_m} \sum_{k\in\mathbb{Z}} \sum_{l=1}^{q} \|\psi_{j,k}\|_{BV} \mathbb{E}\left(|\psi_{j,k}(X_1)|b(\sigma(X_1), X_l)\right) \\
&\leq 2c_\tau K_{BV} \sum_{j=0}^{J_m} 2^{j/2} \left\|\sum_{k\in\mathbb{Z}} |\psi_{j,k}|\right\|_\infty \sum_{l=1}^{q} \tau_{l-1}^{1/3} \\
&\leq 4\left(c_\tau A K_\infty K_{BV} \sum_{l=0}^{\infty} \tau_l^{1/3}\right) 2^{J_m}.
\end{aligned}$$

When $\theta > 2$, the series $\sum_{l=0}^{\infty} \tau_l^{1/3}$ is convergent and we obtain the inequality on $D_{A,m}$ with $c_D = 4\left(c_\tau A K_\infty K_{BV} \sum_{l=0}^{\infty} \tau_l^{1/3}\right)$.
As the models are nested, we only have to compare, for all m in $\mathcal{M}_n$, $b^2_{A,m}$ and $v^2_{A,m}$ with 2^{J_m}. From **[T2]**, $b^2_m \leq \Phi^2 2^{J_m}$, this proves the inequality on $b^2_{A,m,m'}$ with $c_b = \Phi^2$.
For all t in B_m,

$$q\operatorname{Var}(L_q(t)(A_0)) \leq 2 \sum_{l=1}^{q} |\operatorname{Cov}(t(X_1), t(X_l))|. \tag{4.53}$$

Let X_l^* be a random variable, independent of X_1, with law P, such that

$$\mathbb{E}\left(|X_l - X_l^*|\right) \leq \tau_{l-1}.$$

This random variable can be defined thanks to the coupling lemma of Dedecker & Prieur [25] (section 7.1).

$$\begin{aligned}
|\operatorname{Cov}(t(X_1), t(X_l))| &= |\operatorname{Cov}(t(X_1), t(X_l) - t(X_l^*))| \\
&\leq \sqrt{\operatorname{Var}(t(X_1))\mathbb{E}\left((t(X_l) - t(X_l^*))^2\right)} \\
&\leq \sqrt{2\operatorname{Var}(t(X_1)) \|t\|_\infty \mathbb{E}\left(|t(X_l) - t(X_l^*)|\right)} \\
&\leq \sqrt{2\|s\| \|t\|_\infty^2 \operatorname{Lip}(t)\tau_{l-1}}.
\end{aligned}$$

Since t belongs to B_m, $\|t\|_\infty^2 \leq \Phi^2 2^{J_m}$. Moreover, let $a_{j,k} = \int_{\mathbb{R}} t\psi_{j,k} d\mu$, then

$$\begin{aligned} \mathrm{Lip}(t) &= \sup_{x \neq y \in \mathbb{R}} \frac{|t(x) - t(y)|}{|x-y|} \leq \sum_{j=0}^{J_m} \sup_{x \neq y \in \mathbb{R}} \sum_{k \in \mathbb{Z}} |a_{j,k}| \frac{|\psi_{j,k}(x) - \psi_{j,k}(y)|}{|x-y|} \\ &\leq 2AK_L \sum_{j=0}^{J_m} 2^{3j/2} \sup_{k \in \mathbb{Z}} |a_{j,k}|. \end{aligned} \tag{4.54}$$

The last inequality holds since, for all x, y in $\mathbb{R}$ there is less than $2A$ indices k in $\mathbb{Z}$ such that $|\psi_{j,k}(x) - \psi_{j,k}(y)| \neq 0$. Since t belongs to B_m, $\sum_{(j,k) \in m} a_{j,k}^2 \leq 1$, in particular, for all j, $\sup_{k \in \mathbb{Z}} |a_{j,k}| \leq 1$. Thus, there exists a constant c such that $\mathrm{Lip}(t) \leq c2^{3J_m/2}$. Hence, there exists a constant c such that, for all t in B_m and all l in $\mathbb{N}^*$

$$|\mathrm{Cov}(t(X_1), t(X_l))| \leq c2^{5J_m/4}\sqrt{\tau_{l-1}}.$$

Remark that we also have

$$|\mathrm{Cov}(t(X_1), t(X_l))| \leq \|t\|_\infty \|t\| \|s\| \leq c2^{J_m/2}.$$

Recall that $u = 3/(1+\theta)$, there exist constants c, which may vary from line to line such that

$$\begin{aligned} \sum_{l=1}^{q} |\mathrm{Cov}(t(X_1), t(X_l))| &\leq c2^{J_m/2} \sum_{l=1}^{\infty} (2^{3J_m/4}\sqrt{\tau_{l-1}} \wedge 1) \\ &\leq c2^{J_m/2} \sum_{l=1}^{\infty} (2^{3J_m/4} l^{-(1+\theta)/2} \wedge 1) \\ &\leq c2^{J_m/2} \left(\sum_{l=1}^{2^{uJ_m/2}} 1 + \sum_{l=2^{uJ_m/2}}^{\infty} 2^{3J_m/4} l^{-(1+\theta)/2} \right) \\ &\leq c2^{\frac{J_m}{2}(1+u)}. \end{aligned}$$

We deduce the inequality on $v_{A,m,m'}^2$ from (4.53) and this last inequality.
It remains to control MC_n, recall that

$$MC_n = \sum_{m \in \mathcal{M}_n} \left(\left\| \sum_{\lambda \in m} |\psi_\lambda| \right\|_\infty \sup_{\lambda \in m} \mathrm{Lip}_d(\psi_\lambda) + \|s\| |\mathcal{M}_n| \sup_{t \in B_m} \mathrm{Lip}(t) \right).$$

From (4.20, 4.21), for all m in $\mathcal{M}_n$,

$$\left\| \sum_{\lambda \in m} |\psi_\lambda| \right\|_\infty \leq \frac{\sqrt{2}}{\sqrt{2}-1} AK_\infty 2^{J_m/2}, \ \sup_{\lambda \in m} \mathrm{Lip}_d(\psi_\lambda) \leq K_L 2^{3J_m/2}.$$

From (4.54), for all m in $\mathcal{M}_n$, there exists a constant c such that $\sup_{t \in B_m} \mathrm{Lip}(t) \leq c2^{3J_m/2}$. Since $\mathrm{Card}(\mathcal{M}_n) \leq \ln n / \ln 2$, and $2^{\max_{m \in \mathcal{M}_n} J_m} \leq n$, there exists a constant c_M such that $MC_n \leq c_M n^2$.

4.6 Appendix

In this section, we recall some technical lemmas proved in Chapter 2.

Lemma 4.6.1 *For all $\alpha \geq 0$, $K > \alpha + 1$,*

$$\Sigma(K,\alpha) = \sum_{k\in\mathbb{N}} \sum_{m\in\mathcal{M}_n^k} (1+k)^\alpha e^{-K[\ln(1+\mathrm{Card}(\mathcal{M}_\mathrm{n}^\mathrm{k}))+\ln(1+k)]} < \infty.$$

For all m in $\mathcal{M}_n$, let $l_m = l_{n,\gamma}(R_{A,m}, R_{A,m})$. Then, for all $\alpha \geq 0$, for all $K > \sqrt{(1+\alpha)/2}$,

$$\sum_{m\in\mathcal{M}_n} R_{A,m}^\alpha e^{K^2 l_m} \leq \Sigma(K^2,\alpha) e^{-K^2(\ln n)^\gamma}.$$

For all m in $\mathcal{M}_n$, let $l_{m,m'} = l_{n,\gamma}(R_{A,m}, R_{A,m'})$. Then, for all $\alpha \geq 0$, $\alpha' \geq 0$ and all $K > \sqrt{1+\alpha\vee\alpha'}$,

$$\sum_{(m,m')\in(\mathcal{M}_n)^2} R_{A,m}^\alpha R_{A,m'}^{\alpha'} e^{-K^2 l_{m,m'}} = \Sigma(K^2,\alpha)\Sigma(K^2,\alpha') e^{-K^2(\ln n)^\gamma}.$$

Lemma 4.6.2 *Let n be an integer and let $X_1, ..., X_n$ be identically distributed random variables valued in a measurable space $(\mathbb{X}, \mathcal{X})$ with common law P. Let $(t_\lambda)_{\lambda\in\Lambda}$ be a collection of functions in $L^2(\mu)$. Let $p(\Lambda) = \sum_{\lambda\in\Lambda}(\nu_n(t_\lambda))^2$. Let $(W_1, ..., W_n)$ be a resampling scheme, let $\bar{W}_n = \sum_{i=1}^n W_i/n$ and let $v_W^2 = Var(W_1 - \bar{W}_n)$. Let*

$$p^W(\Lambda) = (v_W^2)^{-1} \sum_{\lambda\in\Lambda} \mathbb{E}^W\left((\nu_n^W(t_\lambda))^2\right),$$

$T = \sum_{\lambda\in\Lambda}(t_\lambda - Pt_\lambda)^2$ and

$$U = \frac{1}{n(n-1)} \sum_{i\neq j=1}^n \sum_{\lambda\in\Lambda} (t_\lambda(X_i) - Pt_\lambda)(t_\lambda(X_j) - Pt_\lambda).$$

Then

$$p(\Lambda) = \frac{1}{n}P_n T + \frac{n-1}{n}U,\ p^W(\Lambda) = \frac{1}{n}P_n T - \frac{1}{n}U,\ p(\Lambda) - p^W(\Lambda) = U.$$

Proposition 4.6.3 *Let $X, X_1, ..., X_n$ be i.i.d random variables with common law P. Let B be a symetric class of functions bounded by b. Let $Z = \sup_{t\in B}(\nu_n t)$, $\epsilon = b^2/n$, $v^2 = \sup_{t\in B} Var(t(X))$, $D = \mathbb{E}\left(\sup_{t\in B}(t(X) - Pt)^2\right)$. For all $x > 0$, we have*

$$\mathbb{P}\left(Z^2 - \frac{D}{n} > \frac{D^{3/4}(\epsilon(19x)^2)^{1/4} + 3\sqrt{Dv^2x} + 3v^2x + \epsilon(19x)^2}{n}\right) \leq e^{-x}.$$

$$\mathbb{P}\left(Z^2 - \frac{D}{n} < -\frac{8D^{3/4}(\epsilon x^2)^{1/4} + 7.61\sqrt{v^2Dx} + \epsilon(40.25x)^2}{n}\right) \leq 2.8e^{-x}.$$

Lemma 4.6.4 *Let $X, X_1, ..., X_n$ be i.i.d random variables taking value in a measurable space $(\mathbb{X}, \mathcal{X})$ with common law P. Let μ be a measure on $(\mathbb{X}, \mathcal{X})$ and let $(t_\lambda)_{\lambda\in\Lambda}$ be a set of functions in $L^2(\mu)$. Let $B = \{t = \sum_{\lambda\in\Lambda} a_\lambda t_\lambda,\ \sum_{\lambda\in\Lambda} a_\lambda^2 \leq 1\}$, $D = \mathbb{E}\left(\sup_{t\in B}(t(X) - Pt)^2\right)$, $v^2 = \sup_{t\in B} Var(t(X))$, $b = \sup_{t\in B} \|t\|_\infty$ and $\epsilon = b^2/n$. Let*

$$U = \frac{1}{n(n-1)} \sum_{i\neq j=1}^{n} \sum_{\lambda\in\Lambda} (t_\lambda(X_i) - Pt_\lambda)(t_\lambda(X_j) - Pt_\lambda).$$

Then the following inequality holds

$$\forall x > 0,\ \mathbb{P}\left(U > \frac{5.31D^{3/4}(\epsilon x^2)^{1/4} + 3\sqrt{v^2Dx} + 3v^2x + \epsilon(19.1x)^2}{n-1}\right) \leq 2e^{-x}.$$

$$\forall x > 0,\ \mathbb{P}\left(U < -\frac{9D^{3/4}(\epsilon x^2)^{1/4} + 7.61\sqrt{v^2Dx} + \epsilon(40.3x)^2}{n-1}\right) \leq 3.8e^{-x}.$$

Lemma 4.6.5 *Let $X, X_1, ..., X_n$ be i.i.d random variables taking value in a measurable space $(\mathbb{X}, \mathcal{X})$ with common law P. Let μ be a measure on $(\mathbb{X}, \mathcal{X})$ and let $(\psi_\lambda)_{\lambda\in\Lambda}$ be an orthonormal system in $L^2(\mu)$. Let L be a linear functional in $L^2(\mu)$ and let $B = \{t = \sum_{\lambda\in\Lambda} a_\lambda L(\psi_\lambda),\ \sum_{\lambda\in\Lambda} a_\lambda^2 \leq 1\}$, $v^2 = \sup_{t\in B} Var(t(X))$, $b = \sup_{t\in B} \|t\|_\infty$ and $\epsilon = b^2/n$. Let s be a function in S, the linear space spanned by the functions $(t_\lambda)_{\lambda\in\Lambda}$ and let $\eta > 0$. Then the following inequality holds*

$$\forall x > 0,\ \mathbb{P}\left(\nu_n(L(s)) > \frac{\eta}{2}\|s\|^2 + \frac{2v^2x + \epsilon x^2/9}{\eta n}\right) \leq e^{-x}.$$

Chapter 5

Confidence balls in density estimation

Abstract:

We build non asymptotic confidence balls in L^2-norm for the unknown density s of a real valued random variable X. We first use resampling methods to obtain confidence balls for the orthogonal projection of s onto a finite dimensional linear space. We give an application of this result to model selection theory. Then we investigate the problem of adaptation over a collection of linear subaspaces for confidence balls. We build adaptive confidence balls when it is possible. We deduce adaptive confidence balls over a class of Besov balls. We prove lower bounds and show the optimality of all our results.

Key words: Confidence Balls, Density estimation, Resampling methods, Hypothesis testing.
AMS subject classification: 62G07, 62G09, 62G10, 62G15.

5.1 Introduction

Let $X, X_1, ..., X_n$ be i.i.d. real valued random variables, defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, with common law P. We assume that P has a density s with respect to a positive measure μ and that s belongs to $L^2(\mu)$. In this article, we consider the problem of adaptation for confidence balls in L^2-norm for s. More precisely, let $(S_m)_{m\in\mathcal{M}_n}$ and S be a collection of subsets of $L^2(\mu)$ such that, for all m in $\mathcal{M}_n$, $S_m \subset S$. Let α be a real number in $(0,1)$ and let $\|.\|_2$ be the usual L^2-norm on $L^2(\mu)$. Let $\hat{s}$, $\hat{r}$ be random variables, measurable with respect to $\sigma(X_1, ..., X_n)$, valued in $L^2(\mu)$ and in $\mathbb{R}$ respectively. We say that the L^2-random ball $B(\hat{s}, \hat{r})$ is a $(1-\alpha)$-confidence ball on S if it satisfies the following covering property

$$\forall s \in S,\ \mathbb{P}\left(\|s-\hat{s}\|_2 \leq \hat{r}\right) \geq 1-\alpha. \tag{5.1}$$

Let $B(\hat{s}, \hat{r})$ be a $(1-\alpha)$-confidence ball on S_m. We say that $B(\hat{s}, \hat{r})$ is optimal in expectation, respectively in probability, on S_m if there exists a constant C such that

$$\sup_{s\in S_m} \mathbb{E}(\hat{r}^2) \leq C \inf_{\tilde{r}} \sup_{s\in S_m} \mathbb{E}(\tilde{r}^2), \tag{5.2}$$

respectively if, there exists a constant C such that, for all β in $(0,1)$ the supremum q_β over S_m of the $(1-\beta)$-quantiles of $\hat{r}^2$ satisfies

$$\inf_{\tilde{r}} \sup_{s\in S_m} \mathbb{P}(\tilde{r}^2 > Cq_\beta) \leq \beta. \tag{5.3}$$

In (5.2) and (5.3), the infimum is taken over all the radius $\tilde{r}$ of $(1-\alpha)$-confidence balls on S_m. When a confidence ball is optimal in expectation and in probability, we simply say that it is optimal. A confidence ball $B(\hat{s},\hat{r})$ on S is adaptive over the collection $(S_m)_{m\in\mathcal{M}_n}$ if there exists a constant C such that, for all m in $\mathcal{M}_n$, (5.2) and (5.3) hold.

This problem has been studied in various frameworks. Hoffmann & Lepskii [38], Cai & Low [20] and Juditsky & Lambert-Lacroix [43] worked in the Gaussian white noise model, Li [50] and Baraud [8] in the finite dimensional Gaussian regression model, Beran [11], Beran & Dümberg [12] and Genovese & Wasserman [36] in the fixed design regression framework and Robins & Van der Vaart [61] in a general setting including the density estimation framework. Baraud [8] built non-asymptotic confidence balls in Euclidian norm for a $\mathbb{R}^n$-vector and obtained (5.1) on the whole space $S=\mathbb{R}^n$. In infinite dimensional settings like the Gaussian white noise framework or the density estimation framework, one cannot choose S as the all set $L^2(\mu)$ if we want to ensure that $\hat{r}$ is almost surely finite. It is classical to suppose that S is some ball in a Sobolev or a Besov space. Approximation theory allows then to work in finite dimensional spaces. This idea was developed to build asymptotic confidence balls, see for example Hoffmann & Lepskii [38], Juditsky & Lambert-Lacroix [43], Robins & Van der Vaart [61] and the references therein. In this paper, we always assume that S is a linear space with finite dimension D. As our results are non asymptotic, this dimension can grow with the sample size n and we can prove results on Besov balls (see Section 3.3.1).

First, we consider the problem of confidence balls on a linear space S_m with finite dimension D_m. In Section 2.2, we propose a procedure to obtain such balls. A natural choice for the center of the ball is the least-squares estimator $\hat{s}_m$ of s onto S_m. Thus, (5.1) holds for any upper bound $\hat{r}^2$ on the $(1-\alpha)$-quantile of $\|\hat{s}_m - s\|_2^2$. $\|\hat{s}_m - s\|_2^2$ is a functional of the centered empirical process ν_n. Efron's resampling heuristic (see Efron [30]) provides a natural estimator of this quantity. The difference U_m between $\|\hat{s}_m - s\|_2^2$ and this estimator is a totally degenerate U-statistic of order 2. We give upper bounds on U_m. They are derived from a concentration inequality for U-statistics of order 2 proved by Houdré & Reynaud-Bouret [39]. We deduce $(1-\alpha)$-confidence balls $B(\hat{s},\hat{r})$ for s in S_m. $\mathbb{E}(\hat{r}^2)$ and the quantiles of $\hat{r}^2$ are upper bounded by CD_m/n.

