Estimation non paramétrique adaptative pour les chaînes de Markov et les chaînes de Markov cachées

Claire Lacour

UNIVERSITÉ PARIS DESCARTES
UFR de Mathématiques et Informatique
École doctorale Mathématiques Paris Centre

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PARIS DESCARTES

Discipline : Mathématiques

Présentée par

Claire LACOUR

Estimation non paramétrique adaptative pour les chaînes de Markov et les chaînes de Markov cachées

Soutenue le 1er octobre 2007 devant le jury composé de :

Lucien BIRGÉ	Président du jury
Stéphan CLÉMENÇON	
Fabienne COMTE	Directrice de thèse
Valentine GENON-CATALOT	
Marc HOFFMANN	Rapporteur
Pascal MASSART	Rapporteur
Judith ROUSSEAU	

Remerciements

Je tiens en premier lieu à remercier chaleureusement Fabienne, qui fut en tous points une excellente directrice de thèse, sa modestie dût-elle en souffrir. Dès le stage de DEA et par la suite dans la thèse, elle fut toujours très disponible sans être contraignante. Son dynamisme a rendu chaque rendez-vous particulièrement stimulant et encourageant. Je lui suis également redevable d'avoir trouvé un sujet tout à la fois intéressant et à ma portée. Enfin, je la remercie d'avoir patienté un an que je passe l'agrégation avant de pouvoir commencer la thèse.

Je remercie vivement Marc Hoffmann pour avoir consenti à évaluer cette thèse, mais aussi pour son expertise concernant les bornes inférieures et son enthousiasme en général. Je suis également reconnaissante à Pascal Massart d'avoir consacré une partie de son temps à rapporter cette thèse. Judith Rousseau, Stéphan Clémençon et Valentine Genon-Catalot m'ont fait l'honneur de faire partie du jury, qu'ils en soient remerciés ici. Ma gratitude se porte enfin à Lucien Birgé qui a bien voulu marquer son intérêt pour mon travail en présidant ce jury.

Au cours de cette thèse plusieurs personnes m'ont fait profiter de leurs compétences, je remercie en particulier Eva Löscherbach, ainsi qu'Erwan Lepennec sans qui le chapitre 5 ne contiendrait peut-être aucune simulation. Un grand merci également à tous ceux qui ont participé à la relecture du manuscrit, ils se reconnaîtront.

Mon intêret pour les Statistiques vient sans aucun doute du cours de maîtrise de Dominique Picard, que je remercie pour sa rigueur, doublé des travaux dirigés de Marc Hoffmann dont je garde un excellent souvenir. Cet interêt n'a pu que se renforcer à l'écoute de l'excellent cours de préparation à l'agrégation de Patricia Reynaud-Bouret.

Je profite de cette tribune pour exprimer ma gratitude aux nombreux professeurs qui m'ont fait partager leur amour des mathématiques : Mme Renaud pour sa clairvoyance, Mme Petazzoni pour son intransigeance, Mme Brémond pour son sérieux, M. Randé pour son extravagance ainsi que Mme Picaronny pour sa pédagogie.

Le MAP5 est un havre de paix, particulièrement accueillant pour les doctorants. J'ai toujours pu disposer de tout le matériel dont j'avais besoin, et partir régulièrement en congrès. Je voudrais en particulier exprimer toute ma reconnaissance à Marie-Hélène Gbaguidi et Laurent Moineau que j'ai si souvent sollicités. Mais je tiens à remercier tout le laboratoire, avec une pensée particulière pour Antoine, Elodie, Hermine et Adeline.

Ce travail m'aurait sans doute paru beaucoup plus laborieux sans le soutien au quotidien de Béatrice, Javiera, Amandine, Gwendoline, Cécile, Claire J., Arno, Sylvain et

Mohamed. Je n'oublie pas les thésards de Chevaleret et d'ailleurs, avec qui j'ai partagé de nombreux séminaires : Matthieu, Séverine, Ismaël, Olivier, Guillaume, Pierre, Thomas, Etienne, Christophe, Alexandre, Stéphane, Jean-Patrick, Tu, Fanny.

Je voudrais remercier particulièrement Nathalie, Marie et Cécile qui ont apporté un peu de finesse dans ce monde d'hommes, et avec qui j'ai pu partager les bons moments comme les difficultés. Je pense également à tous mes amis rôlistes : Sophie, Cédric, Marc et Héloïse, Claire et Benjamin, Benny, Silvère, Tonio, Seb, Erwan et Yiming. Que serait la vie sans les soirées jeux, les randos, restos, pique-niques et autres enterrements de vie de célibataire ?

Je ne peux conclure cette page sans adresser une pensée à mes parents, en particulier pour m'avoir toujours laissé une entière liberté dans mes choix. Je remercie également Philippe pour son aide informatique et ses questions mathématiques impromptues, ainsi qu'Aline qui a toujours su me conseiller et me réconforter.

Enfin, merci à Nicolas sans qui tout cela n'aurait aucune importance.

Table des matières

Introduction

Généralités

Le but de cette thèse est d'estimer la transition d'une chaîne de Markov observée ou cachée. Les chaînes de Markov que nous considérons étant à espace d'états continu, la densité de transition apparaît comme une fonction. L'estimation d'une telle quantité relève donc des statistiques non paramétriques, qui ont pour but d'estimer une fonction, plutôt qu'un paramètre de $\mathbb{R}^d$.

En statistique, toute quantité calculable à partir des observations est appelée estimateur. Il s'agit ensuite de mesurer la qualité de cet estimateur. Pour cela, on évalue une distance entre l'estimateur et l'objet à estimer. Dans le cadre de l'estimation d'une fonction, les distances les plus couramment employées sont la distance ponctuelle ou bien les normes L^p. Plus précisément, si $\hat{s}$ est un estimateur de $s : \mathbb{R}^d \mapsto \mathbb{R}$, les fonctions de perte sont $\rho(s, \hat{s}) = |s(x_0) - \hat{s}(x_0)|$ ou bien

$$\|s - \hat{s}\|_p^p = \int_A |s(x) - \hat{s}(x)|^p dx \quad \text{avec } A \subset \mathbb{R}^d.$$

L'erreur d'estimation est alors $\mathbb{E}\rho(s, \hat{s})$; elle est appelée erreur ponctuelle dans le premier cas, erreur $L^p(A)$ dans le second. Dans ce travail, nous ne nous intéresserons qu'à la norme L^2 ($L^2(A)$ ou $L^2(\mathbb{R}^d)$ avec $d = 1$ ou 2 selon les cas)[(1)]. L'erreur est alors appelée risque quadratique, ou MISE, abréviation pour *mean integrated squared error*, et s'écrit

$$\mathbb{E}\|s - \hat{s}\|_2^2.$$

Pour apprécier la qualité de nos estimateurs, nous nous plaçons dans cette thèse dans le cadre minimax décrit ci-dessous.

Si n est le nombre d'observations, il est naturel d'espérer que le risque décroisse en fonction de n : plus on dispose d'information, mieux on estime la fonction. Mais on souhaite quantifier la vitesse de décroissance du risque en fonction de n. Si V est une classe de fonctions à laquelle s est supposée appartenir, la quantité d'intérêt est

$$R_n(V) = \inf_{\hat{s}_n} \sup_{s \in V} \mathbb{E}\|s - \hat{s}\|_2^2$$

où l'infimum est pris sur tous les estimateurs $\hat{s}_n = \hat{s}_n(X_1, \ldots, X_n)$ de s. Ce nombre est appelé risque minimax pour des raisons évidentes. On dit que l'estimateur $\hat{s}$ atteint la vitesse r_n s'il existe une constante C strictement positive telle que

$$\sup_{s \in V} \mathbb{E}\|s - \hat{s}\|_2^2 \leq C r_n.$$

[(1)] En effet nous utilisons le cadre hilbertien des estimateurs par projection. Le lecteur intéressé par des risques L^p avec $p \geq 1$ peut se reporter à Clémençon (1999). A la différence de ces travaux, nous nous sommes davantage intéressés à la faisabilité des procédures d'estimation ainsi qu'à l'obtention d'un estimateur sans perte logarithmique dans la vitesse.