In Section 2.4, we study the optimality of these balls. We give an example of space S_m, where the radius $\tilde{r}$ of any confidence ball satisfies $\mathbb{E}(\tilde{r}^2) \geq CD_m/n$. This proves that the bounds given in Corollary 5.2.2 cannot be improved in general. Section 2.5 is devoted to a short simulation study. In a first part, we illustrate the main theorem of Section 2, then we test another procedure to obtain confidence balls, still based on resampling ideas. In Section 2.6, we introduce an application to model selection that will be fully developed in a forthcoming article. Resampling based confidence balls have recently been studied from a non-asymptotic point of view in the regression framework by Arlot, Blanchard and Roquain [6]. Resampling penalties have been used in model selection by Arlot [5] in the regression framework and by Fromont

[32] in classification.
In Section 3, we consider the problem of adaptation for confidence balls. In order to build such balls, we need informations on the distance between s_n and S_m, where s_n denotes the projection of s onto S. The necessary estimation of this bias leads to an irreductible limitation in the adaptive property. More precisely, we prove in Theorem 5.3.1 that all $(1-\alpha)$-confidence balls $B(\hat{s},\hat{r})$ over S satisfy $\mathbb{E}(\hat{r}^2) \geq C\sqrt{D}/n$. From Section 2, for all s in S_m, an adaptive confidence ball should satisfy $\mathbb{E}(\hat{r}^2) \leq CD_m/n$. Thus, we can build confidence balls on S that are adaptive over $(S_m)_{m\in\mathcal{M}_n}$ only if all the models S_m satisfy $D_m \geq C\sqrt{D}$. The main problem is that we cannot build a test of null hypothesis $s \in S_m$ against the alternative $s \in S$ and $s \notin S_m$ with asymptotic rate of convergence (as defined in Ingster [40]-[41]-[42]) smaller than $C\sqrt{D}/n$. Then, we discuss the link with adaptive estimation. Actually, the problem of adaptive confidence balls can be viewed as the problem of giving empirical bounds on $\|\tilde{s}-s\|_2^2$ for an adaptive estimator $\tilde{s}$ as mentioned in Section 3.1.2 and fully discussed in Hoffmann & Lepskii [38]. We adapt Baraud's model selection algorithm to the density estimation framework and obtain non asymptotic confidence balls on S, adaptive over $(S_m)_{m\in\mathcal{M}_n}$ when it is possible. Finally, we derive some applications, in particular, we build non asymptotic confidence balls on Besov balls $B_{w,2,\infty}(M)$ (for a precise definition see Section 3.3.1). In this framework, the problem of adaptation can be stated as follows. It is well known that

$$\forall w > 0,\ \lim_{n\to\infty} n^{2w/(2w+1)} \inf_{\hat{s}} \sup_{s\in B_{w,2,\infty}(M)} \mathbb{E}\,\|s-\hat{s}\|_2^2 = C_M.$$

Therefore, there is no hope to build confidence balls on $B_{w,2,\infty}(M)$ with radius $\hat{r}$ satisfying asymptotically $\mathbb{E}\hat{r}^2 \leq C_M n^{-2w/(2w+1)}$. A $(1-\alpha)$-confidence ball on $B_{w,2,\infty}(M)$ is adaptive over the class of Besov balls $(B_{v,2,\infty}(M))_{v\geq w}$ if, for all $v \geq w$, $\sup_{s\in B_{v,2,\infty}(M)} \mathbb{E}\hat{r}^2 \leq C_M n^{-2v/(2v+1)}$. Robins & Van der Vaart [61] proved that all $(1-\alpha)$-confidence balls over $B_{w,2,\infty}(M)$ satisfy $\mathbb{E}\hat{r}^2 \geq C_M\left(n^{-2v/(2v+1)} \vee n^{-4w/(4w+1)}\right)$. Thus, we can build $(1-\alpha)$-confidence balls on $B_{w,2,\infty}(M)$ that are adaptive only over the class $(B_{v,2,\infty}(M))_{w\leq v\leq 2w}$. Moreover, Robins & Van der Vaart built such adaptive $(1-\alpha)$-confidence balls. The same phenomenon occurs in various statistical frameworks, see for example, Hoffman & Lepskii [38] and Juditsky & Lambert-Lacroix [43], Robins & Van der Vaart [61] and Baraud [8].
In this paper, we consider the problem of adaptive confidence balls in L^2-norm. It is not clear if this work can be done for other norms. Low [52] proved for density estimation at a single point, that fixed radius confidence intervals perform as well as random length intervals, that is, the data do not help to reduce the length of the balls. In particular, adaptation is impossible for confidence balls in L^∞-norm. This result has been extended by Genovese and Wasserman [35] to non parametric regression and other frameworks. They proved that adaptation on $S_m \subset S$ is impossible because of the functions that are close to S_m in L^2 norm but far in L^∞ norm, and develop a theory of adaptation with a weaker definition of covering (in particular, they did not guarantee the covering property (5.1) on the whole space S).
The paper is organized as follows. In Section 2, we study the problem of confidence balls on a finite dimensional linear subspace S_m. First, we present our resampling algorithm to estimate $\|s-\hat{s}_m\|_2^2$. Then, in Section 2.3, we give upper bounds on the difference between $\|s-\hat{s}_m\|_2^2$ and this estimator and we use them to build our

confidence balls. We prove the optimality of this procedure in Section 2.4. In a short simulation study, we compute the difference between $\|s - \hat{s}_m\|_2^2$ and its resampled estimator and remark that the upper bounds given in Theorem 5.2.1 seem to be sharp. Finally, we give an application of our main theorem in model selection theory.
In Section 3, we study the problem of adaptation for confidence balls. We prove lower bounds on the size of an adaptive confidence ball and use it to discuss the links with adaptive estimation. Then, we build an optimal test for the null hypothesis $s \in S_m$ against the alternative $s \in S$ and $s \notin S_m$, we use it to extend Baraud's algorithm to the density estimation framework and we derive adaptive confidence balls. These balls happen to be optimal in view of the lower bounds. We conclude the paper with two applications. We give adaptive confidence balls for a regular density and for the vector $(\mathbb{P}(X \in I_\lambda))_{\lambda \in \Lambda}$, where $(I_\lambda)_{\lambda\in\Lambda}$ is a partition of $\mathbb{X}$. All the proofs are postponed to Section 4.

5.2 Confidence balls for s_m by resampling methods

5.2.1 Definition and assumptions

Hereafter, $\mathbb{X}$ denotes a measurable subspace of $\mathbb{R}$ and $\mathcal{X}$ denotes the Borel σ-algebra on $\mathbb{X}$. Let μ be a fixed measure on $(\mathbb{X}, \mathcal{X})$. Let $L^2(\mu)$ be the linear space of functions t from $\mathbb{X}$ to $\mathbb{R}$ such that $\int_{\mathbb{X}} t^2 d\mu < \infty$. For all linear subspaces S in $L^2(\mu)$, we denote by $\bar{S}$ the set of all densities s with respect to μ in S, that is, the set of all non negative functions s in S such that $\int_{\mathbb{X}} s d\mu = 1$. Let $X, X_1, X_2, ..., X_n$ be iid random variables defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, taking value in $(\mathbb{X}, \mathcal{X})$, with common law $dP_s = s d\mu$ for some s in $L^2(\mu)$. Let P_s, P_n and ν_n be the following processes, defined for every functions t in $L^2(\mu)$ by $P_s t = \mathbb{E}t(X) = \int_x t s d\mu$, $P_n t = \sum_{i=1}^n t(X_i)/n$ and $\nu_n t = (P_n - P_s)t$. P_s is defined on $L^2(\mu)$ thanks to Cauchy-Schwarz inequality. When no confusion can occur, we simply write P instead of P_s. Let $\mathbb{P}_s$ denote the product measure $P_s^{\otimes \mathbb{N}}$.
Let S_m be a linear subspace of $L^2(\mu)$, with finite dimension D_m. We assume that S_m satisfies the following property.
$[M_1]$: $\sqrt{D_m} \leq n$, the constant functions belong to S_m and, for all t in S_m, $\|t\|_\infty \leq \Phi_0 \sqrt{D_m}\, \|t\|_2$.
Four examples are usually developed as fulfilling this set of assumptions:
[H] histogram spaces: S_m is the space of all the functions constant on a finite partition $(I_\lambda)_{\lambda \in m}$ of $\mathbb{X}$ such that, for all λ in m, $\mu(I_\lambda) \geq 1/(\Phi_0^2 |m|)$. $D_m = |m|$.
In the three other examples, $\mathbb{X} = [0,1]$ and μ is the Lebesgue measure on $(\mathbb{X}, \mathcal{X})$.
[T] trigonometric spaces: S_m is the linear span of $\psi_{0,0}(x) = 1$, $\psi_{j,1}(x) = \cos(2\pi jx)$ and $\psi_{j',2}(x) = \sin(2\pi jx)$ for all $1 \leq j, j' \leq J_m$. $D_m = 2J_m + 1$;
[P] regular piecewise polynomial spaces: S_m is the linear span of the functions $(\psi_{j,k})$ for $j = 1, ..., J_m$, $k = 0, ..., r-1$, where, for all $j = 1, ..., J_m$ and $k = 0, ..., r-1$, $\psi_{j,k}$ is a polynomial of degree k on $[(j-1)/J_m, j/J_m]$. $D_m = rJ_m$;
[W] spaces spanned by dyadic wavelets with regularity r as described in Section 3.3.
For a precise description of those spaces and their properties, we refer to Birgé & Massart [15].

5.2.2 Resampling estimators of the quantiles of $\|s_m - \hat{s}_m\|_2^2$

Let s_m be the orthogonal projection of s onto S_m and let $\hat{s}_m$ be the projection estimator of s onto S_m. In this section, for a real number α in $(0,1)$, we want to estimate the $(1-\alpha)$-quantile of $\|s_m - \hat{s}_m\|_2^2$. Let $(\psi_\lambda)_{\lambda\in m}$ be an orthonormal basis of S_m. We have $s_m = \sum_{\lambda\in m}(P\psi_\lambda)\psi_\lambda$, $\hat{s}_m = \sum_{\lambda\in m}(P_n\psi_\lambda)\psi_\lambda$ and

$$\|s_m - \hat{s}_m\|_2^2 = \sum_{\lambda\in m}(P\psi_\lambda - P_n\psi_\lambda)^2 = \sum_{\lambda\in m}[\nu_n\psi_\lambda]^2.$$

In order to estimate the quantiles of this functional $F(\nu_n)$, we apply Efron's resampling heuristic. We introduce a resampling scheme $W_1, ...W_n$, ie a vector of random variables independent of $X, X_1, ..., X_n$ and exchangeable, that is, for all permutations τ of $(1, ..., n)$,

$$(W_1, ..., W_n) \text{ has the same law as } (W_{\tau(1)}, ..., W_{\tau(n)}).$$

Let $\bar{W}_n = \sum_{1=1}^n W_i/n$. For all functions t, let $P_n^W t = \sum_{i=1}^n W_i t(X_i)/n$ and $\nu_n^W t = (P_n^W - \bar{W}_n P_n)t$. Let $\mathbb{E}^W$ and $\mathcal{L}^W$ be respectively the expectation and the law with respect to the resampling scheme $W_1, ..., W_n$, i.e. conditionally to $X_1, ..., X_n$. Efron's heuristic states that the law of the functional $F(\nu_n)$ is close to its resampled conterpart, that is $\mathcal{L}^W(F[C_W\nu_n^W])$ where C_W is a normalizing constant depending only on the resampling scheme $W_1, ..., W_n$ (see Theorem 5.2.1). We can use this heuristic in two different ways.

The quantiles of the law of $\|s_m - \hat{s}_m\|_2^2$ can be estimated by those of the conditional law $\mathcal{L}^W\left(C_W\sum_{\lambda\in m}[(\nu_n^W)\psi_\lambda]^2\right)$. We do not use this estimation in this note. Nevertheless, we show in a short simulation study that it seems to give very good results in practice.

We use another approach based on the concentration of measure phenomenon. In Lemma 5.4.3, we prove a concentration inequality for $\|s_m - \hat{s}_m\|_2^2$ around its expectation. Basically, it says that, with large probability, $\|s_m - \hat{s}_m\|_2^2$ is close to $\mathbb{E}(\|s_m - \hat{s}_m\|_2^2) = \mathbb{E}(\sum_{\lambda\in m}[\nu_n(\psi_\lambda)]^2)$. Hence, an estimator of $\|s_m - \hat{s}_m\|_2^2$ is the resampled conterpart of this expectation, that is $\mathbb{E}^W\left(C_W\sum_{\lambda\in m}[\nu_n^W(\psi_\lambda)]^2\right)$. Then, we have to control the error made in this estimation. In the next section, we obtain this control thanks to a concentration inequality (see Lemma 5.26) for U-statistics based on an orthonormal system.

5.2.3 Confidence balls on S_m

Let us first state the main result of this Section. It is a concentration inequality for $\|s_m - \hat{s}_m\|_2^2$ around the resampling estimator of its expectation.

Theorem 5.2.1 *Let $X_1, ..., X_n$ be i.i.d real valued random variables with common law P_s. Assume that s belongs to $L^2(\mu)$. Let S_m be a linear subspace in $L^2(\mu)$ with dimension D_m satisfying Assumption $[M_1]$ and let $(\psi_\lambda)_{\lambda\in m}$ be an orthonormal basis of S_m. Let s_m be the orthogonal projection of s onto S_m and let $\hat{s}_m$ be the projection estimator of s onto S_m. Let $W_1, ..., W_n$ be a resampling scheme. Let $v_W^2 = Var(W_1 - \bar{W}_n)$ and*

$$R_W^2 = \frac{1}{v_W^2}\mathbb{E}^W\left(\sum_{\lambda\in m}[\nu_n^W\psi_\lambda]^2\right).$$

For all $\epsilon > 0$, for all s in $\bar{L}^2(\mu)$ and for all $x > 0$, we have

$$\mathbb{P}_s\left(\|s_m - \hat{s}_m\|_2^2 > (1+\epsilon)^3 R_W^2 + \eta(\epsilon)\frac{(\Phi_0\sqrt{D_m}\,\|s\|_2\,x)\vee 1}{n} + K_n(\epsilon)\frac{D_m x^2}{n^2}\right) \le 4.8e^{-x} \tag{5.4}$$

and

$$\mathbb{P}_s\left(\|s_m - \hat{s}_m\|_2^2 > (1+\epsilon)^3 R_W^2 + \eta(\epsilon)\frac{\Phi_0^2 D_m x}{n} + K_n(\epsilon)\frac{D_m x^2}{n^2}\right) \le 4.8e^{-x}, \tag{5.5}$$

where $\eta(\epsilon) = 4(1+\epsilon^{-1}) + (1+\epsilon)^3$ and

$$K_n(\epsilon) = \Phi_0^2\left((1+\epsilon)^3\left(14.7 + \frac{384}{\sqrt{n}} + \frac{1020}{n}\right) + 2(1+\epsilon^{-1})(\epsilon^{-1}+1/3)^2\right).$$

Comments:
Inequality (5.4) gives a control of $\|s_m - \hat{s}_m\|_2^2$ with a remainder term of order $\sqrt{D_m}/n$. This is sufficient to derive sharp oracle inequalities in model selection as mentioned in the next section. However, Inequality (5.4) involves the unknown $\|s\|_2$, thus, we cannot derive from it a confidence ball for s_m unless we have an upper bound on $\|s\|_2$. This is why we give Inequality (5.5) where the control of $\|s_m - \hat{s}_m\|_2^2$ is totally data driven. The main drawback is that the remainder term in (5.5) is of order D_m/n instead of $\sqrt{D_m}/n$. Hence, Inequality (5.5) is too conservative and the confidence balls derived from it also. Nevertheless, we will prove in Section 2.3 that these balls are optimal, at least, up to the constant. Let us now give the confidence balls that we obtain in the following corollary.

Corollary 5.2.2 *Let us keep the notations of Theorem 5.2.1. Let α be a real number in $(0,1)$ such that $\ln(1/\alpha) \le C_* n$ for some constant C_*. Let*

$$r_m(x)^2 = (1+\epsilon)^3 R_W^2 + \eta(\epsilon)\frac{\Phi_0^2 D_m x}{n} + K_n(\epsilon)\frac{D_m x^2}{n^2}.$$

Then, the set $B(\hat{s}_m, r_m(\ln(4.8/\alpha)))$ is a $(1-\alpha)$-confidence ball for s_m, that is

$$\forall s \in \bar{L}^2(\mu),\ \mathbb{P}_s\left(s_m \in B(\hat{s}_m, r_m(\ln(4.8/\alpha)))\right) \ge 1-\alpha. \tag{5.6}$$

There exists a constant C such that $\mathbb{E}\left(r_m^2(\ln(4.8/\alpha))\right) \le C\ln(\alpha^{-1})D_m/n$. Moreover, there exists a constant C such that, for all β in $(0,1)$ such that $\ln(1/\beta) \le C_* n$,

$$\mathbb{P}_s\left(r_m^2(\ln(4.8/\alpha)) \le C\ln(\alpha^{-1}\vee\beta^{-1})\frac{D_m}{n}\right) \ge 1-\beta.$$

Assume that there exists $M \ge 1$ such that $\|s\|_2 \le M$ and let

$$\hat{r}_m(x)^2 = (1+\epsilon)^3 R_W^2 + \eta(\epsilon)\frac{\Phi_0 M\sqrt{D_m}x}{n} + K_n(\epsilon)\frac{D_m x^2}{n^2}.$$

Then, the set $B(\hat{s}_m, \hat{r}_m(\ln(4.8/\alpha)))$ is a $(1-\alpha)$-confidence ball for s_m, that is

$$\forall s \in \bar{L}^2(\mu),\ \mathbb{P}_s\left(s_m \in B(\hat{s}_m, \hat{r}_m(\ln(4.8/\alpha)))\right) \ge 1-\alpha. \tag{5.7}$$

If $\ln(1/\alpha) \le C_* n/\sqrt{D_m}$*, there exists a constant* C *such that*

$$\mathbb{E}\left(\hat{r}_m^2(\ln(4.8/\alpha))\right) \le \frac{C}{n}\left(D_m + \ln(\alpha^{-1})\sqrt{D_m}\right).$$

Moreover, there exists a constant C *such that, for all* β *in* $(0,1)$ *such that* $\ln(1/\beta) \le C_* n/\sqrt{D_m}$*, we have*

$$\mathbb{P}_s\left(\hat{r}_m^2(\ln(4.8/\alpha)) \le \frac{C}{n}(D_m + \ln(\alpha^{-1} \vee \beta^{-1})\sqrt{D_m})\right) \ge 1-\beta. \tag{5.8}$$

Comments:
The upper bounds given on the radius of the confidence balls (5.6) and (5.7) have the same order. Thus there is only a loss in the constants when we use (5.6) instead of (5.7). The main advantage of (5.8) is that the term $\ln(\alpha^{-1} \vee \beta^{-1})\sqrt{D_m}$ will be small compared with D_m in our application in Section 3.3.1.
We can derive from this result two statistical applications:
Let $(I_\lambda)_{\lambda\in m}$ be a finite partition of $\mathbb{X}$, then we can obtain a confidence set for the vector $(\mathbb{P}(X \in I_\lambda))_{\lambda\in m}$.
If we have some regularity assumptions on s, we can obtain a confidence ball for s. These applications will be developed in Section 3.3.