Cette vitesse est dite optimale s'il n'existe pas d'estimateur ayant une meilleure vitesse, c'est-à-dire s'il existe une constante C' strictement positive telle que

$$R_n(V) \geq C' r_n.$$

La notion de vitesse optimale n'est donc définie qu'à une constante près. Dans ce travail, on considère différentes classes de régularité V :

- Des boules de l'espace de Besov $B^{\alpha}_{2,\infty}$ en dimension 1 ou 2 selon les cas. En effet cet espace contient l'espace de Sobolev W^{α}_2 et tous les espaces de Besov $B^{\alpha}_{2,q}$ pour $q \geq 1$. De plus on a $B^{\alpha}_{p,\infty} \subset B^{\alpha}_{2,\infty}$ pour tout $p \geq 2$ (on peut se référer à DeVore et Lorentz (1993) pour une définition des espaces de Besov et à Triebel (1983) pour les inclusions entre espaces de Besov).
- Des classes de régularités de type

$$\mathcal{A}_{\delta,r,a}(l) = \left\{ f : \mathbb{R} \mapsto \mathbb{R}, \quad \int |f^*(x)|^2 (x^2+1)^{\delta} \exp(2a|x|^r) dx \leq l \right\}$$

  et leur équivalent en dimension 2. Lorsque $r = 0$, il s'agit d'une classe de Sobolev. Pour $r > 0$, la fonction est dite super régulière ; elle est infiniment différentiable et analytique dès que $r > 1$.

En règle générale, nous n'estimerons s que sur un compact A, ce qui revient à estimer $s\mathbb{1}_A$ à la place de s. Dans ce cas, on utilise la norme L^2 sur A.

Adaptation et sélection de modèles

Dans cette thèse, tous les estimateurs considérés sont des estimateurs par projection. Le principe de ce type d'estimation est d'estimer une fonction s en approchant sa projection sur un espace d'approximation appelé modèle. Si S_m est un sous-espace du Hilbert L^2 de dimension D_m engendré par une base orthonormale $(e_1, \ldots, e_{D_m})$, on note s_m la projection orthogonale de s sur S_m : $s_m = \sum_{j=1}^{D_m} a_j e_j$ avec $a_j =< s, e_j >$. L'estimateur de s est alors de la forme $\hat{s}_m = \sum_{j=1}^{D_m} \hat{a}_j e_j$. Le risque peut s'écrire

$$\mathbb{E}\|s - \hat{s}_m\|_2^2 = \|s - s_m\|_2^2 + \mathbb{E}\|s_m - \hat{s}_m\|_2^2 = \|s - s_m\|_2^2 + \sum_{j=1}^{D_m} \mathbb{E}|\hat{a_j} - a_j|^2.$$

Cette décomposition divise le risque quadratique en deux termes :

- le biais ou erreur d'approximation $\|s - s_m\|_2^2$ due à la méthode d'approximation utilisée,
- la variance ou erreur stochastique $\mathbb{E}\|s_m - \hat{s}_m\|_2^2$ due au caractère aléatoire des observations.

Le terme de biais décroît en fonction de D_m car la fonction est de mieux en mieux approchée par s_m. En revanche le terme de variance augmente en fonction de D_m parce que le nombre de coefficients à estimer est de plus en plus grand. Il y a donc un arbitrage à faire sur D_m pour que l'erreur soit minimale, le D_m^* optimal étant celui qui réalise le meilleur compromis biais-variance. Si $D_m^* = \arg\min_{D_m} \mathbb{E}\|s - \hat{s}_m\|_2^2$, l'estimateur $\hat{s}_{D_m^*}$ est appelé oracle.