5.2.4 Optimality of the confidence balls

The aim of this section is to prove that Corollary 5.2.2 provides optimal confidence sets for s_m from a minimax point of view. More precisely, we want to prove that the upper bounds given on the radius of our balls cannot be improved in general. In order to see this, let us consider the following particular case.

Proposition 5.2.3 *Let* $\mathbb{X} = [0,1)$ *and let* μ *be the Lebesgue measure on* $\mathbb{X}$*. Let* S_m *be the set of functions constant on the intervals* $[k/D_m, (k+1)/D_m)$*, for all* $k = 0, ..., D_m - 1$*. Let* $\bar{S}_m$ *be the set of all densities with respect to* μ *in* S_m*. Let* $X_1, ..., X_n$ *be i.i.d random variables with common law* P_s*. Let* α, β *be real numbers in* $(0,1)$ *such that* $\alpha+\beta<1$ *and let* $\hat{s}, \hat{r}$ *be random variables satisfying*

$$\forall s \in \bar{S}_m,\ \mathbb{P}_s(s \in B(\hat{s}, \hat{r})) \ge 1-\alpha, \tag{5.9}$$

$$\forall s \in \bar{S}_m,\ \mathbb{P}_s(\hat{r} \le d_m) \ge 1-\beta. \tag{5.10}$$

Then, if $D_m \ge 3 + 18\log(\sqrt{2}/(1-\alpha-\beta))$ *and if* $n \ge D_m + 1$*, we have*

$$d_m^2 \ge \frac{D_m - 1}{6n}.$$

In particular $\inf_{\hat{r}} \sup_{s\in\bar{S}_m} \mathbb{E}_s \hat{r}^2 \ge \beta(D_m-1)/(6n)$.

Comments:
The space S_m described in Proposition 5.2.3 satisfies Assumption $[M_1]$. Thus, we can build confidence balls for s_m thanks to Corollary 5.2.2. These balls satisfy both conditions (5.9) and (5.10) with $d_m^2 = C_{\alpha,\beta} D_m/n$. Proposition 5.2.3 ensures that this bound is optimal, at least, up to the constant $C_{\alpha,\beta}$. This justifies the choice of the least-squares estimator as the center of the ball.
It is well known that the minimax rate of convergence in L^2-norm over S_m is of order D_m/n. Therefore, there is intuitively no hope to build confidence balls over S_m such that $\mathbb{E}(\hat{r}^2) \le C D_m/n$. Proposition 5.2.3 shows that this intuition is correct.

5.2.5 Simulation study

In this section, our first goal is to illustrate Theorem 5.2.1. In Proposition 5.2.3, we proved that the choice of $\hat{s}_m$ as the center of our balls is optimal. Actually, we give upper bounds on $\|s_m - \hat{s}_m\|_2^2$ that are of optimal order D_m/n. Moreover, Inequality (5.4) in Theorem 5.2.1 ensures that R_W^2 is a tight estimate of $\|s_m - \hat{s}_m\|_2^2$ since we provide a control of the difference $\|s_m - \hat{s}_m\|_2^2 - R_W^2$ of order $\sqrt{D_m}/n$, which is smaller than D_m/n. We want here to illustrate that this control seems to be sharp also. Then, we consider the second possible use of Efron heuristics that we mentioned in Section 2.1. Recall that it states that the quantiles of $\|s_m - \hat{s}_m\|_2^2$ are close to their resampled conterpart: the quantiles of the conditional law $\mathcal{L}^W\left(C_W \sum_{\lambda\in m}[(\nu_n^W)\psi_\lambda]^2\right)$. In a second simulation, we test this method and remark that it seems to give very good results.

Illustration of Theorem 5.2.1

In this simulation, μ is the Lebesgue measure on $\mathbb{R}$ and the density $s = 1_{[0,1)}$. S_m is the set of histograms on the partition $([(k-1)/D_m, k/D_m))_{k=1,\dots,D_m}$. The resampling scheme $(W_1, \dots, W_n)$ is given by Efron's weights, which means that the law $\mathcal{L}(W_1, \dots, W_n)$ is the multinomial law $\mathcal{M}(n, 1/n, \dots, 1/n)$. In order to compute R_W^2, we estimate the conditional expectation $\mathbb{E}^W(\sum_{\lambda\in m}[\nu_n^W\psi_\lambda]^2)$ by a Monte Carlo method with nb repetitions. Finally, we repeat $p = 1000$ times the experiment. We plot the histograms of the p values of the normalized difference $n(\|s_m - \hat{s}_m\|_2^2 - R_W^2)/\sqrt{D_m}$ that we obtained. The first histogram is obtained with $n = 50, D_m = 10, nb = 100$ and the second for $n = 200, D_m = 50, nb = 500$. We remark that the law of $n(\|s_m - \hat{s}_m\|_2^2 - R_W^2)/\sqrt{D_m}$ does not seem to change with n or D_m, thus $\sqrt{D_m}/n$ seems to be the correct order of the difference $\|s_m - \hat{s}_m\|_2^2 - R_W^2$. In particular, the control (5.4) in Theorem 5.2.1 seems to be sharp, at least, up to the constant in front of the remainder term.

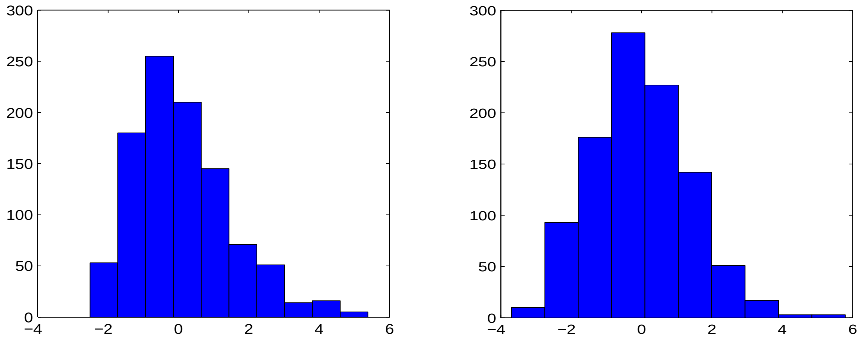


Figure 5.1: $\frac{n}{\sqrt{D_m}}(\|s_m - \hat{s}_m\|_2^2 - R_W^2)$.

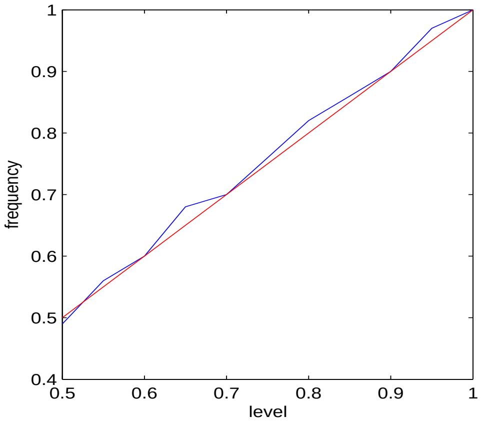
Illustration of the second Efron's heuristic

In this simulation, we keep the same s and the same law of $W_1, \dots, W_n$ as in the previous simulation. S_m is the set of functions constant on the partition $([(k -$

$1)/D_m, k/D_m))_{k=1,...,D_m}$, with $D_m = 50$. $n = 100$, $N = 100$ and $((X_i^J)_{i=1,...,n})_{J=1,...,N}$ are N independent samples with common law $P_s^{\otimes n}$. For all $J = 1, ..., N$, we compute the projection estimator $\hat{s}_m^J$ on S_m with the sample $(X_i^J)_{i=1,...,n}$. Then, we take $nb = 10000$ resampling schemes $(W_1, ..., W_n)$. For all resampling schemes, we compute the quantity

$$Q_{W,m}^J = \frac{1}{v_W^2}\left(\sum_{\lambda\in m}[\nu_n^{J,W}\psi_\lambda]^2\right)$$

and we obtain an approximation of the $(1-\alpha)$-quantiles $\hat{q}_\alpha^J$ of its conditional law $\mathcal{L}^W(Q_{W,m}^J)$. We plot the frequency of J such that $\left\|s_m - \hat{s}_m^J\right\|^2 \leq \hat{q}_\alpha^J$ and the function $f(\alpha) = \alpha$ when α varies in $(0.5, 1)$ in the following curves.



Comments:
We see that the covering property of this empirical ball is very close to the one we would like to obtain. Hence, this method seems to give sharp confidence balls for s_m. However, we do not prove any theoretical evidence of this fact. In particular, we cannot guarantee that the covering property $\mathbb{P}(\|s_m - \hat{s}_m\|_2^2 \leq \hat{q}_\alpha) \geq 1-\alpha$ occurs in general. The practical user can evaluate this estimation of the quantiles of the conditional law $\mathcal{L}^W(Q_{W,m})$ (the computation time is the same as the one needed to evaluate R_W^2). He can compare it with the majorations given in Corollary 5.2.2 and have an idea of the size of $\|s_m - \hat{s}_m\|_2^2$ but the covering property is only guaranteed when one uses a radius given in Corollary 5.2.2.

5.2.6 Application to model selection

Let $(S_m)_{m\in\mathcal{M}_n}$ be a collection of finite dimensional linear subspaces in $L^2(\mu)$ and let $(s_m)_{m\in\mathcal{M}_n}$ and $(\hat{s}_m)_{m\in\mathcal{M}_n}$ be respectively the collection of the orthogonal projections of s and the collection of the projection estimators onto the linear spaces $(S_m)_{m\in\mathcal{M}_n}$. The problem of model selection is to select a model $\hat{m}$ in the collection $\mathcal{M}_n$ satisfying a so called oracle inequality, that is an inequality of the form

$$\mathbb{P}\left(\|s - \hat{s}_{\hat{m}}\|_2^2 \leq C \inf_{m\in\mathcal{M}_n}\left\{\|s - \hat{s}_m\|_2^2 + R(m,n)\right\}\right) \geq 1 - c_n, \qquad (5.11)$$

where C is a constant, $R(m,n)$ is a (possibly random) remainder term and $c_n \to 0$. A classical way to obtain $\hat{m}$ (see for example Birgé & Massart [15]) is to build a function $\mathrm{pen}(m)$ measuring the complexity of the model S_m and to choose

$$\hat{m} \in \arg\min_{m\in\mathcal{M}_n} P_n\gamma(\hat{s}_m) + \mathrm{pen}(m),$$

where for all t in $L^2(\mu)$, $\gamma(t) = \|t\|_2^2 - 2t$. This way we have, for all m in $\mathcal{M}_n$,

$$\begin{aligned} P_n\gamma(\hat{s}_{\hat{m}}) + \mathrm{pen}(\hat{m}) &\leq P_n\gamma(\hat{s}_m) + \mathrm{pen}(m) \\ P\gamma(\hat{s}_{\hat{m}}) + \nu_n\gamma(\hat{s}_{\hat{m}}) + \mathrm{pen}(\hat{m}) &\leq P\gamma(\hat{s}_m) + \nu_n\gamma(\hat{s}_m) + \mathrm{pen}(m) \\ \|\hat{s}_{\hat{m}} - s\|_2^2 - 2\nu_n(\hat{s}_{\hat{m}} - s_{\hat{m}}) + \mathrm{pen}(\hat{m}) &\leq \|\hat{s}_m - s\|_2^2 - 2\nu_n(\hat{s}_m - s_m) \\ &\quad +\mathrm{pen}(m) - 2\nu_n(s_m - s_{\hat{m}}). \qquad (5.12) \end{aligned}$$

The last inequality comes from $P\gamma(t) = \|t - s\|_2^2 - \|s\|_2^2$ and from $\nu_n\gamma(t) = -2\nu_n(t)$ for all t in $L^2(\mu)$.
We can prove with Bernstein's inequality that $\nu_n(s_m - s_{\hat{m}})$ is small compared with $\|\hat{s}_{\hat{m}} - s\|_2^2 + \|\hat{s}_m - s\|_2^2$. Thus, a good penalty is a tight estimate of

$$\mathrm{pen}_{id}(m) = 2\nu_n(\hat{s}_m - s_m) = 2\,\|\hat{s}_m - s_m\|_2^2\,.$$

Assume that, for all m in $\mathcal{M}_n$, $\sqrt{D_m} \leq n$, then we can apply inequality (5.4) and obtain, for $\epsilon > 0$,

$$\forall x > 0,\ \mathbb{P}\left(\|s_m - \hat{s}_m\|_2^2 \leq (1+\epsilon)^3 R_W^2 + (\eta(\epsilon)\Phi_0\,\|s\|_2 + K_n(\epsilon))\,\frac{\sqrt{D_m}}{n}x^2\right) \leq 4.8e^{-x} \qquad (5.13)$$

Suppose that there exists $\gamma > 1$ such that, for all $c > 0$, the series $\sum_{m\in\mathcal{M}_n} e^{-c(\log(D_m))^\gamma}$ is convergent. Then, inequality (5.13) applied with

$$x = \frac{(\log\sqrt{D_m})^\gamma + (\log n)^\gamma}{\sqrt{\eta(\epsilon)\Phi_0\,\|s\|_2 + K_n(\epsilon)}}$$

gives

$$\mathbb{P}\left(\forall m \in \mathcal{M}_n,\ \|s_m - \hat{s}_m\|_2^2 \leq (1+\epsilon)^2 R_W^2 + \frac{4\sqrt{D_m}(\log n)^{2\gamma}}{n}\right) \leq L_{s,\epsilon}e^{-c_{s,\epsilon}(\log n)^\gamma},$$

with $L_{s,\epsilon} > 0$ and $c_{s,\epsilon} > 0$. In particular, (5.12) applied with

$$\mathrm{pen}(m) = 2(1+\epsilon)^3 R_W^2 + \frac{8\sqrt{D_m}(\log n)^{2\gamma}}{n},$$

leads to

$$\mathbb{P}\left(\|s - \hat{s}_{\hat{m}}\|_2^2 + 2\nu_n s_{\hat{m}} \leq \inf_{m\in\mathcal{M}_n}\left\{\|s - \hat{s}_m\|_2^2 + R(m,n)\right\}\right) \geq 1 - c_n, \qquad (5.14)$$

where $R(m,n) = \mathrm{pen}(m) - 2\,\|s_m - \hat{s}_m\|^2 - 2\nu_n(s_m)$ and $c_n = L_{s,\epsilon}e^{-c_{s,\epsilon}(\log n)^\gamma}$. We can derive an oracle inequality from (5.14) thanks to a control of $\nu_n s_{\hat{m}}$ by Bernstein's inequality. Moreover, we can prove a lower bound on our penalty and obtain under reasonable assumptions on the collection $(S_m)_{m\in\mathcal{M}_n}$ that, with large probability, $|R(m,n)| = \epsilon_n\,\|s - \hat{s}_m\|_2^2$ with $\epsilon_n \to 0$. The proof of these results is beyond the scope of this paper. It will be fully developed in a forthcoming article (see Chapter 2).

5.3 Adaptive confidence balls

Let S_m be a linear subspace of $L^2(\mu)$ with finite dimension D_m and let α be a real number in $(0,1)$. In Section 2, we proved that any $(1-\alpha)$-confidence ball of s_m, i.e. any random set $B(\hat{s},\hat{r}_\alpha)$ such that

$$\forall s \in \bar{L}^2(\mu),\ \mathbb{P}_s\left(s_m \in B(\hat{s},\hat{r}_\alpha)\right) \geq 1-\alpha$$

satisfies $\mathbb{E}(\hat{r}_\alpha^2) \geq C_\alpha D_m/n$. Moreover, we built $(1-\alpha)$-confidence balls such that $\mathbb{E}(\hat{r}_\alpha^2) \leq C_\alpha D_m/n$. In this framework, the problem of adaptation can be stated as follows. Let S be a linear subspace in $L^2(\mu)$ with dimension D and let $(S_m)_{m\in\mathcal{M}_n}$ be a collection of linear subspaces in S. A $(1-\alpha)$-confidence ball $B(\hat{s},\hat{r}_\alpha)$ for the projection s_n of s onto S is said to be adaptive over the collection $(S_m)_{m\in\mathcal{M}_n}$ if there exists a constant C_α such that, for all m in $\mathcal{M}_n$, for all s in $L^2(\mu)$ such that s_n belongs to S_m, we have $\mathbb{E}(\hat{r}_\alpha^2) \leq C_\alpha D_m/n$. In order to build such balls, we need informations on the distance between s_n and S_m $d(s_n,S_m) = \inf_{t\in S_m}\|s-t\|_2$. The necessary estimation of this bias leads to an irreductible limitation in the adaptive property. More precisely, we will prove in Theorem 5.3.1 that all $(1-\alpha)$-confidence balls $B(\hat{s},\hat{r})$ over S satisfy $\mathbb{E}(\hat{r}^2) \geq C(D_m \vee \sqrt{D})/n$. Thus, we can build confidence balls on S that are adaptive over $(S_m)_{m\in\mathcal{M}_n}$ only if all the models S_m satisfy $D_m \geq C\sqrt{D}$. The main problem is that we cannot build a test of null hypothesis $s \in S_m$ against the alternative $s \in S$ and $s \notin S_m$ with asymptotic rate of convergence (as defined in Ingster [40, 41, 42] smaller than $C\sqrt{D}/n$. In this section, we prove non asymptotic lower bounds on the radius of a confidence ball in our finite dimensional setting. Then, we adapt Baraud's model selection algorithm to the density estimation framework and obtain non asymptotic adaptive confidence balls for s_n. Finally, we derive some applications, in particular, we build non asymptotic confidence balls for s under regularity assumptions.

5.3.1 Minimax lower bounds

The result

Theorem 5.3.1 *Let $\mathbb{X} = [0,1)$ and let μ be the Lebesgue measure on $\mathbb{X}$. Let S be the linear subspace of $L^2(\mu)$ spanned by the functions $1_{[k/D,(k+1)/D)}$ for $k = 0,...,D-1$ and let S_m be the linear subspace of S with dimension D_m spanned by the functions $1_{[k/D_m,(k+1)/D_m)}$ for $k = 0,...,D_m-1$. We assume that there exists an integer l such that $D = lD_m$, thus S_m is a subset of S. As before, we denote by $\bar{S}$, respectively $\bar{S}_m$, the set of all densities in S, respectively in S_m. Let α,β be real numbers in $(0,1)$ such that $2\alpha+\beta < 1$. Let $(X_1,...,X_n)$ be i.i.d random variables with common law P_s and let $\hat{s},\hat{r}$ be random variables measurable with respect to $\sigma(X_1,...,X_n)$ such that*

$$\forall s \in \bar{S},\ \mathbb{P}_s(s \in B(\hat{s},\hat{r})) \geq 1-\alpha, \tag{5.15}$$

$$\forall s \in \bar{S}_m,\ \mathbb{P}_s(\hat{r} \leq r_m) \geq 1-\beta. \tag{5.16}$$

Assume that $D \geq 10$ and that $n \geq \sqrt{D}$. Then, there exists a positive constant $C_{\alpha,\beta}$ such that

$$r_m^2 \geq C_{\alpha,\beta}\frac{\sqrt{D}}{n}.$$

In particular, $\inf_{\hat{r}}\sup_{s\in S_m}\mathbb{E}_s\hat{r}^2 \geq \beta C_{\alpha,\beta}\sqrt{D}/n$.

Comments:
We want to build confidence balls $B(\hat{s},\hat{r})$ on $\bar{S}$ such that, when s belongs to $\bar{S}_m$, $\mathbb{E}(\hat{r}^2) \leq CD_m/n$. Theorem 5.3.1 shows that we can achieve this goal only if $D_m \geq C\sqrt{D}$.
Actually, we prove a slightly more general result since we obtain that the separation rate (as defined for example in Ingster [40, 41, 42]) for the test of null hypothesis $H_0 : s \in S_m$ against the alternative $H_1 : s \in \bar{S} - S_m$ is lower bounded by $C\sqrt{D}/n$. Thus, we extend the asymptotic result of Ingster since we prove it for any fixed $n \geq 10$.

Connection with adaptive estimation

There is a large amount of litterature on adaptive estimation (see for example [15] or [16] and the references therein). In our framework, an estimator $\tilde{s}_m$ of s is minimax on S_m if there exists a constant C such that

$$\sup_{s\in S_m} \mathbb{E}\left(\|\tilde{s}_m - s\|_2^2\right) \leq C \inf_{\tilde{s}} \sup_{s\in S_m} \mathbb{E}\left(\|\tilde{s} - s\|_2^2\right),$$

where the infimum is taken over all the possible estimators of s. The quantity $M_m = \inf_{\tilde{s}} \sup_{s\in S_m} \mathbb{E}\left(\|\tilde{s} - s\|_2^2\right)$ is called the minimax rate of convergence on S_m. In our framework, it is well known that this minimax rate of convergence is of order D_m/n and that the projection estimator is minimax over S_m. We proved in Sections 2.2 and 2.3 that D_m/n is also the minimax size for a confidence ball on S_m. An estimator $\hat{s}$ is said to be adaptive over a collection of models $(S_m)_{m\in\mathcal{M}_n}$ if there exists a constant C such that

$$\forall m \in \mathcal{M}_n, \; \sup_{s\in S_m} \mathbb{E}\left(\|\tilde{s} - s\|_2^2\right) \leq CM_m.$$

An adaptive estimator behaves as well as possible over all the models $(S_m)_{m\in\mathcal{M}_n}$ simultaneously. We can prove under reasonnable assumptions on the collection $(S_m)_{m\in\mathcal{M}_n}$ that the penalized least-squares estimator defined in Section 2.5 is adaptive over all the collection $(S_m)_{m\in\mathcal{M}_n}$. On the other hand, given an adaptive estimator $\tilde{s}$ over a collection $(S_m)_{m\in\mathcal{M}_n}$, one may wonder if we can learn from the data the actual size of $\|\tilde{s} - s\|_2^2$. More precisely, given $\alpha \in (0,1)$, the problem is to build an (eventually random) bound $\hat{r}^2$ such that, for all s in $(S_m)_{m\in\mathcal{M}_n}$ $\mathbb{P}(\|\tilde{s} - s\|_2^2 > \hat{r}^2) \leq 1-\alpha$ and such that, for all s in S_m, $\mathbb{E}(\hat{r}^2)$ is of order D_m/n. In other words, the problem is to build adaptive confidence balls centered in $\tilde{s}$. Theorem 5.3.1 can be interpreted in this context. In order to build $\hat{r}^2$, we have to choose a subset $\tilde{\mathcal{M}}_n$ of $\mathcal{M}_n$ such that the linear span S_n of $(S_m)_{m\in\tilde{\mathcal{M}}_n}$ has finite dimension D. Then, every bound $\hat{r}^2$ on $\|\tilde{s} - s\|_2^2$ satisfy $\mathbb{E}(\hat{r}^2) \geq C\sqrt{D}/n$. In particular, $\mathbb{E}(\hat{r}^2)$ does not give the actual order of $\mathbb{E}(\|\tilde{s} - s\|_2^2)$ when s is closed to a model S_m such that $D_m \leq C\sqrt{D}$. This is a fundamental difference with the problem of adaptive estimation where we do not need to specify a space S_n and where adaptation is possible without limitations on the dimensions of the models involved. This idea was extensively discussed in Hoffmann & Lepskii [38]. Nevertheless, adaptive confidence balls can be informative for applications, this is why we explain now how to obtain such adaptive confidence balls.

5.3.2 An adaptive algorithm

In this section, α and β denote some real numbers in $(0,1)$. Let S_n be a linear subspace of $L^2(\mu)$, with finite dimension D. Let $(S_m)_{m\in\mathcal{M}_n}$ be a collection of subspaces in S_n with respective dimensions $(D_m)_{m\in\mathcal{M}_n}$. Assume that Assumption $[M_1]$ is fulfilled in all the spaces $(S_m)_{m\in\mathcal{M}_n}$ and in S_n with the same constant Φ_0. As in the previous sections, we denote by $\bar{S}$ and $\bar{S}_m$ the set of all densities in S and S_m. Given i.i.d observations $X_1,...,X_n$, we want to build random variables $\tilde{s},\tilde{R}_\alpha$ measurable with respect to $\sigma(X_1,...,X_n)$ such that

$$\forall s\in\bar{S},\ \mathbb{P}_s\left(s\in B(\tilde{s},\tilde{R}_\alpha)\right)\geq 1-\alpha.$$

$$\forall m\in\mathcal{M}_n,\ \forall s\in\bar{S}_m,\ \mathbb{P}_s\left(\tilde{R}_\alpha^2\leq C_{\alpha,\beta}\left(\frac{D_m}{n}\vee\frac{\sqrt{D}}{n}\right)\right)\geq 1-\beta.$$

As in the previous sections, for all m in $\mathcal{M}_n$, we denote by s_m and $\hat{s}_m$ the orthogonal projection of s onto S_m and the projection estimator of s onto S_m. Moreover, let s_n be the orthogonal projection of s onto S_n. From Pythagoras Theorem, we have

$$\forall s\in\bar{L}^2(\mu),\ \forall m\in\mathcal{M}_n,\ \|\hat{s}_m-s\|_2^2=\|\hat{s}_m-s_m\|_2^2+\|s_m-s_n\|_2^2+\|s_n-s\|_2^2.$$

From Corollary 5.2.1, for all m in $\mathcal{M}_n$, we can build v_m^2 such that

$$\mathbb{P}\left(\|\hat{s}_m-s_m\|_2^2>v_m^2\right)\leq\alpha.$$

Moreover, there exists a constant $C_{\alpha,\beta}$, such that $\mathbb{P}(v_m^2>C_{\alpha,\beta}D_m/n)\leq\beta/2$. We do not deal with the bias of our collection of models and we assume that some bound $\|s_n-s\|_2^2\leq\eta^2$ is known. We will see in the applications in Section 3.3 that there exist some classical problems where such bound is available. Thus, in order to achieve our goal, it remains to control the quadratic functional $\|s_m-s_n\|_2^2$. This is the aim of the following subsection.

An estimation of $d(s_n,S_m)$ and a test

In order to build estimators of quadratic functionals such as $\|s_m-s_n\|_2^2$, Laurent & Massart [45] and Fromont & Laurent [33] used U-statistics of order 2. We use the same idea here. We deduce from Theorem 3.4 in Houdré & Reynaud-Bouret [39] a concentration inequality for U-statistics of order 2. We use this inequality to obtain a control of $\|s_m-s_n\|_2^2$. In order to prove an optimal result, we need another assumption on the density s.
[D]: $s\in L^\infty(M)=\{t\in L^2(\mu),\ \|t\|_\infty\leq M\}$.
We prove the following result.

Proposition 5.3.2 *Let $X_1,...,X_n$ be i.i.d random variables valued in a measurable subspace $(\mathbb{X},\mathcal{X})$ of $\mathbb{R}$, with common law P_s. Assume that s satisfies Assumption* **[D]** *for some constant M. Let S_n and S_m be linear subspaces in $L^2(\mu)$, with respective dimension D and D_m. Assume that S_n and all the S_m's satisfy Assumption $[M_1]$ with the same constant Φ_0. Let s_n and s_m be respectively the orthogonal projections* *of s onto S_n and S_m. Let $(\psi_\lambda)_{\lambda\in\Lambda}$ be an orthonormal basis of S_n such that $(\psi_\lambda)_{\lambda\in m}$ is a basis of S_m. Let*

$$U_m = \frac{1}{n(n-1)} \sum_{i\neq j=1}^{n} \sum_{\lambda\in\Lambda-m} (\psi_\lambda(X_i) - P_s\psi_\lambda)(\psi_\lambda(X_j) - P_s\psi_\lambda),$$

$$V_m = \frac{1}{n(n-1)} \sum_{i\neq j=1}^{n} \sum_{\lambda\in\Lambda-m} \psi_\lambda(X_i)\psi_\lambda(X_j). \tag{5.17}$$

Then, for all ξ in $\{-1,1\}$, for all $x>0$ and all $\epsilon>0$, we have

$$\mathbb{P}_s\left(\xi U_m > C_1\frac{\sqrt{Dx}}{n} + C_2(\epsilon)\frac{x}{n} + C_3(\epsilon)D\left(\frac{x}{n}\right)^2\right) \leq 2.8e^{-x}, \tag{5.18}$$

$$\mathbb{P}_s\left(\xi\left(V_m - \|s_n - s_m\|_2^2\right) > \frac{1}{2}\|s_n - s_m\|_2^2 + C(\epsilon,D,n,x)\right) \leq 3.8e^{-x}. \tag{5.19}$$

where $C_1 = 5.7\sqrt{M}\Phi_0$, $C_2(\epsilon) = (8+192\epsilon)M$, $C_3(\epsilon) = 192\Phi_0^2(5.4+\epsilon^{-1})$ and

$$C(\epsilon,D,n,x) = C_1(\epsilon)\frac{\sqrt{Dx}}{n} + (C_2(\epsilon)+8M)\frac{x}{n} + (C_3(\epsilon)+2\Phi_0^2)\frac{Dx^2}{n^2}. \tag{5.20}$$

Comments:

We deduce from Inequality (5.19) the control of $\|s_n - s_m\|_2^2$ that we need. It is consequence of Inequality (5.18) and of Bernstein's inequality. Inequality (5.18) can be used to build a test for the null hypothesis $H_0 : s \in S_m$ against the alternative $H_1 : s \in S - S_m$. Let us now briefly explain how.

We keep the notations and assumptions of Proposition 5.3.2. Under H_0, for all λ in $\Lambda - m$, we have $P_s(\psi_\lambda) = \int_{\mathbb{X}} s\psi_\lambda d\mu = 0$. Thus

$$V_m = \frac{1}{n(n-1)} \sum_{i\neq j=1}^{n} \sum_{\lambda\in\Lambda-m} \psi_\lambda(X_i)\psi_\lambda(X_j)$$

is a totally degenerate U-statistic of order 2 and from Inequality (5.18), we have, for all α in $(0,1)$,

$$\mathbb{P}_s\left(|V_m| > C_1\frac{\sqrt{D\ln(5.6/\alpha)}}{n} + C_2(\epsilon)\frac{\ln(5.6/\alpha)}{n} + C_3(\epsilon)\frac{D(\ln(5.6/\alpha))^2}{n^2}\right) \leq \alpha.$$

Thus a test for H_0 against H_1 with confidence level α is given by

$$\phi = 1_{|V_m|\leq C_1\sqrt{D\ln(5.6/\alpha)}/n + C_2(\epsilon)\ln(5.6/\alpha)/n + C_3(\epsilon)D(\ln(5.6/\alpha))^2/n^2}.$$

This test cannot have a well controlled error of second kind when $d(s,S_m)^2 \leq C_\alpha\sqrt{D}/n$ because of Inequality (5.19) but we show in the proof of Proposition 5.3.1 that this is the separation rate for this test. Thus, this test is rate-optimal. We can now turn to the problem of adaptive confidence balls.

Adaptive confidence balls

We apply the algorithm of Baraud [8] to obtain adaptive confidence balls in a density model. Let us recall this algorithm. We fix α in $(0,1)$ and we choose a collection $(\alpha_m)_{m\in\mathcal{M}_n}$ of positive real numbers such that $\sum_{m\in\mathcal{M}_n}\alpha_m=\alpha/4$. Assume moreover that we can choose the weights α_m such that $\ln(1/\alpha_m)\leq C_*(n/\sqrt{D})^{2/3}$ for some constant C_*. Let $\epsilon>0$ and let $(W_1,...,W_n)$ be a resampling scheme as defined in Section 2.2. Let $v_W^2=\mathrm{Var}(W_1-\sum_{i=1}^n W_i/n)$. For all m in $\mathcal{M}_n$, let $(\psi_\lambda^m)_{\lambda\in m}$ denotes an orthonormal basis that we complete to obtain an orthonormal basis $(\psi_\lambda^m)_{\lambda\in\Lambda}$ of S_n. When no confusion can occur, we simply write ψ_λ instead of ψ_λ^m. Let $y_m=\ln(4.8/\alpha_m)$. Let us define the following random variables:

$$v_m^2=\frac{(1+\epsilon)^3}{v_W^2}\mathbb{E}^W\left(\sum_{\lambda\in m}[\nu_n^W\psi_\lambda^m]^2\right)+\eta(\epsilon)\frac{\Phi_0\sqrt{MD_m}y_m}{n}+K_n(\epsilon)\frac{D_my_m^2}{n^2},$$

$$b_m^2=2V_m+2C(\epsilon,D,n,y_m),$$

where $\eta(\epsilon)$, $K_n(\epsilon)$ are defined in Theorem 5.2.1, $C(\epsilon,D,n,y_m)$ is defined in (5.20) and V_m is defined as in (5.17) with the basis $(\psi_\lambda^m)_{\lambda\in\Lambda-m}$. Remark that the assumption on α_m implies that there exists a constant C such that $y_m\leq Cn/\sqrt{D_m}$. Thus, there exists a constant $C(\epsilon)$ such that

$$v_m^2=\frac{(1+\epsilon)^3}{v_W^2}\mathbb{E}^W\left(\sum_{\lambda\in m}[\nu_n^W\psi_\lambda^m]^2\right)+C(\epsilon)\frac{\sqrt{D_m}y_m}{n}.$$

Moreover, there exists a constant C such that $Dy_m^2/n\leq C\sqrt{Dy_m}$. Thus, there exists a constant $C(\epsilon)$ such that

$$b_m^2=2V_m+C(\epsilon)\frac{\sqrt{Dy_m}+y_m}{n}. \tag{5.21}$$

Let $\eta>0$ and, for all m in $\mathcal{M}_n$, let $R_m^2=b_m^2+v_m^2+\eta^2$. Let $(\psi_\lambda)_{\lambda\in\Lambda}$ be an orthonormal basis of S_n, let $y_\alpha=\ln(9.6/\alpha)$ and let

$$R_n^2=\frac{(1+\epsilon)^3}{v_W^2}\mathbb{E}^W\left(\sum_{\lambda\in\Lambda}[\nu_n^W\psi_\lambda]^2\right)+\eta(\epsilon)\frac{\Phi_0\sqrt{MD}y_\alpha}{n}+K_n(\epsilon)\frac{Dy_\alpha^2}{n^2}+\eta^2.$$

Finally, for $\tilde{\mathcal{M}}_n=\mathcal{M}_n\cup\{n\}$, we choose

$$\hat{m}\in\mathrm{Arg}\min_{m\in\tilde{\mathcal{M}}_n}R_m^2,\ \tilde{R}^2=R_{\hat{m}}^2,\ \tilde{s}=\hat{s}_{\hat{m}}. \tag{5.22}$$

The following result holds.

Theorem 5.3.3 *Let $X_1,...,X_n$ be i.i.d random variables valued in a measurable subspace $(\mathbb{X},\mathcal{X})$ of $\mathbb{R}$, with common law P_s. Let $L^\infty(M)=\{t\in L^2(\mu),\ \|t\|_\infty\leq M\}$. Let S_n be a linear subspace of $L^2(\mu)$ with finite dimension D. Let $(S_m)_{m\in\mathcal{M}_n}$ be a collection of linear subspaces of S_n with respective dimensions $(D_m)_{m\in\mathcal{M}_n}$. Assume that S_n and all the S_m's satisfy Assumption $[M_1]$ with the same constant Φ_0. Let s_n and s_m be respectively the orthogonal projections of s onto S_n and S_m. Let*

$(\alpha_m)_{m\in\mathcal{M}_n}$ be a collection of positive real numbers such that $\sum_{m\in\mathcal{M}_n}\alpha_m = \alpha/4$ and such that $\ln(1/\alpha_m) \leq C_*(n/\sqrt{D})^{2/3}$ for some constant C_*. Let $\eta > 0$ and let $\tilde{s}, \tilde{R}$ be the random variables defined in (5.22). Then, for all s in $L^\infty(M)$ such that $d(s, S_n) \leq \eta$, we have

$$\mathbb{P}_s\left(s \in B(\tilde{s}, \tilde{R})\right) \geq 1-\alpha. \tag{5.23}$$

Let $(\eta_m)_{m\in\mathcal{M}_n}$ be a collection of positive real numbers and let β be a real number in $(0,1)$ such that $\ln(1/\beta) \leq C_*(n/\sqrt{D})^{2/3}$. There exists a constant C such that, for all m in $\mathcal{M}_n$, for all s in $L^\infty(M)$ satisfying $d(s_n, S_m) \leq \eta_m$ we have

$$\mathbb{P}_s\left(\tilde{R}^2 > \frac{C}{n}\left(D_m + \sqrt{D_m}x_m + \sqrt{Dx_m}\right) + \eta^2 + 3\eta_m^2\right) \leq \beta, \tag{5.24}$$

where $x_m = \ln(\alpha_m^{-1} \vee \beta^{-1})$.