En règle générale, le biais est de l'ordre de $D_m^{-2\alpha}$ où α est la régularité de la fonction s estimée, et la variance est de l'ordre d'une fonction croissante de D_m (par exemple une puissance) divisée par le nombre d'observations n. Le D_m optimal est donc fonction de n et de la régularité α. Cependant, puisque la fonction s n'est pas connue (on cherche à l'estimer à partir des observations), sa régularité α ne l'est *a priori* pas non plus. On ne peut donc pas choisir le D_m^* optimal évoqué précédemment. On retrouve le même problème de choix optimal, cette fois de la fenêtre, dans les méthodes d'estimation par noyau.

Pour pallier cette difficulté, de nouvelles procédures se sont développées dans les années quatre-vingt-dix. Ces procédures permettent de trouver un estimateur ayant les mêmes performances que l'oracle mais construit seulement à partir des données, c'est-à-dire ne nécessitant aucune connaissance préalable de la régularité de la fonction. Ces estimateurs sont appelés adaptatifs car ils s'adaptent à la régularité de la fonction à estimer.

Les deux grandes familles de procédures adaptatives sont le seuillage en ondelettes et la sélection de modèles. La première se fonde sur un estimateur par projection sur un espace d'ondelettes. Mais les coefficients d'ondelettes de l'estimateur les plus petits sont abandonnés pour ne garder que ceux qui dépassent un certain seuil (on peut se reporter à Donoho *et al.* (1996) ou Härdle *et al.* (1998) par exemple). C'est cette méthode qui a été utilisée par Clémençon (1999). Elle a cependant le désagrément de produire un estimateur qui en général n'est pas exactement minimax car on observe une perte logarithmique dans la vitesse de convergence.

Dans cette thèse, nous avons utilisé la procédure plus générale de sélection de modèles (voir Birgé et Massart (1997), Barron *et al.* (1999) ou encore Massart (2007)). Expliquons en quoi elle consiste. On estime la fonction s par projection sur différents modèles S_m (dont le nombre peut dépendre de n) en écrivant

$$\hat{s}_m = \arg\min_{t \in S_m} \gamma_n(t).$$

Ici γ_n est un certain contraste, c'est-à-dire que la fonction $t \mapsto \mathbb{E}[\gamma_n(t)]$ atteint son minimum au point s. On dispose ainsi d'une famille d'estimateurs $(\hat{s}_m)_{m \in \mathcal{M}_n}$. Puisqu'on ne peut pas minimiser directement le risque des estimateurs de cette famille (seul l'oracle peut le faire), la première idée pour trouver le meilleur estimateur serait de minimiser en m la version empirique du risque $\gamma_n(\hat{s}_m)$. Le problème est que γ_n et $\hat{s}_m$ sont étroitement liés. En effet chaque S_m contient les estimateurs $\hat{s}_{m'}$ pour $S_{m'} \subset S_m$ et donc $\gamma_n(\hat{s}_m) \leq \gamma_n(\hat{s}_{m'})$. Cela montre que le risque empirique décroît en fonction de la taille du modèle, ce qui n'est pas le cas du vrai risque. Pour les gros modèles, le risque empirique sous-évalue le

vrai risque. On va donc compenser en *pénalisant* le risque empirique. On choisit un $\hat{m}$ (ne dépendant que des données) qui minimise le critère

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

où pen : $\mathcal{M}_n \mapsto \mathbb{R}^+$ est une fonction de pénalité.

Si la pénalité est bien choisie, on peut alors prouver que l'estimateur $\tilde{s} = \hat{s}_{\hat{m}}$ vérifie

$$\mathbb{E}\|s - \tilde{s}\|_2^2 \leq C \inf_m \{\|s - s_m\|_2^2 + \text{pen}(m)\} + R_n \tag{1}$$

où C est une constante strictement positive que l'on espère proche de 1, et R_n est un reste. Si le reste est d'ordre faible et si la pénalité a le même ordre que le terme de variance, on obtient pour l'estimateur $\tilde{s}$ la même vitesse de convergence que celle de l'oracle, qui est souvent la vitesse minimax optimale. Pour obtenir (1) on utilise des inégalités de concentration car R_n correspond généralement au contrôle d'un processus empirique. Dans ce travail, nous utilisons fréquemment l'inégalité suivante, adaptée des travaux de Talagrand (1996) et prouvée au chapitre 4 section 4.7.8.

Lemme 1 *Soient $T_1, \ldots, T_n$ des variables aléatoire indépendantes. Pour tout r appartenant à une classe dénombrable de fonctions $\mathcal{R}$, on pose $\nu_n(r) = (1/n)\sum_{i=1}^n [r(T_i) - \mathbb{E}(r(T_i))]$. Alors, pour $\epsilon > 0$,*

$$\mathbb{E}[\sup_{r\in\mathcal{R}} |\nu_n(r)|^2 - 2(1+2\epsilon)H^2]_+ \leq K\left(\frac{v}{n} e^{-K_1 \epsilon \frac{nH^2}{v}} + \frac{M_1^2}{n^2 C^2(\epsilon)} e^{-K_2 C(\epsilon)\sqrt{\epsilon}\frac{nH}{M_1}}\right)$$

avec $K_1 = 1/6$, $K_2 = 1/(21\sqrt{2})$, $C(\epsilon) = \sqrt{1+\epsilon} - 1$, K une constante universelle et où

$$\sup_{r\in\mathcal{R}} \|r\|_\infty \leq M_1, \quad \mathbb{E}\left(\sup_{r\in\mathcal{R}} |\nu_n(r)|\right) \leq H, \quad \sup_{r\in\mathcal{R}} \frac{1}{n}\sum_{i=1}^n \text{Var}(r(T_i)) \leq v.$$

Des arguments de densités classiques permettent d'étendre ce résultat à une classe non-dénombrable de fonctions $\mathcal{R}$.

En ce qui concerne le choix des modèles S_m, il est important qu'ils vérifient une hypothèse de connexion de normes du type

$$\exists \phi_0 > 0 \quad \forall t \in S_m \quad \|t\|_\infty \leq \phi_0 \sqrt{D_m} \|t\|_2 \tag{2}$$

où D_m désigne la dimension de S_m. Dans cette thèse, nous avons considéré différents types de modèles selon les contraintes du modèle statistique utilisé, mais ils vérifient toujours la propriété (2).

Chaînes de Markov

a/ Définition

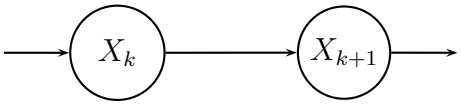
Puisque notre travail consiste à estimer la transition d'une chaîne de Markov, rappelons ici ce qu'est une chaîne de Markov et quelles sont ses propriétés. Pour une référence générale sur les chaînes de Markov, on peut citer Meyn et Tweedie (1993) ou Revuz (1984).

Une chaîne de Markov est un processus stochastique à temps discret[(2)] dont la dépendance entre les éléments est traduite de la façon la plus simple mais cependant très riche.

Définition *Une suite de variables aléatoires $(X_i)_{i\geq 0}$ à valeurs dans un espace d'états mesurable $(E, \mathcal{E})$ est une chaîne de Markov si pour tout $Y \in \sigma(X_i, i \geq n)$*

$$\mathbb{E}(Y|X_0, \ldots, X_n) = \mathbb{E}(Y|X_n).$$

Autrement dit, l'état du processus à l'instant $n+1$ ne dépend que de l'instant n précédent et non de tout le passé (« le futur est indépendant du passé conditionnellement au présent »). On représente généralement une chaîne de Markov par un schéma du type



Parmi les chaînes de Markov, nous allons considérer en particulier des chaînes de Markov homogènes, c'est-à-dire telles que la probabilité de passage d'un état à l'autre ne dépende pas de l'instant n mais seulement des états considérés. Le comportement de la chaîne est alors entièrement défini par son noyau de transition (et la loi initiale).