Comments:
Baraud [8] proved the same kind of result in the regression framework and was the first, up to our knowledge, to study adaptive confidence balls from a nonasymptotic point of view. Theorem 5.3.3 can be viewed as an extension of his results to the density estimation framework.
The introduction of the weights $(\eta_m)_{m\in\mathcal{M}_n}$ prove the robustness of the procedure. Moreover, it allows us to replace the assumption $s \in \bar{S}$ by the more classical one that s belongs to some Besov space as we will see in Section 3.3.1.

5.3.3 Applications

Adaptive confidence balls under regularity assumptions

It is common to replace the assumption that s is close to some finite dimensional linear space by some regularity assumption on s. Let us recall some definitions and some results of approximation theory here.
Wavelet basis.
Hereafter, we work with an r-regular orthonormal multiresolution analysis of $L^2(\mu)$, associated with a compactly supported scaling function ϕ and a compactly supported mother wavelet ψ. Without loss of generality, we suppose that the support of the functions ϕ and ψ is an interval $[A_1, A_2]$ where A_1 and A_2 are two integers such that $A_2 - A_1 = A \geq 1$. Let us recall that ϕ and ψ generate an orthonormal basis by dilatations and translations.
For all k in $\mathbb{Z}$ and all j in $\mathbb{N}^*$, let $\psi_{0,k} : x \to \sqrt{2}\phi(2x-k)$ and $\psi_{j,k} : x \to 2^{j/2}\psi(2^jx-k)$. The family $\{(\psi_{j,k})_{j\geq 0,k\in\mathbb{Z}}\}$ is an orthonormal basis of $L^2(\mu)$. We assume moreover that the constant functions belong to the linear span of $(\psi_{0,k})_{k\in\mathbb{Z}}$.
Besov balls.
Let w, p be two positive numbers such that $w + 1/2 - 1/p > 0$. For all functions t in $L^2(\mu)$, $t = \sum_{j\geq 0,k\in\mathbb{Z}}\beta_{j,k}\psi_{j,k}$, we say that t belongs to the Besov ball $B_{w,p,\infty}(M_1)$ on the real line if $\|t\|_{w,p,\infty} \leq M_1$ where

$$\|t\|_{w,p,\infty} = \sup_{j\in\mathbb{N}} 2^{j(w+1/2-1/p)}\left(\sum_{k\in\mathbb{Z}}|\beta_{j,k}|^p\right)^{1/p}.$$

It is easy to check that if $p \geq 2$ $B_{w,p,\infty}(M_1) \subset B_{w,2,\infty}(M_1)$ so that upper bounds on $B_{w,2,\infty}(M_1)$ yield upper bounds on $B_{w,p,\infty}(M_1)$.

The framework.

Let $J \in \mathbb{N}$ and for all $J_m \leq J$, let

$$\Lambda_J = \{(0,k), -A_2 < k < 2 - A_1\} \cup \{(j,k),\ 1 \leq j \leq J,\ -A_2 < k < -A_1 + 2^j\}$$

$$m = \{(0,k), -A_2 < k < 2 - A_1\} \cup \{(j,k),\ 1 \leq j \leq J_m,\ -A_2 < k < -A_1 + 2^j\}.$$

Let S be the linear span of $(\psi_\lambda)_{\lambda\in\Lambda}$ and, for all $J_m = 1, ..., J$, let S_m be the linear span of $(\psi_\lambda)_{\lambda\in m}$. In particular, we have $D_m = A(J_m+1) + 2^{J_m+1} - J_m$ and thus $2^{J_m+1} \leq D_m \leq A(J_m+1)+2^{J_m+1} \leq (A+1)2^{J_m+1}$. Moreover, S_m satisfies Assumption $[M_1]$.

Approximation results on Besov spaces.

We have the following result (Birgé & Massart [15] Section 4.7.1). Suppose that the support of s equals $[0,1]$ and that s belongs to the Besov ball $B_{w,2,\infty}(M_1)$, then whenever $r > w - 1$,

$$\|s - s_m\|_2^2 \leq \frac{\|s\|_{w,2,\infty}^2}{4(4^w-1)} 2^{-2J_m w} \leq \frac{[2(1+A)]^{2w}\|s\|_{w,2,\infty}^2}{4(4^w-1)} D_m^{-2w}. \tag{5.25}$$

We can now state this straightforward application of Proposition 5.3.3.

Corollary 5.3.4 *Let S_n be the linear span $(\psi_\lambda)_{\lambda\in\Lambda_J}$ where $D = |\Lambda_J|$ satisfies*

$$\frac{1}{2}\left(\frac{n}{\sqrt{\log(\log n)}}\right)^{1/(2w+1/2)} \leq D \leq \left(\frac{n}{\sqrt{\log(\log n)}}\right)^{1/(2w+1/2)}.$$

We have $\sqrt{D} \leq n$. Let $(S_m)_{m\in\mathcal{M}_n}$ be the wavelet spaces defined above for all J_m in $1, ..., J$. Let $\alpha_m = \alpha/(4J) \leq C_\alpha/\ln(n)$. Let $\tilde{s}, \tilde{R}$ be the random variables given by (5.22) with $\eta^2 = M_1^2 2^{-2Jw}/(4(4^w-1))$. We have

$$\forall s \in B_{w,2,\infty}(M_1) \cap L^\infty(M),\ \mathbb{P}_s(s \in B(\tilde{s}, \tilde{R})) \geq 1 - \alpha.$$

Moreover, there exists a constant $C_{\alpha,\beta}$ such that, for all $v > w$, for all s in $B_{v,2,\infty}(M_1) \cap L^\infty(M)$, we have

$$\mathbb{P}_s\left(\tilde{R}^2 \leq C_{\alpha,\beta}\left[\left(\frac{1}{n}\right)^{\frac{2v}{2v+1}} \vee \left(\frac{\sqrt{\log(\log n)}}{n}\right)^{\frac{4w}{4w+1}}\right]\right) \geq 1 - \beta.$$

Comments:

This result extends the asymptotic result of Robins & Van der Vaart [61] since it is valid for any fixed n.

Adaptive confidence balls for a vector $((\mathbb{P}(X \in I_\lambda))_{\lambda\in\Lambda})$

In this section, we suppose that a partition $(I_\lambda)_{\lambda\in\Lambda}$ with finite cardinality D of $\mathbb{X}$ is given. We assume that there exists a positive constant Φ_0 such that, for all λ in Λ, $\mu(I_\lambda) > 1/(\Phi_0^2 D)$. This way, the space S_n of real valued functions constant on the partition $(I_\lambda)_{\lambda\in\Lambda}$ satisfies Assumption $[M_1]$. For all α in $(0,1)$, we can use Theorem 5.3.3 to build adaptive $(1-\alpha)$-confidence balls for the vector $((\mathbb{P}(X \in I_\lambda)/\sqrt{\mu(I_\lambda)})_{\lambda\in\Lambda})$. The first step is to see that the function

$$s_n = \sum_{\lambda\in\Lambda} \mathbb{P}(X \in I_\lambda) \frac{1_{I_\lambda}}{\sqrt{\mu(I_\lambda)}}$$

is the orthogonal projection of s onto the linear space S_n. Consider a family $(S_m)_{m\in\mathcal{M}_n}$ of linear subspaces of S. For all m in $\mathcal{M}_n$, S_m is a space of function constant on a finite partition $(I^m_\lambda)_{\lambda\in m}$. Thus all the spaces $(S_m)_{m\in\mathcal{M}_n}$ satisfy Assumption $[M_1]$ for the same constant Φ_0 if and only if for all m in $\mathcal{M}_n$ and all λ in m, we have $\mu(I^m_\lambda) > 1/(\Phi_0^2 D_m)$. Hereafter, we assume that this condition as fulfilled. We can apply the procedure of Section 3.2, with $\eta = 0$ to provide random variables $\tilde{s}$ and $\tilde{R}$ satisfying

$$\mathbb{P}_s\left(s_n \in B(\tilde{s}, \tilde{R})\right) \geq 1-\alpha$$

with $\mathbb{E}\tilde{R}^2 \leq C_{\alpha,\beta}\left(D_m + \sqrt{D_m}x_m + \sqrt{Dx_m} + \eta_m^2\right)/n$ when $d(s, s_m) \leq \eta_m$. The function $\tilde{s}$ is constant on the partition $(I_\lambda)_{\lambda\in\Lambda}$ like all the functions in S_n, thus we can consider the vector of its value on $(I_\lambda)_{\lambda\in\Lambda}$ and the ball (for the euclidian norm in $\mathbb{R}^D$) centered on this vector with radius $\tilde{R}$ is an adaptive $(1-\alpha)$-confidence ball for the vector $((\mathbb{P}(X \in I_\lambda)/\sqrt{\mu(I_\lambda)})_{\lambda\in\Lambda})$.

5.4 Proofs

We start this section with a very useful lemma. It is a concentration inequality for U-statistics based on an orthonormal basis of $L^2(\mu)$ derived from Theorem 3.4 in Houdré & Reynaud-Bouret [39].

Lemma 5.4.1 *Let $X, X_1, ..., X_n$ be i.i.d random variables, valued in a measurable subspace $(\mathbb{X}, \mathcal{X})$ of $\mathbb{R}$, with common density s with respect to a measure μ. Assume that s belongs to $L^2(\mu)$. Let S_m be a linear subspace of $L^2(\mu)$ and let $(\psi_\lambda)_{\lambda\in\Lambda_m}$ be an orthonormal basis of S_m. Let $B_m = \{t \in S_m,\ \|t\|_2 \leq 1\}$, $v_m^2 = \sup_{t\in B_m} Var(t(X))$ and $b_m = \sup_{t\in B_m}\|t - Pt\|_\infty$. Let*

$$U(\Lambda_m) = \frac{1}{n(n-1)} \sum_{i\neq j=1}^{n} \sum_{\lambda\in\Lambda_m} (\psi_\lambda(X_i) - P\psi_\lambda)(\psi_\lambda(X_j) - P\psi_\lambda).$$

Then for all $x > 0$ and all ξ in $\{-1, 1\}$, there exists a space Ω_x such that $\mathbb{P}(\Omega_x^c) \leq 2.8e^{-x}$ and such that, on Ω_x,

$$\xi U(\Lambda_m) \leq 5.7 v_m b_m \frac{\sqrt{x}}{n} + 8v_m^2 \frac{x}{n} + 384 v_m b_m \left(\frac{x}{n}\right)^{3/2} + 1020\left(\frac{b_m x}{n}\right)^2. \tag{5.26}$$

Proof :

We apply Theorem 3.4 in Houdré & Reynaud-Bouret [39]. We have, for all $x > 0$

$$\mathbb{P}\left(\xi U(\Lambda_m) > \frac{1}{n^2}\left(5.7B_1\sqrt{x} + 8B_2 x + 384 B_3 x^{3/2} + 1020 B_4 x^2\right)\right) \leq 2.8e^{-x}, \tag{5.27}$$

where

$$U(x,y) = \textstyle\sum_{\lambda\in\Lambda_m}(\psi_\lambda(x) - P\psi_\lambda)(\psi_\lambda(y) - P\psi_\lambda),$$
$$B_1^2 = n^2\mathbb{E}\left[(U(X_1,X_2))^2\right],\ B_3^2 = n\sup_x \mathbb{E}\left[(U(x,X_2))^2\right],\ B_4 = \sup_{x,y} U(x,y),$$
$$B_2 = \sup\left\{\left|\mathbb{E}\sum_{i=1}^{n}\sum_{j=1}^{i-1}U(X_1,X_2)\alpha_i(X_1)\beta_j(X_2)\right|,\ \mathbb{E}\sum_{i=1}^{n}\alpha_i^2(X_1)\le 1,\ \mathbb{E}\sum_{j=1}^{n}\beta_j^2(X_1)\le 1\right\}.$$

Let $T_m = \sum_{\lambda\in\Lambda_m}(\psi_\lambda - P\psi_\lambda)^2$ and let $b_m = \sup_{t\in B_m}\|t - Pt\|_\infty$. From Cauchy-Schwarz inequality, we have, for all real numbers $(b_\lambda)_{\lambda\in\Lambda_m}$

$$\sum_{\lambda\in\Lambda_m} b_\lambda^2 = \left(\sup_{\sum a_\lambda^2\le 1}\sum_{\lambda\in\Lambda_m} a_\lambda b_\lambda\right)^2. \tag{5.28}$$

In particular, since the system $(\psi_\lambda)_{\lambda\in\Lambda_m}$ is orthonormal, we have, for all x in $\mathbb{X}$, $T_m(x) = (\sup_{t\in B_m}(t(x) - Pt))^2$. Thus

$$PT_m \le \|T_m\|_\infty \le b_m^2. \tag{5.29}$$

Let us now evaluate B_1, B_2, B_3 and B_4.

Evaluation of B_1:

$$\begin{aligned}\frac{B_1^2}{n^2} &= \sum_{\lambda,\lambda'\in\Lambda_m}\left(P\left((\psi_\lambda - P\psi_\lambda)(\psi_{\lambda'} - P\psi_{\lambda'})\right)\right)^2\\ &= \sum_{\lambda\in\Lambda_m}\left(\sup_{\sum a_{\lambda'}^2\le 1} P\left((\psi_\lambda - P\psi_\lambda)\left[\sum_{\lambda'\in\Lambda_m} a_{\lambda'}\psi_{\lambda'} - P\left(\sum_{\lambda'\in\Lambda_m} a_{\lambda'}\psi_{\lambda'}\right)\right]\right)\right)^2\\ &= \sum_{\lambda\in\Lambda_m}\left(\sup_{t\in B_m} P\left((\psi_\lambda - P\psi_\lambda)(t - Pt)\right)\right)^2 \le PT_m v_m^2,\end{aligned}$$

where we use successively the independence of X_1 and X_2, Inequality (5.28), the orthonormality of the system $(\psi_\lambda)_{\lambda\in\Lambda_m}$ and Cauchy-Schwarz inequality. Thus we obtain

$$B_1 \le n v_m b_m. \tag{5.30}$$

Evaluation of B_2: For all real numbers y, z, we have $2yz \le y^2 + z^2$, thus, for all i, j in $\{1, ..., n\}$, we have

$$2P\left((\psi_\lambda - P\psi_\lambda)\alpha_i\right)P\left((\psi_{\lambda'} - P\psi_{\lambda'})\beta_j\right) \le \left(P\left((\psi_\lambda - P\psi_\lambda)\alpha_i\right)\right)^2 + \left(P\left((\psi_{\lambda'} - P\psi_{\lambda'})\beta_j\right)\right)^2.$$

We apply (5.28) with $b_\lambda = P\left((\psi_\lambda - P\psi_\lambda)\alpha_i\right)$, since the system $(\psi_\lambda)_{\lambda\in\Lambda_m}$ is orthonormal, we have for all i in $\{1, ..., n\}$,

$$\sum_{\lambda\in\Lambda_m}\left(P\left((\psi_\lambda - P\psi_\lambda)\alpha_i\right)\right)^2 = \left(\sup_{t\in B_m} P(t - Pt)\alpha_i\right)^2 \le v_m^2 P\alpha_i^2.$$

Since $\sum_{i=1}^n P\alpha_i^2 \le 1$ we deduce that

$$\sum_{i,j=1}^{n}\sum_{\lambda\in\Lambda_m}\left(P\left((\psi_\lambda - P\psi_\lambda)\alpha_i\right)\right)^2 \le n v_m^2.$$

The same inequality holds for β_j, thus we obtain

$$B_2 \leq n v_m^2. \tag{5.31}$$

Evaluation of B_3: For all x in $\mathbb{X}$, $\mathbb{E}[(U(x, X_2))^2]$ is the variance of the function $t_x = \sum_{\lambda \in m}(\psi_\lambda(x) - P\psi_\lambda)\psi_\lambda$. t_x is a function in S_m and, from inequality (5.28), we have

$$\|t_x\|_2^2 = \sum_{\lambda \in \Lambda_m} (\psi_\lambda(x) - P\psi_\lambda)^2 = \left(\sup_{t \in B_m} (t(x) - Pt)\right)^2 \leq b_m^2.$$

Thus $\mathbb{E}[(U(x, X_2))^2] = \mathrm{Var}(t_x(X)) = b_m^2 \mathrm{Var}(t_x(X)/b_m) \leq b_m^2 v_m^2$. Thus

$$B_3 \leq \sqrt{n} v_m b_m. \tag{5.32}$$

Evaluation of B_4: We apply Cauchy-Schwarz inequality and we obtain

$$B_4 \leq \|T_m\|_\infty \leq b_m^2. \tag{5.33}$$

Let Ω_x^c be the event defined by inequality (5.27). From (5.30), (5.31), (5.32) and (5.33), we have, on Ω_x,

$$\xi U(\Lambda_m) \leq \frac{5.7 v_m b_m \sqrt{x}}{n} + \frac{8 v_m^2 x}{n} + 384 v_m b_m \left(\frac{x}{n}\right)^{3/2} + 1020 b_m^2 \left(\frac{x}{n}\right)^2. \tag{5.34}$$

5.4.1 Proof of Theorem 5.2.1:

Inequalities (5.4) and (5.5) are trivial when $x \leq 1$ so that we only have to prove them for all $x \geq 1$. We decompose the proof into several lemmas.