Définition *Une fonction P de $E \times \mathcal{E}$ dans $[0,1]$ est appelée noyau de transition si*
- *pour tout $A \in \mathcal{E}$, $P(., A)$ est une fonction mesurable sur E,*
- *pour tout $x \in E$, $P(x, .)$ est une mesure de probabilité sur E.*

A partir du noyau de transition, on peut définir ses itérés par la relation de récurrence suivante

$$P^k(x, A) = \int_E P^{k-1}(x, dy) P(y, A).$$

Une chaîne de Markov de noyau de transition P est définie de la façon suivante

(2) Le terme de « chaîne » renvoie selon les auteurs à un processus à temps discret ou à un espace d'états discret. Ici le temps sera discret et l'espace d'état continu et on parlera indifféremment de chaîne de Markov ou de processus de Markov.

Définition *Le processus $(X_i)_{i\geq 0}$ est une chaîne de Markov homogène de loi initiale μ et de noyau de transition $P(.,.)$ si pour tout n et tous ensembles $A_0, A_1, ..., A_n$,*

$$\mathbb{P}(X_0 \in A_0, ..., X_n \in A_n) = \int_{y_0 \in A_0} ... \int_{y_{n-1} \in A_{n-1}} \mu(dy_0) P(y_0, dy_1) ... P(y_{n-1}, A_n).$$

On peut également écrire $P^k(x, A) = \mathbb{P}(X_{n+k} \in A | X_n = x)$ et pour toute fonction g mesurable

$$\mathbb{E}[g(X_{n+1})|X_n] = \int P(X_n, dy) g(y).$$

Dans la théorie des chaînes de Markov, deux cas sont à distinguer : le cas d'un espace E discret, et le cas d'un espace E continu. Si E est discret, la transition est une matrice (finie si E est finie, dénombrable sinon) de coefficients $p_{ij} = P(X_{n+1} = j | X_n = i)$. L'estimation d'une telle quantité est alors paramétrique. C'est par exemple le cas des travaux sur la génomique en biologie, l'ADN étant construit sur un alphabet fini. On peut trouver un aperçu des travaux en statistiques paramétriques des chaînes de Markov dans Clémençon (1999). Dans cette thèse, on considère le cas où E est continu, et plus précisément $E = \mathbb{R}$.

b/ Irréductibilité

Définissons ici la notion d'irréductibilité. Une chaîne est irréductible si tous ses états communiquent. Dans le cas d'une chaîne à états discrets, il est facile de définir la notion de communication entre états et par suite d'irréductibilité. Un état y est accessible depuis x s'il existe un $n \geq 0$ tel que $\mathbb{P}(X_{n+k} = y | X_n = x) = P^n(x, y) > 0$, c'est-à-dire s'il existe un chemin depuis x vers y. On dit alors que deux états x et y communiquent si x est accessible depuis y et réciproquement. Une chaîne est dite irréductible si tous ses états communiquent entre eux. Pour une chaîne à espace d'états continu, on ne peut pas définir l'irréductibilité de la même façon. En effet la probabilité d'atteindre un point fixé de E étant nulle pour une variable à densité, il faut plutôt considérer $P^n(x, A)$ avec A un ensemble non négligeable.