Lemma 5.4.2 *Let*

$$U_m = \frac{1}{n(n-1)} \sum_{\lambda \in m} \sum_{l \neq l'=1}^{n} (\psi_\lambda(X_l) - P\psi_\lambda)(\psi_\lambda(X_{l'}) - P\psi_\lambda).$$

Then we have

$$\|s_m - \hat{s}_m\|_2^2 = R_W^2 + U_m. \tag{5.35}$$

In particular,

$$\mathbb{E}\left(\|s_m - \hat{s}_m\|_2^2\right) = \mathbb{E}\left(R_W^2\right). \tag{5.36}$$

Proof of Lemma 5.4.2:

For all $i = 1, ..., n$, let $\tilde{W}_i = W_i - \bar{W}_n$. For all functions t, we have $\nu_n^W(t) = \sum_{i=1}^n \tilde{W}_i t(X_i)/n$. $\sum_{i=1}^n \tilde{W}_i = 0$, thus, for all λ in m we have $\nu_n^W(P\psi_\lambda) = 0$. Moreover,

$$0 = \mathbb{E}\left[\left(\sum_{i=1}^n \tilde{W}_i\right)^2\right] = \sum_{i=1}^n \mathbb{E}\left(\tilde{W}_i^2\right) + \sum_{i \neq j=1}^n \mathbb{E}\left(\tilde{W}_i \tilde{W}_j\right) = n\mathbb{E}(\tilde{W}_1^2) + n(n-1)\mathbb{E}(\tilde{W}_1 \tilde{W}_2).$$

The second equality comes from the exchangeability of the weights. Thus, we have

$$v_W^2 = \mathbb{E}(\tilde{W}_1^2) = -(n-1)\sum_{i \neq j} \mathbb{E}(\tilde{W}_1 \tilde{W}_2).$$

Hence,

$$\begin{aligned}
v_W^2 R_W^2 &= \sum_{\lambda\in m} \mathbb{E}^W\left([\nu_n^W(\psi_\lambda)]^2\right) = \sum_{\lambda\in m} \mathbb{E}^W\left([\nu_n^W(\psi_\lambda - P\psi_\lambda)]^2\right)\\
&= \sum_{\lambda\in m} \mathbb{E}^W\left(\frac{1}{n^2}\sum_{i,j=1}^n \tilde{W}_i\tilde{W}_j(\psi_\lambda(X_i)-P\psi_\lambda)(\psi_\lambda(X_j)-P\psi_\lambda)\right)\\
&= \frac{1}{n^2}\sum_{\lambda\in m}\sum_{i=1}^n \mathbb{E}\left(\tilde{W}_i^2\right)(\psi_\lambda(X_i)-P\psi_\lambda)^2\\
&\quad +\frac{1}{n^2}\sum_{\lambda\in m}\sum_{i\neq j=1}^n \mathbb{E}(\tilde{W}_i\tilde{W}_j)(\psi_\lambda(X_i)-P\psi_\lambda)(\psi_\lambda(X_j)-P\psi_\lambda)\\
&= \frac{v_W^2}{n}\left(P_n\left(\sum_{\lambda\in m}(\psi_\lambda - P\psi_\lambda)^2\right) - U_m\right). \qquad (5.37)
\end{aligned}$$

On the other hand, easy algebra leads to

$$\|s_m - \hat{s}_m\|_2^2 = \sum_{\lambda\in m}\left([\nu_n(\psi_\lambda)]^2\right) = \frac{1}{n}\left(P_n\left(\sum_{\lambda\in m}(\psi_\lambda - P\psi_\lambda)^2\right) + (n-1)U_m\right).$$

Thus, we have $\|s_m - \hat{s}_m\|_2^2 - R_W^2 = U_m$. This proves (5.35). In order to prove (5.36), just remark that U_m is a totally degenerate U-statistic of order 2.

The next lemma is a straightforward application of Bousquet's version of Talagrand's Theorem (see [18]).

Lemma 5.4.3 *Let $\epsilon > 0$, $\alpha(\epsilon) = 4(1+\epsilon^{-1})$, $\beta(\epsilon) = 2(1+\epsilon^{-1})(\epsilon^{-1}+1/3)^2$. Then, for all $x > 0$, with probability larger than $1 - e^{-x}$,*

$$\|s_m - \hat{s}_m\|_2^2 \le (1+\epsilon)^3\mathbb{E}\left(\|s_m - \hat{s}_m\|_2^2\right) + \alpha(\epsilon)\frac{\Phi_0\sqrt{D_m}\,\|s\|_2\,x}{n} + \beta(\epsilon)\frac{\Phi_0^2 D_m x^2}{n^2}. \quad (5.38)$$

and, with probability larger than $1 - e^{-x}$,

$$\|s_m - \hat{s}_m\|_2^2 \le (1+\epsilon)^3\mathbb{E}\left(\|s_m - \hat{s}_m\|_2^2\right) + \alpha(\epsilon)\frac{\Phi_0^2 D_m x}{n} + \beta(\epsilon)\frac{\Phi_0^2 D_m x^2}{n^2}. \quad (5.39)$$

Proof of Lemma 5.4.3:

We apply Inequality (5.28) with $b_\lambda = \nu_n(\psi_\lambda)$, and we obtain

$$\|s_m - \hat{s}_m\|_2 = \sqrt{\sum_{\lambda\in m}[\nu_n(\psi_\lambda)]^2} = \sup_{\sum a_\lambda^2\le 1}\sum_{\lambda\in m} a_\lambda[\nu_n(\psi_\lambda)] = \sup_{t\in S_m,\ \|t\|_2\le 1}\nu_n(t).$$

Let $\epsilon > 0$, $\kappa(\epsilon) = \epsilon^{-1} + 1/3$, $B_m = \{t \in S_m,\ \|t\|_2 \le 1\}$, $v_m^2 = \sup_{t\in B_m}\mathrm{Var}(t(X))$ and $b_m = \sup_{t\in B_m}\|t\|_\infty$. Talagrand's Theorem (see for example Bousquet's version in Bousquet [18]) states that

$$\forall x > 0,\ \mathbb{P}\left(\|s_m - \hat{s}_m\|_2 > (1+\epsilon)\mathbb{E}\left(\|s_m - \hat{s}_m\|_2\right) + v_m\sqrt{\frac{2x}{n}} + \kappa(\epsilon)\frac{b_m x}{n}\right) \le e^{-x}.$$

For all real numbers y, z, we have $(y+z)^2 \leq (1+\epsilon)y^2 + (1+\epsilon^{-1})z^2$. Thus for all $x > 0$, we obtain that, with probability less than e^{-x},

$$\|s_m - \hat{s}_m\|_2^2 > (1+\epsilon)^3 \left(\mathbb{E}\left(\|s_m - \hat{s}_m\|_2\right)\right)^2 + \left(1+\frac{1}{\epsilon}\right)\left(v_m\sqrt{\frac{2x}{n}} + \kappa(\epsilon)\frac{b_m x}{n}\right)^2. \tag{5.40}$$

From Jensen inequality, we have $\left(\mathbb{E}\left(\|s_m - \hat{s}_m\|_2\right)\right)^2 \leq \mathbb{E}\left(\|s_m - \hat{s}_m\|_2^2\right)$, thus using again the inequality $(y+z)^2 \leq 2y^2 + 2z^2$, we obtain

$$\forall x > 0,\ \mathbb{P}\left(\|s_m - \hat{s}_m\|_2^2 > (1+\epsilon)^3 \mathbb{E}\left(\|s_m - \hat{s}_m\|_2^2\right) + \alpha(\epsilon)\frac{v_m^2 x}{n} + \beta(\epsilon)\frac{b_m^2 x^2}{n^2}\right) \leq e^{-x}. \tag{5.41}$$

From Assumption $[M_1]$, we have $b_m \leq \Phi_0\sqrt{D_m}$ and by Cauchy-Schwarz inequality and Assumption $[M_1]$, we have, for all t in S_m such that $\|t\|_2 \leq 1$

$$\mathrm{Var}(t(X_1)) \leq \mathbb{E}t^2(X_1) \leq \|t\|_\infty P|t| \leq \Phi_0\sqrt{D_m}\,\|t\|_2\,\|s\|_2 \leq \Phi_0\sqrt{D_m}\,\|s\|_2\,.$$

Thus $v_m^2 \leq \Phi_0\sqrt{D_m}\,\|s\|_2$. Plugging the evaluation of b_m and v_m^2 in (5.41), we obtain (5.38). Using the inequality $\mathrm{Var}(t(X_1)) \leq \mathbb{E}t^2(X_1) \leq \|t\|_\infty^2 \leq \Phi_0^2 D_m$, we obtain (5.39).□

The next lemma is a consequence of (5.26) and of Bernstein's inequality.

Lemma 5.4.4 *We keep the notations of Lemmas 5.4.2 and 5.4.3. For all $x \geq 1$, and all ξ in $\{-1, 1\}$, there exists an event Ω_x with $\mathbb{P}(\Omega_x^c) \leq 3.8e^{-x}$ such that, on Ω_x, we have*

$$\xi(R_W^2 - \mathbb{E}R_W^2) \leq \frac{1}{n} + \frac{\Phi_0^2 D_m}{n}\left(5.7\frac{\sqrt{x}}{n} + 9\frac{x}{n} + 384\left(\frac{x}{n}\right)^{3/2} + 1020\left(\frac{x}{n}\right)^2\right). \tag{5.42}$$

Proof :

From (5.37), we have $\mathbb{E}R_W^2 = P\left(\sum_{\lambda\in m}(\psi_\lambda - P\psi_\lambda)^2\right)$, thus

$$R_W^2 - \mathbb{E}R_W^2 = \frac{1}{n}\left(\nu_n\left(\sum_{\lambda\in m}(\psi_\lambda - P\psi_\lambda)^2\right) - U_m\right). \tag{5.43}$$

Let $T_m = \sum_{\lambda\in m}(\psi_\lambda - P\psi_\lambda)^2$. Since the constant functions are assumed to belong to S_m, for all function t in S_m, $t - Pt$ belongs to S_m. Thus, from assumption $[M_1]$, $b_m \leq \Phi_0^2 D_m$. We apply inequality (5.29) with $\Lambda_m = m$ and we obtain

$$n\mathbb{E}(R_W^2) = PT_m \leq \|T_m\|_\infty \leq \Phi_0^2 D_m. \tag{5.44}$$

Thus, from Bernstein's inequality, we have

$$\forall x > 0,\ \mathbb{P}\left(\xi\nu_n(T_m) > \sqrt{\frac{2\Phi_0^2 D_m x}{n}} + \frac{\Phi_0^2 D_m x}{3n}\right) \leq e^{-x}.$$

We apply the inequality $yz \leq y^2/4 + z^2$ to $y = \sqrt{2\Phi_0^2 D_m x/n}$ and $z = 1$ and we obtain

$$\forall x > 0,\ \mathbb{P}\left(\xi\nu_n(T_m) > 1 + \frac{\Phi_0^2 D_m x}{n}\right) \leq e^{-x}. \tag{5.45}$$

In order to control the U-statistic, we use inequality (5.26) with $\Lambda_m = m$ and $v_m^2 = \sup_{t\in B_m} \mathrm{Var}(t(X))$. For all $x > 0$, there exists an event Ω_x with $\mathbb{P}(\Omega_x^c) \leq 2.8e^{-x}$ such that, on Ω_x

$$\xi U_m \leq 5.7 v_m \frac{\Phi_0\sqrt{D_m x}}{n} + 8v_m^2\frac{x}{n} + 384\sqrt{v_m}\left(\frac{\Phi_0\sqrt{D_m}x}{n}\right)^{3/2} + 1020\left(\frac{\Phi_0\sqrt{D_m}x}{n}\right)^2. \tag{5.46}$$

For all t in B_m, from Assumption $[M_1]$, we have $\|t\|_\infty \leq \Phi_0\sqrt{D_m}$, thus $\mathrm{Var}(t(X)) \leq \|t\|_\infty^2 \leq \Phi_0^2 D_m$. In particular $v_m^2 \leq \Phi_0^2 D_m$. We plug (5.46) and (5.45) in (5.43) to conclude the proof of Lemma 5.4.4.

Conclusion of the proof of Theorem 5.2.1.

Let $x \geq 1$, $\epsilon > 0$. From Inequality (5.38) of Lemma 5.4.3, we have, on an event Ω_2 such that $\mathbb{P}(\Omega_2^c) \leq e^{-x}$,

$$\|s_m - \hat{s}_m\|_2^2 \leq (1+\epsilon)^3\mathbb{E}\left(\|s_m - \hat{s}_m\|_2^2\right) + \alpha(\epsilon)\frac{\Phi_0\sqrt{D_m}\,\|s\|_2\, x}{n} + \beta(\epsilon)\frac{\Phi_0^2 D_m x^2}{n^2}.$$

From Lemma 5.4.2, we have $\mathbb{E}\left(\|s_m - \hat{s}_m\|_2^2\right) = \mathbb{E}\left(R_W^2\right)$, thus, on Ω_2

$$\|s_m - \hat{s}_m\|_2^2 \leq (1+\epsilon)^3\mathbb{E}\left(R_W^2\right) + \alpha(\epsilon)\frac{\Phi_0\sqrt{D_m}\,\|s\|_2\, x}{n} + \beta(\epsilon)\frac{\Phi_0^2 D_m x^2}{n^2}.$$

From Lemma 5.4.4, there exists an event Ω_3 with $\mathbb{P}(\Omega_3^c) \leq 3.8e^{-x}$ such that, on Ω_3

$$\begin{aligned}\mathbb{E}\left(R_W^2\right) &\leq R_W^2 + \frac{1}{n} + \frac{\Phi_0^2 D_m}{n}\left(5.7\frac{\sqrt{x}}{n} + 9\frac{x}{n} + 384\left(\frac{x}{n}\right)^{3/2} + 1020\left(\frac{x}{n}\right)^2\right)\\ &\leq R_W^2 + \frac{1}{n} + \frac{\Phi_0^2 D_m}{n^2}\left(14.7 + \frac{384}{\sqrt{n}} + \frac{1020}{n}\right)x^2.\end{aligned}$$

In the last inequality, we use the inequality $x \geq 1$. Therefore, on $\Omega_2 \cap \Omega_3$, we have

$$\|s_m - \hat{s}_m\|_2^2 \leq (1+\epsilon)^3(R_W^2 + \frac{1}{n}) + \alpha(\epsilon)\frac{\Phi_0\sqrt{D_m}\,\|s\|_2\, x}{n} + K_n(\epsilon)\frac{D_m x^2}{n^2}.$$

This concludes the proof of (5.4). In order to prove (5.5), we use inequality (5.39) of Lemma 5.4.3 rather than (5.38).

5.4.2 Proof of Corollary 5.2.2:

The covering properties are straightforward from Theorem 5.2.1. When $\ln(1/\alpha) \leq C_* n$, we have $D_m(\ln(1/\alpha))^2/n^2 \leq C_* D_m \ln(1/\alpha)/n$ and when $\ln(1/\alpha) \leq C_* n/\sqrt{D_m}$, we have

$$\frac{D_m(\ln(1/\alpha))^2}{n^2} \leq C_*\frac{\sqrt{D_m}\ln(1/\alpha)}{n}.$$

Moreover, from Inequality (5.44), we have $\mathbb{E}(R_W^2) \leq \Phi_0^2 D_m/n$. Thus the bound in expectation is proved. From inequalities (5.42) and (5.44), there exists a constant C such that

$$\mathbb{P}\left(R_W^2 > C\frac{D_m}{n}\left(1 + \frac{(\ln(1/\beta))^2}{n}\right)\right) \leq \beta.$$

When $\ln(1/\beta) \leq C_* n$, we have $(\ln(1/\beta))^2/n \leq C_* \ln(1/\beta)$ and when $\ln(1/\beta) \leq C_* n/\sqrt{D_m}$, we have $D_m(\ln(1/\beta))^2/n \leq C_*\sqrt{D_m}\ln(1/\beta)/n$. This concludes the proof of the upper bound in probability.

5.4.3 Proof of Proposition 5.2.3:

The proof is decomposed in two lemmas.

Lemma 5.4.5 *Let us assume (5.9-5.10). For all s in S_m, we have*

$$\mathbb{P}_s\left(\|s-\hat{s}\|_2^2 > d_m^2\right) \leq \alpha+\beta. \tag{5.47}$$

Proof of Lemma 5.4.5:

$$\begin{aligned}\mathbb{P}_s\left[\|s-\hat{s}\|_2 > d_m\right] &= \mathbb{P}_s\left[\|s-\hat{s}\|_2 > d_m \cap d_m \geq \hat{r}\right] + \mathbb{P}_s\left[\|s-\hat{s}\|_2 > d_m \cap d_m < \hat{r}\right]\\ &\leq \mathbb{P}_s\left[\|s-\hat{s}\|_2 > \hat{r}\right] + \mathbb{P}_s\left[d_m < \hat{r}\right] \leq \alpha+\beta.\end{aligned}$$

Lemma 5.4.6 *Let $\delta = \alpha+\beta$ and let $d_m(\delta)$ be any real number satisfying (5.47). Then we have*

$$d_m^2(\delta) \geq \frac{D_m-1}{2n} - \frac{1}{n}\sqrt{2(D_m+1)\ln\left[\frac{\sqrt{1+(D_m+1)n^{-1}}}{1-\delta}\right]}.$$

Remark: When $D_m \geq 3+18\log(\sqrt{2}/(1-\delta))$ and $n \geq D_m+1$, we have

$$\sqrt{2(D_m+1)\ln\left[\frac{\sqrt{1+(D_m+1)n^{-1}}}{1-\delta}\right]} \leq \frac{D_m-1}{3},$$

thus $d_m^2(\delta) \geq (D_m-1)/(6n)$.

Proof :

We prove that if

$$d_m^2(\delta) = \frac{D_m-1}{2n} - \frac{1}{n}\sqrt{2(D_m+1)\ln\left[\frac{\sqrt{1+(D_m+1)n^{-1}}}{1-\delta}\right]}$$

then

$$\inf_{s\in S_m} \mathbb{P}_s\left[\|s-\hat{s}\|_2 \leq d_m(\delta)\right] \leq 1-\delta.$$

Let $s_0 = 1_{[0,1)}$, $m = \{1,...,[D_m/2]\}$ and for all λ in m, let

$$\psi_\lambda = \sqrt{\frac{D_m}{2}}\left(1_{[2(\lambda-1)/D_m,(2\lambda-1)/D_m)} - 1_{[(2\lambda-1)/D_m,2\lambda/D_m)}\right).$$

It is easy to check that $(\psi_\lambda)_{\lambda\in m}$ is an orthonormal system in S_m, orthogonal to s_0 such that, for all λ in m, $\|\psi_\lambda\|_\infty \leq \sqrt{D_m/2}$. Let $\hat{s}_0 = \int \hat{s}s_0 d\mu$ and for all λ in m, let $\hat{\beta}_\lambda = \int \hat{s}\psi_\lambda d\mu$. Let $(\xi_\lambda)_{\lambda\in m}$ be independent Rademacher random variables, independent of $X_1,...,X_n$, let ρ be some real number to be chosen later and let

$s_\xi = s_0 + \rho \sum_{\lambda \in m} \xi_\lambda \psi_\lambda$. The ψ_λ have distinct support, thus $\left\| \sum_{\lambda \in m} |\psi_\lambda| \right\|_\infty \leq \sqrt{D_m/2}$ and s_ξ is a density if $-\sqrt{2/D_m} \leq \rho \leq \sqrt{2/D_m}$. We have

$$\begin{aligned} \inf_{s \in S_m} \mathbb{P}_s \left[\|s - \hat{s}\|_2 \leq d_m(\delta) \right] &\leq \mathbb{P}_{s_\xi} \left[\|s_\xi - \hat{s}\|_2^2 \leq d_m^2(\delta) \right] \\ &= \mathbb{E}_{s_\xi} \left(\mathbf{1}_{(1-\hat{s}_0)^2 + \sum_{\lambda \in m} \left(\rho \xi_\lambda - \hat{\beta}_\lambda \right)^2 \leq d_m^2(\delta)} \right) \\ &= \int_0^1 \left(\mathbf{1}_{(1-\hat{s}_0)^2 + \sum_{\lambda \in m} \left(\rho \xi_\lambda - \hat{\beta}_\lambda \right)^2 \leq d_m^2(\delta)} \right) s_\xi d\mu. (5.48) \end{aligned}$$

We have

$$\sum_{\lambda \in m} \left(\rho \xi_\lambda - \hat{\beta}_\lambda \right)^2 = \sum_{\lambda \in m} \rho^2 - 2\rho \xi_\lambda \hat{\beta}_\lambda + \hat{\beta}_\lambda^2 \geq \sum_{\lambda \in m,\ \rho \xi_\lambda \hat{\beta}_\lambda \leq 0} \rho^2 - 2\rho \xi_\lambda \hat{\beta}_\lambda + \hat{\beta}_\lambda^2 \geq \rho^2 N(\xi, \hat{s}). \tag{5.49}$$

where $N(\xi, \hat{s}) = \mathrm{Card}(\{\lambda \in m,\ \rho \xi_\lambda \hat{\beta}_\lambda \leq 0\}) = \sum_{\lambda \in m} 1_{\{\rho \xi_\lambda \hat{\beta}_\lambda \leq 0\}}$. If we plug (5.49) in (5.48), we obtain

$$\inf_{s \in S_m} \mathbb{P}_s \left[\|s - \hat{s}\|_2 \leq d_m(\delta) \right] \leq \int_0^1 \mathbf{1}_{\rho^2 N(\xi, \hat{s}) \leq d_m^2(\delta)} s_\xi d\mu.$$

We integrate with respect to ξ and we apply Fubbini's theorem, we obtain

$$\inf_{s \in S_m} \mathbb{P}_s \left[\|s - \hat{s}\|_2 \leq d_m(\delta) \right] \leq \int_0^1 \mathbb{E}_\xi \left(\mathbf{1}_{\rho^2 N(\xi, \hat{s}) \leq d_m^2(\delta)} s_\xi \right) d\mu. \tag{5.50}$$

From Cauchy-Schwarz inequality, we have

$$\mathbb{E}_\xi^2 \left(\mathbf{1}_{\rho^2 N(\xi, \hat{s}) \leq d_m^2(\delta)} s_\xi \right) \leq \mathbb{P}_\xi \left(\rho^2 N(\xi, \hat{s}) \leq d_m^2(\delta) \right) \mathbb{E}_\xi \left(s_\xi^2 \right). \tag{5.51}$$

We have $\mathbb{E}_\xi s_\xi^2 = s_0^2 + \rho^2 \sum_{\lambda \in m} \psi_\lambda^2$. For all λ in m, $\int_0^1 \psi_\lambda^2 = 1$, thus

$$\int_0^1 \mathbb{E}_\xi s_\xi^2 d\mu = 1 + \rho^2 [D_m/2]. \tag{5.52}$$

Moreover, conditionally to $\hat{s}$, $N(\xi, \hat{s})$ is a sum of $[D_m/2]$ independent random variables valued in $\{0, 1\}$. Thus, from Hoeffding's inequality, we have

$$\forall t > 0,\ \mathbb{P}_\xi \left(N(\xi, \hat{s}) \leq \mathbb{E}_\xi \left(N(\xi, \hat{s}) \right) - \sqrt{[D_m/2] t} \right) \leq e^{-2t}. \tag{5.53}$$

We have

$$E_\xi \left(N(\xi, \hat{s}) \right) = \sum_{\lambda \in m} \mathbb{E}_\xi \left(\mathbf{1}_{\xi_\lambda \hat{\beta}_\lambda \leq 0} \right) \geq \frac{[D_m/2]}{2}.$$

We choose

$$t = \ln \left[\frac{\sqrt{1 + \rho^2 [D_m/2]}}{1 - \delta} \right],\ \rho = \sqrt{\frac{2}{n}} \leq \sqrt{\frac{2}{D_m}}.$$

Since $(D_m - 1)/2 \leq [D_m/2] \leq (D_m + 1)/2$, we have

$$t \leq \ln \left[\frac{\sqrt{1 + (D_m + 1)/n}}{1 - \delta} \right],\ E_\xi \left(N(\xi, \hat{s}) \right) \geq \frac{D_m - 1}{4}.$$

Thus

$$\{\rho^2 N(\xi,\hat{s}) \le d_m^2(\delta)\} \subset \{N(\xi,\hat{s}) \le \mathbb{E}_\xi\left(N(\xi,\hat{s})\right) - \sqrt{[D_m/2]t}\}.$$

Hence, from (5.53), we have

$$\mathbb{P}_\xi\left(\rho^2 N(\xi,\hat{s}) \le d_m^2(\delta)\right) \le \frac{(1-\delta)^2}{1+\rho^2[D_m/2]}. \tag{5.54}$$

We plug inequalities (5.52) and (5.54) in (5.51) to obtain

$$\int_0^1 \mathbb{E}_\xi^2\left(\mathbf{1}_{D_m\rho^2 N(\xi,\hat{s})\le d_m^2(\delta)} s_\xi\right) \le (1-\delta)^2.$$

Thus, from (5.50) and Jensen inequality, we have

$$\inf_{s\in S_m} \mathbb{P}_s\left[\|s-\hat{s}\|_2 \le d_m(\delta)\right] \le 1-\delta.$$

5.4.4 Proof of Theorem 5.3.1:

We decompose the proof into two lemmas.

Lemma 5.4.7 *Let $S(2r_m) = \{t \in S - S_m\ ;\ \|t-s_0\|_2 \ge 2r_m\}$. Under the assumptions of Theorem 5.3.1, there exists a test ϕ such that*

$$\mathbb{P}_{s_0}(\phi) \ge 1-\alpha, \ \inf_{s\in S(2r_m)} \mathbb{P}_s(1-\phi) \ge 1-(\alpha+\beta).$$

Proof of Lemma 5.4.7:
Let $\phi = 1_{s_0\in B(\hat{s},\hat{r})}$. From Inequality (5.15), we have $\mathbb{P}_{s_0}(\phi) \ge 1-\alpha$. Moreover, for all s in $S(2r_m)$, we have

$$\begin{aligned}
\mathbb{P}_s(\phi) &= \mathbb{P}_s(s_0 \in B(\hat{s},\hat{r})) = \mathbb{P}_s(\|s_0-\hat{s}\|_2 \le \hat{r}) \\
&\le \mathbb{P}_s(\|s_0-s\|_2 - \|s-\hat{s}\|_2 \le \hat{r}) \le \mathbb{P}_s(\|s-\hat{s}\|_2 \ge 2r_m - \hat{r}) \\
&= \mathbb{P}_s(\|s-\hat{s}\|_2 \ge 2r_m - \hat{r} \cap \hat{r} > r_m) + \mathbb{P}_s(\|s-\hat{s}\|_2 \ge 2r_m - \hat{r} \cap \hat{r} \le r_m) \\
&\le \mathbb{P}_s(\hat{r} > r_m) + \mathbb{P}_s(\|s-\hat{s}\|_2 \ge \hat{r}) \le \beta + \alpha. \square
\end{aligned}$$

The second lemma gives the separation rate for the test of null hypothesis $H_0 : s = s_0$

Lemma 5.4.8 *Let $\eta = 2(1-2\alpha-\beta)$, let Φ_α be the set of functions ϕ_α taking value in $\{0,1\}$ such that $P_{s_0}^{\otimes n}(\phi_\alpha) \ge 1-\alpha$, let r be a positive real number and $S(r)$ be the set of all densities s in S such that $\|s-s_0\|_2 \ge r$.*
Let $\beta\left(S(r)\right) = \inf_{\phi_\alpha\in\Phi_\alpha} \sup_{s\in S(r)} P_s^{\otimes n}(\phi_\alpha)$.
If $D \ge 10$ and $r^2 \le \sqrt{\ln(1+\eta^2)/3.2}(\sqrt{D-1}/n)$ then $\beta\left(S(r)\right) \ge \beta+\alpha$.

Comments: From Lemmas 5.4.7 and 5.4.8, we deduce that

$$r_m^2 \ge \sqrt{\frac{\ln(1+\eta^2)}{3.2}}\frac{\sqrt{D-1}}{4n}.$$

Thus the proof of Lemma 5.4.8 concludes the proof of Theorem 5.3.1.

Proof of lemma 5.4.8:

It is clear that the function $\beta\left(S(r)\right)$ is non-increasing with r. Thus we take $r^2 = \sqrt{\ln(1+\eta^2)/3.2}\sqrt{D-1}/n$ and we want to prove that $\beta\left(S(r)\right) \geq \alpha + \beta$. Let μ_r be a probability measure on $S(r)$ and let $P_{\mu_r} = \int P_s d\mu_r$, then we have

$$\begin{aligned}
\beta\left(S(r)\right) &\geq \inf_{\phi_\alpha \in \Phi_\alpha} P_{\mu_r}^{\otimes n}(\phi_\alpha) \\
&= \inf_{\phi_\alpha \in \Phi_\alpha} \left(P_{\mu_r}^{\otimes n}(\phi_\alpha) - P_{s_0}^{\otimes n}(\phi_\alpha) + P_{s_0}^{\otimes n}(\phi_\alpha)\right) \\
&\geq 1 - \alpha + \inf_{\phi_\alpha \in \Phi_\alpha} \left(P_{\mu_r}^{\otimes n}(\phi_\alpha) - P_{s_0}^{\otimes n}(\phi_\alpha)\right) \qquad (5.55) \\
&\geq 1 - \alpha - \sup_{A\,;\,P_{s_0}^{\otimes n}(A) \leq \alpha} \left|P_{\mu_r}^{\otimes n}(A) - P_{s_0}^{\otimes n}(A)\right| \\
&\geq 1 - \alpha - 1/2 \left\|P_{\mu_r}^{\otimes n} - P_{s_0}^{\otimes n}\right\|_{TV} \qquad (5.56)
\end{aligned}$$

where $\|.\|_{TV}$ denote the total variation distance. We can easily compute this distance if $P_{\mu_r}^{\otimes n}$ is absolutely continuous with respect to $P_{s_0}^{\otimes n}$. If $L_{\mu_r} = dP_{\mu_r}^{\otimes n}/dP_{s_0}^{\otimes n}$, we have

$$\left\|P_{\mu_r}^{\otimes n} - P_{s_0}^{\otimes n}\right\|_{TV} = \mathbb{E}_{s_0}\left|L_{\mu_r}(X_1,...,X_n) - 1\right| \leq \left(P_{s_0}^{\otimes n}\left(L_{\mu_r}^2\right) - 1\right)^{1/2}$$

and then

$$\beta\left(S(r)\right) \geq 1 - \alpha - \frac{\sqrt{P_{s_0}^{\otimes n}\left(L_{\mu_r}^2\right) - 1}}{2}. \qquad (5.57)$$

From (5.56) and (5.57), $\beta\left(S(r)\right) \geq \alpha + \beta$ if $P_{s_0}^{\otimes n}\left(L_{\mu_r}^2\right) \leq 1 + \eta^2$. Let us now give a probability measure on $S(r)$ such that $P_{s_0}^{\otimes n}\left(L_{\mu_r}^2\right) \leq 1 + \eta^2$.

Let $(\psi_\lambda)_{\lambda=1,...,[D/2]}$ be the following orthonormal system. Let $\psi_0 = s_0$, $\phi = 1_{[0,1/2)} - 1_{[1/2,1)}$ and for all $\lambda = 1,...,[D/2]$, $\psi_\lambda = \sqrt{D/2}\phi(Dx/2 - (\lambda - 1))$. We have $\left\|\sum_{\lambda=1}^{[D/2]} |\psi_\lambda|\right\|_\infty \leq \sqrt{D/2}$. We denote by μ_r the law of $s_\xi = s_0 + r\sum_{\lambda=1}^{[D/2]} \xi_\lambda \psi_\lambda / \sqrt{[D/2]}$ where $\xi = (\xi_\lambda)_{\lambda=1,...,[D/2]}$ are independent Rademacher random variables. Since $2\alpha + \beta < 1$, $\eta^2 \leq 4$ and $\ln(1+\eta^2) \leq \ln(5)$. $\sqrt{D} \leq n$, hence

$$r^2 \leq \sqrt{\frac{\ln(5)}{3.2}} \frac{\sqrt{D-1}}{n} \leq 1.$$

Finally, we have $\left\|\sum_{\lambda=1}^{[D/2]} \xi_\lambda \psi_\lambda\right\|_\infty / \sqrt{[D/2]} \leq 1$, thus s_ξ is a real density. Since $(\psi_\lambda)_{\lambda=1,..,[D/2]}$ is an orthonormal system, we have $\|s_\xi - s_0\|_2 = r$, thus s_ξ belongs to $S(r)$ and μ_r is a law on $S(r)$. Moreover

$$\frac{dP_{s_\xi}^{\otimes n}}{dP_{s_0}^{\otimes n}}(x_1,..,x_n) = \prod_{\alpha=1}^{n}\left(1 + \frac{r}{\sqrt{[D/2]}}\sum_{\lambda=1}^{[D/2]} \xi_\lambda \psi_\lambda(x_\alpha)\right).$$

Thus

$$L_{\mu_r}(x_1,..,x_n) = \frac{1}{2^{[D/2]}} \sum_{\xi \in \{-1,1\}^{[D/2]}} \prod_{\alpha=1}^{n}\left(1 + \frac{r}{\sqrt{[D/2]}}\sum_{\lambda=1}^{[D/2]} \xi_\lambda \psi_\lambda(x_\alpha)\right).$$

Hereafter, in order to symplify the notations, we write $\sum_\xi$ instead of $\sum_{\xi\in\{-1,1\}^{[D/2]}}$ and $\sum_\lambda$ instead of $\sum_{\lambda=1}^{[D/2]}$. Let $\Phi(r,\xi)=r\sum_\lambda \xi_\lambda\psi_\lambda/\sqrt{[D/2]}$, we have

$$\begin{aligned}L^2_{\mu_r}(x_1,..,x_n) &= \frac{1}{2^{2([D/2])}}\sum_{\xi,\xi'}\prod_{\alpha=1}^n\left(1+\Phi(r,\xi)(x_\alpha)\right)\left(1+\Phi(r,\xi')(x_\alpha)\right).\\ P_{s_0}^{\otimes n}(L^2_{\mu_r}) &= \frac{1}{2^{2[D/2]}}\sum_\xi\sum_{\xi'}\prod_{\alpha=1}^n P_{s_0}\left(1+\Phi(r,\xi)+\Phi(r,\xi')+\Phi(r,\xi)\Phi(r,\xi')\right).\end{aligned}$$

For all $\lambda\neq\lambda'=1,...,[D/2]$, $\psi_\lambda\psi_{\lambda'}=0$, thus

$$\Phi(r,\xi)\Phi(r,\xi')=\frac{r^2}{[D/2]}\left(\sum_\lambda \xi_\lambda\psi_\lambda\right)\left(\sum_\lambda \xi'_\lambda\psi_\lambda\right)=\frac{r^2}{[D/2]}\sum_\lambda \xi_\lambda\xi'_\lambda\psi^2_\lambda.$$

For all $\lambda=1,...,[D/2]$ and all $\alpha=1,...,n$, $P_{s_0}(\psi_\lambda)=0$, $P_{s_0}(\psi^2_\lambda)=1$, thus

$$\begin{aligned}P_{s_0}^{\otimes n}(L^2_{\mu_r}) &\leq \frac{1}{2^{2[D/2]}}\sum_\xi\sum_{\xi'}\left(1+\frac{r^2}{[D/2]}\sum_\lambda \xi_\lambda\xi'_\lambda\right)^n\\ &= \frac{1}{2^{2[D/2]}}\sum_\xi\sum_{l=0}^{[D/2]}\sum_{\xi';\mathrm{Card}(\lambda,\,\xi'_\lambda=\xi_\lambda)=l}\left[1+\frac{r^2}{[D/2]}(2l-[D/2])\right]^n\\ &= \frac{1}{2^{[D/2]}}\sum_{l=0}^{[D/2]}C^l_{[D/2]}\left[1+\frac{r^2 2l}{[D/2]}-r^2\right]^n\end{aligned}$$

For all real numbers $u\geq -1$, we have $0\leq 1+u\leq e^u$, thus $(1+u)^n\leq e^{nu}$. Since $r^2\leq 1$, we can apply this inequality to all the $u_l=(2l/[D/2]-1)r^2$ and we obtain

$$P_{s_0}^{\otimes n}(L^2_{\mu_\rho})\leq \frac{1}{2^{[D/2]}}\sum_{l=0}^{[D/2]}C^l_{[D/2]}\exp\left(\frac{r^2 2nl}{[D/2]}-nr^2\right)=\frac{e^{-nr^2}}{2^{[D/2]}}\left(\exp\left(\frac{r^2 2n}{[D/2]}\right)+1\right)^{[D/2]}$$

Thus, $P_{s_0}^{\otimes n}\left(L^2_{\mu_r}\right)\leq 1+\eta^2$ if

$$-nr^2+([D/2])\ln\left(\frac{\exp\left(\frac{r^2 2n}{[D/2]}\right)+1}{2}\right)\leq \ln(1+\eta^2).$$

For all positive u, $\ln(1+u)\leq u$, thus, we only have to prove that

$$-nr^2+\frac{[D/2]}{2}\left(\exp\left(\frac{r^2 2n}{[D/2]}\right)-1\right)\leq \ln(1+\eta^2).$$

$[D/2]\geq (D-1)/2$ and $D\geq 10$, thus

$$\frac{r^2 2n}{[D/2]}=2\sqrt{\frac{\ln(1+\eta^2)}{3.2}}\frac{\sqrt{D-1}}{[D/2]}\leq\frac{4*0.71}{\sqrt{D-1}}\leq 1.$$

For all real numbers x in $[0,1]$, we have $e^x\leq 1+x+3.2x^2$, thus $\exp\left(r^2 2n/([D/2])\right)-1\leq r^2 2n/([D/2])+3.2\left(r^2 n/([D/2])\right)^2$. Hence

$$-nr^2+\frac{[D/2]}{2}\left(\exp\left(\frac{r^2 2n}{[D/2]}\right)-1\right)\leq 1.6r^4n^2/([D/2])\leq\frac{D-1}{2[D/2]}\ln(1+\eta^2)\leq\ln(1+\eta^2).$$

5.4.5 Proof of Proposition 5.3.2:

Inequalities (5.18) and (5.19) are trivial when $x \leq 1$, thus we prove them for $x \geq 1$. U_m is a totally degenerate U-statistic of order 2. We want to apply inequality (5.26) with $\Lambda_m = \Lambda - m$. Let S'_m be the linear span of $(\psi_\lambda)_{\lambda \in \Lambda - m}$. Since s satisfies assumption $[D]$, we have $v_m^2 = \sup_{t \in B_1(S'_m)} \mathrm{Var}(t(X)) \leq \sup_{t \in B_1(S'_m)} Pt^2 \leq M$. Since the constant functions belongs to S_n, for all function t in S_n, $t - Pt$ belongs to S_n. Thus, from Assumption $[M_1]$, we have

$$b_m = \sup_{t \in B_1(S'_m)} \|t - Pt\|_\infty \leq \sup_{t \in B_1(S_n)} \|t - Pt\|_\infty \leq \sup_{t \in B_1(S_n)} \|t\|_\infty \leq \Phi_0 \sqrt{D}.$$

From inequality (5.26), we have, with probability larger than $1 - 2.8e^{-x}$

$$\xi U_m \leq 5.7\sqrt{M}\frac{\Phi_0\sqrt{Dx}}{n} + 8M\frac{x}{n} + 384\Phi_0\sqrt{MD}\left(\frac{x}{n}\right)^{3/2} + 1020\left(\frac{\Phi_0\sqrt{Dx}}{n}\right)^2.$$

We apply inequality $2yz \leq \epsilon y^2 + \epsilon^{-1}z^2$ to $y = \sqrt{Mx/n}$ and $z = \Phi_0\sqrt{D}x/n$ and we obtain, with probability larger than $1 - 2.8e^{-x}$

$$\xi U_m \leq 5.7\frac{\Phi_0\sqrt{MDx}}{n} + (8 + 192\epsilon)M\frac{x}{n} + \left(1020 + \frac{192}{\epsilon}\right)\Phi_0^2\frac{Dx^2}{n^2}. \tag{5.58}$$

This proves (5.18). In order to prove (5.19), note that

$$V_m - \|s_m - s_n\|_2^2 = U_m + \frac{2}{n}\sum_{i=1}^{n}\sum_{\lambda \in \Lambda - m} P\psi_\lambda(\psi_\lambda(X_i) - P\psi_\lambda). \tag{5.59}$$

Equality (5.59) corresponds to the Hoeffding's decomposition of the U-statistic V_m. Let

$$T(x) = \sum_{\lambda \in \Lambda - m} P\psi_\lambda(\psi_\lambda(x) - P\psi_\lambda) = (s_n - s_m)(x) - P(s_n - s_m).$$

Since $s_n - s_m$ belongs to S, we have

$$\|s_n - s_m\|_\infty \leq \Phi_0\sqrt{D}\|s_m - s_n\|_2.$$

Thus, $\|T\|_\infty \leq 2\Phi_0\sqrt{D}\|s_m - s_n\|_2$. Moreover, from Assumption **[D]**, we have

$$PT^2 \leq P(s_n - s_m)^2 \leq M\|s_n - s_m\|_2^2.$$

Thus Bernstein's inequality gives

$$\forall x > 0,\ \mathbb{P}\left(\xi P_n T > \sqrt{\frac{2M\|s_n - s_m\|_2^2 x}{n}} + \frac{2\Phi_0\sqrt{D}\|s_m - s_n\|_2 x}{3n}\right) \leq e^{-x}.$$

Since $2ab \leq a^2/4 + 4b^2$, we obtain, for all $x > 0$,

$$\mathbb{P}\left(\xi\frac{2}{n}\sum_{i=1}^{n}\sum_{\lambda \in \Lambda - m} P\psi_\lambda(\psi_\lambda(X_i) - P\psi_\lambda) > \frac{1}{2}\|s_m - s_n\|_2^2 + 8M\frac{x}{n} + \frac{16\Phi_0^2 Dx^2}{9n^2}\right) \leq e^{-x}. \tag{5.60}$$

Plugging inequalities (5.18) and (5.60) in (5.59), we obtain (5.19).

5.4.6 Proof of Theorem 5.3.3:

From Pythagoras theorem, we have, for all $m \in \mathcal{M}_n$

$$\|s-\hat{s}_m\|_2^2 = \|s-s_n\|_2^2 + \|s_n-s_m\|_2^2 + \|s_m-\hat{s}_m\|_2^2 \le \eta^2 + \|s_n-s_m\|_2^2 + \|s_m-\hat{s}_m\|_2^2 . \tag{5.61}$$

Since s satisfies Assumption [D], we have $\|s\|_2^2 \le M$, thus for all m in $\mathcal{M}_n$, Inequality (5.4) in Theorem 5.2.1 ensures that

$$\mathbb{P}_s\left(\|s_m-\hat{s}_m\|_2^2 > v_m^2\right) \le \alpha_m \text{ and } \mathbb{P}_s\left(\|s_n-\hat{s}_n\|_2^2 > R_n^2\right) \le \alpha/2.$$

Inequality (5.19) in Proposition 5.3.2 gives $\mathbb{P}_s\left(\|s_n-s_m\|_2^2 > b_m^2\right) \le \alpha_m$. Hence

$$\begin{aligned}
\mathbb{P}_s\left(s \notin B(\tilde{s},\tilde{R})\right) &\le \mathbb{P}_s\left(\exists m \in \mathcal{M}_n \cup \{n\},\ s \notin B(\hat{s}_m, R_m^2)\right)\\
&= \mathbb{P}_s\left(\|s_n-\hat{s}_n\|_2^2 > R_n^2\right) + \mathbb{P}_s\left(\exists m \in \mathcal{M}_n,\ \|s_n-\hat{s}_m\|_2^2 > b_m^2 + v_m^2\right)\\
&\le \frac{\alpha}{2} + \sum_{m\in\mathcal{M}_n} \mathbb{P}_s\left(\|s_n-s_m\|_2^2 > b_m^2\right) + \mathbb{P}_s\left(\|\hat{s}_m-s_m\|_2^2 > v_m^2\right))\\
&\le \frac{\alpha}{2} + 2\sum_{m\in\mathcal{M}_n} \alpha_m \le \alpha,
\end{aligned}$$

where we use successively the definition of $\tilde{R}$, the decomposition (5.61) and $\sum_{m\in\mathcal{M}_n}\alpha_m \le \alpha/4$. Since $x_m \le C_*(n/\sqrt{D})^{2/3}$, we have $Dx_m^2/n^2 \le C_*\sqrt{Dx_m}/n$. Moreover, when $d(s_n,S_m) \le \eta_m$, Inequality (5.19) in Proposition 5.3.2 gives a constant C such that

$$\mathbb{P}_s\left(V_m > \frac{3}{2}\eta_m^2 + \frac{C}{n}\left(\sqrt{Dx_m} + x_m\right)\right) \le \frac{\beta}{2}. \tag{5.62}$$

Since $\sqrt{D} \le n$, we have $x_m \le C_*(n/\sqrt{D})^{2/3} \le C_* n/\sqrt{D} \le C_* n/\sqrt{D_m}$. Thus, from Corollary (5.2.2), there exists a constant C such that

$$\mathbb{P}\left(v_m^2 > \frac{C}{n}\left(D_m + \sqrt{D_m}x_m\right)\right) \le \frac{\beta}{2}. \tag{5.63}$$

Thus, when $d(s_n,S_m) \le \eta_m$, there exists a constant C which may vary from line to line such that

$$\begin{aligned}
&\mathbb{P}_s\left(\tilde{R}^2 > \frac{C}{n}\left(D_m + \sqrt{D_m}x_m + \sqrt{Dx_m}\right) + \eta^2 + 3\eta_m^2\right)\\
&\le \mathbb{P}_s\left(R_m^2 > \frac{C}{n}\left(D_m + \sqrt{D_m}x_m + \sqrt{Dx_m}\right) + \eta^2 + 3\eta_m^2\right)\\
&\le \mathbb{P}_s\left(v_m^2 + b_m^2 > \frac{C}{n}\left(D_m + \sqrt{D_m}x_m + \sqrt{Dx_m}\right) + 3\eta_m^2\right)\\
&\le \mathbb{P}_s\left(v_m^2 > \frac{C}{n}\left(D_m + \sqrt{D_m}x_m\right)\right) + \mathbb{P}_s\left(b_m^2 > \frac{C}{n}\left(\sqrt{Dx_m} + x_m\right) + 3\eta_m^2\right)\\
&\le \mathbb{P}_s\left(V_m > \frac{3}{2}\eta_m^2 + \frac{C}{n}\left(D_m + \sqrt{D_m}x_m\right)\right) + \frac{\beta}{2} \le \beta,
\end{aligned}$$

where we use successively the inequality $\tilde{R} \le R_m$, the definition of R_m, Inequalities (5.21) and (5.63) and Inequality (5.62).

5.4.7 Proof of Corollary 5.3.4:

Since $\alpha_m = \alpha/(4|\mathcal{M}_n|)$, we have $\ln(1/\alpha_m) \leq C\ln((\ln n)/\alpha) \leq C\sqrt{n/D}$. When s belongs to $B_{w,2,\infty}(M_1)$, inequality (5.25) ensures that

$$d^2(s, S_n) \leq \frac{M_1^2}{4(4^w - 1)} 2^{-2Jw}$$

so that the covering property is straightforward from (5.23). Our choices of α_m and D imply that

$$\eta^2 \leq C(n/\sqrt{\ln(\ln n)})^{-4w/(4w+1)} \text{ and that } \sqrt{Dx_m}/n \leq C(n/\sqrt{\ln(\ln n)})^{-4w/(4w+1)}.$$

When s belongs to $B_{v,2,\infty}(M_1)$ for some $v \geq w$, then inequality (5.25) ensures that

$$d^2(s, S_m) \leq \frac{M_1^2}{4(4^v - 1)} 2^{-2J_m v}.$$

We apply Inequality (5.24) with $\eta_m^2 = M_1^2 2^{-2J_m v}/(4(4^v - 1))$ to the space S_m such that $n^{1/(2v+1)}/2 \leq D_m \leq n^{1/(2v+1)}$. We have $D_m/n \leq n^{-2v/(2v+1)}$ and there exists a constant C such that $\eta_m^2 \leq Cn^{-2v/(2v+1)}$. This proves the result.

Bibliography

[1] H. Akaike. Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22:203–217, 1970.

[2] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[3] D. W. K. Andrews. Nonstrong mixing autoregressive processes. *J. Appl. Probab.*, 21(4):930–934, 1984.

[4] S. Arlot. *Resampling and model selection.* PhD thesis, Université Paris-Sud 11, 2007.

[5] S. Arlot. Model selection by resampling penalization. *Electron. J. Statist.*, 3:557–624, 2009.

[6] S. Arlot, G. Blanchard, and E. Roquain. Resampling-based confidence regions and multiple tests for a correlated random vector. In *Learning theory*, volume 4539 of *Lecture Notes in Comput. Sci.*, pages 127–141. Springer, Berlin, 2007.

[7] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine learning research*, 10:245–279, 2009.

[8] Y. Baraud. Confidence balls in Gaussian regression. *Ann. Statist.*, 32(2):528–551, 2004.

[9] Y. Baraud, F. Comte, and G. Viennet. Adaptive estimation in autoregression or β-mixing regression via model selection. *Ann. Statist.*, 29(3):839–875, 2001.

[10] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[11] R. Beran. REACT scatterplot smoothers: superefficiency through basis economy. *J. Amer. Statist. Assoc.*, 95(449):155–171, 2000.

[12] R. Beran and L. Dümbgen. Modulation of estimators and confidence sets. *Ann. Statist.*, 26(5):1826–1856, 1998.

[13] H. C. P. Berbee. *Random walks with stationary increments and renewal theory*, volume 112 of *Mathematical Centre Tracts*. Mathematisch Centrum, Amsterdam, 1979.

[14] L. Birgé. Model selection for density estimation with l^2-loss. *Preprint*, 2008.

[15] L. Birgé and P. Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[16] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001.

[17] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

[18] O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.

[19] R. C. Bradley. *Introduction to strong mixing conditions. Vol. 1.* Kendrick Press, Heber City, UT, 2007.

[20] T. Cai and M. G. Low. Adaptive confidence balls. *Ann. Statist.*, 34(1):202–228, 2006.

[21] A. Célisse. Density estimation via cross validation: Model selection point of view. *Preprint, downloadable on arXiv.org : 08110802*, 2008.

[22] F. Comte, J. Dedecker, and M. L. Taupin. Adaptive density deconvolution with dependent inputs. *Math. Methods Statist.*, 17(2):87–112, 2008.

[23] F. Comte and F. Merlevède. Adaptive estimation of the stationary density of discrete and continuous time mixing processes. *ESAIM Probab. Statist.*, 6:211–238 (electronic), 2002. New directions in time series analysis (Luminy, 2001).

[24] J. Dedecker, P. Doukhan, G. Lang, J. R. León R., S. Louhichi, and C. Prieur. *Weak dependence: with examples and applications*, volume 190 of *Lecture Notes in Statistics*. Springer, New York, 2007.

[25] J. Dedecker and C. Prieur. New dependence coefficients. Examples and applications to statistics. *Probab. Theory Related Fields*, 132(2):203–236, 2005.

[26] R. A. DeVore and G. G. Lorentz. *Constructive approximation*, volume 303 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1993.

[27] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2):508–539, 1996.

[28] P. Doukhan. *Mixing*, volume 85 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1994. Properties and examples.

[29] R. M. Dudley. A course on empirical processes. In *École d'été de probabilités de Saint-Flour, XII—1982*, volume 1097 of *Lecture Notes in Math.*, pages 1–142. Springer, Berlin, 1984.

[30] B. Efron. Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 1979.

[31] B. Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78(382):316–331, 1983.

[32] M. Fromont. Model selection by bootstrap penalization for classification. *Machine Learning*, 66(2, 3):165–207, 2007.

[33] M. Fromont and B. Laurent. Adaptive goodness-of-fit tests in a density model. *Ann. Statist.*, 34(2):680–720, 2006.

[34] I. Gannaz and O. Wintenberger. Adaptive density estimation under dependence. *forthcoming in ESAIM, Probab. and Statist.*, 2009.

[35] C. Genovese and L. Wasserman. Adaptive confidence bands. *Ann. Statist.*, 36(2):875–905, 2008.

[36] C. R. Genovese and L. Wasserman. Confidence sets for nonparametric wavelet regression. *Ann. Statist.*, 33(2):698–729, 2005.

[37] E. Giné and R. Nickl. Uniform limit theorems for wavelet density estimators. *Annals of Probability*, forthcoming.

[38] M. Hoffmann and O. Lepski. Random rates in anisotropic regression. *Ann. Statist.*, 30(2):325–396, 2002. With discussions and a rejoinder by the authors.

[39] C. Houdré and P. Reynaud-Bouret. Exponential inequalities, with constants, for U-statistics of order two. In *Stochastic inequalities and applications*, volume 56 of *Progr. Probab.*, pages 55–69. Birkhäuser, Basel, 2003.

[40] Y. I. Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. I. *Math. Methods Statist.*, 2(2):85–114, 1993.

[41] Y. I. Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. II. *Math. Methods Statist.*, 2(3):171–189, 1993.

[42] Y. I. Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. III. *Math. Methods Statist.*, 2(4):249–268, 1993.

[43] A. Juditsky and S. Lambert-Lacroix. Nonparametric confidence set estimation. *Math. Methods Statist.*, 12(4):410–428 (2004), 2003.

[44] H.R. Künsch. The jackknife and the bootstrap for general stationary observations. *Ann. Statist.*, 17:1217–1241, 1989.

[45] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000.

[46] M. Ledoux. On Talagrand's deviation inequalities for product measures. *ESAIM Probab. Statist.*, 1:63–87 (electronic), 1995/97.

[47] M. Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.

[48] M. Ledoux and M. Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. Isoperimetry and processes.

[49] M. Lerasle. Adaptive density estimation of stationary β-mixing and τ-mixing sequences. *Mathematical Methods of statistics*, 2009.

[50] K.C. Li. Honest confidence regions for nonparametric regression. *Ann. Statist.*, 17(3):1001–1008, 1989.

[51] R.Y. Liu and K. Singh. Moving block jackknife and bootstrap capture weak dependence. *R.Lepage & L. Billard eds, Exploring the limits of bootstrap (wiley, New York)*, pages 225–248, 1992.

[52] M. G. Low. On nonparametric confidence intervals. *Ann. Statist.*, 25(6):2547–2554, 1997.

[53] C.L. Mallows. Some comments on c_p. *Technometrics*, 15:661–675, 1973.

[54] D. M. Mason and M. A. Newton. A rank statistics approach to the consistency of a general bootstrap. *Ann. Statist.*, 20(3):1611–1624, 1992.

[55] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[56] D. Pollard. *Empirical processes: theory and applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, 2. Institute of Mathematical Statistics, Hayward, CA, 1990.

[57] J. Præstgaard and J. A. Wellner. Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.*, 21(4):2053–2086, 1993.

[58] C. Prieur. Change point estimation by local linear smoothing under a weak dependence condition. *Math. Methods Statist.*, 16(1):25–41, 2007.

[59] D. Radulović. On the bootstrap and empirical processes for dependent sequences. In *Empirical process techniques for dependent data*, pages 345–364. Birkhäuser Boston, Boston, MA, 2002.

[60] E. Rio. *Théorie asymptotique des processus aléatoires faiblement dépendants*, volume 31 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2000.

[61] J. Robins and A. van der Vaart. Adaptive nonparametric confidence sets. *Ann. Statist.*, 34(1):229–253, 2006.

[62] M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.*, 9(2):65–78, 1982.

[63] R. Shibata. An optimal selection of regression variables. *Biometrika*, 68(1):45–54, 1981.

[64] M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147, 1974. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors.

[65] M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505–563, 1996.

[66] A. B. Tsybakov. *Introduction à l'estimation non-paramétrique*, volume 41 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2004.

[67] S. A. van de Geer. *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2000.

[68] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[69] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[70] G. Viennet. Inequalities for absolutely regular sequences: application to density estimation. *Probab. Theory Related Fields*, 107(4):467–492, 1997.

[71] V. A. Volkonskiĭ and Y. A. Rozanov. Some limit theorems for random functions. I. *Teor. Veroyatnost. i Primenen*, 4:186–207, 1959.