Définition *On dit que X est irréductible s'il existe une mesure φ sur $\mathcal{E}$ telle que*

$$\forall A \in \mathcal{E} \quad \varphi(A) > 0 \Rightarrow \forall x \in E \quad \exists n > 0 \quad P^n(x, A) > 0. \tag{3}$$

Comme dans le cas discret, une chaîne est irréductible si elle ne peut pas être « coupée en morceaux séparés ». Dès qu'un ensemble est assez gros (en terme de φ-mesurabilité), il peut être atteint depuis n'importe quel état x. De plus (théorème 4.0.1 dans Meyn et Tweedie (1993)), si X est irréductible, il existe une mesure d'irréductibilité maximale ψ qui domine toutes les autres et telle que pour tout A de mesure $\psi(A) = 0$, l'ensemble des états permettant d'accéder à A est de mesure nulle (il s'agit en quelque sorte de la réciproque de (3)). On se restreint maintenant à l'étude des chaînes de Markov irréductibles.

c/ Récurrence et stationnarité

La communication entre états est précisée par la notion de récurrence.

Définition *Une chaîne de Markov est dite récurrente si elle est irréductible et si pour tout ensemble A de mesure $\psi(A)$ non nulle et pour tout x de E, $\sum_{n=1}^{\infty} P^n(x, A) = \infty$.*

Une chaîne est récurrente si pour tout A, le nombre moyen de visites en A est infini, quel que soit l'état initial x. D'après le théorème 10.4.4 dans Meyn et Tweedie (1993), toute chaîne de Markov récurrente admet une mesure invariante, c'est-à-dire qui reste inchangée sous l'action de P :

Définition *Une mesure σ-finie μ sur $\mathcal{E}$ est dite invariante si*

$$\forall A \in \mathcal{E} \quad \mu(A) = \int \mu(dx) P(x, A).$$

Si cette mesure invariante est finie, la chaîne est dite récurrente positive, sinon elle est dite récurrente nulle. En conséquence,

Définition *Une chaîne de Markov est dite récurrente positive si elle admet une probabilité invariante.*

Dans ce travail nous étudions des chaînes de Markov récurrentes positives, mais aussi stationnaires.

Définition *Un processus est stationnaire si pour tout k la loi du vecteur $(X_n, \ldots, X_{n+k})$ est la même quel que soit n.*

Il est important de remarquer que si la mesure initiale est la mesure invariante alors la chaîne est stationnaire. En effet, si μ est invariante,

$$\begin{aligned} \mu(A) &= \int \mu(dx) P(x, A) = \iint \mu(dy) P(y, dx) P(x, A) \\ &= \int \mu(dy) P^2(y, A) = \ldots = \int \mu(dx) P^n(x, A) \\ &= \mathbb{P}_\mu(X_n \in A). \end{aligned}$$

C'est pourquoi la probabilité invariante est aussi appelée loi stationnaire du processus.

d/ Ergodicité et mélange

Dans cette thèse, la plupart des démonstrations s'appuient sur des inégalités exponentielles qui contrôlent la déviation de processus empiriques issus de la chaîne (voir lemme 1). Ces démonstrations nécessitent une convergence rapide de la loi de X_n vers la loi stationnaire.

L'ergodicité traduit la convergence des itérés du noyau de transition vers la mesure stationnaire. Dans le chapitre 1, la chaîne est supposée géométriquement ergodique, c'est-à-dire qu'il existe $\rho \in]0,1[$ et une fonction $V > 0$ finie tels que pour tout $n \geq 1$,

$$\forall x \in E \quad \|P^n(x,.) - \mu\|_{TV} \leq V(x)\rho^n \tag{4}$$

où $\|.\|_{TV}$ est la norme en variation totale. Plusieurs travaux antérieurs (Roussas (1969), Prakasa Rao (1978), Gillert et Wartenberg (1984)) concernant l'estimation de la transition ou de la loi stationnaire d'une chaîne de Markov considèrent une chaîne satisfaisant la condition de Doeblin. En réalité, cette condition équivaut à l'ergodicité uniforme $\sup_{x\in E} \|P^n(x,.)-\mu\|_{TV} \to 0$, qui implique l'ergodicité géométrique (théorème 16.0.2 dans Meyn et Tweedie (1993)).

Une autre façon très répandue de traiter les processus dépendants est d'utiliser le concept de mélange. La notion de mélange permet de quantifier la dépendance des variables aléatoires d'un processus par des coefficients appelés coefficients de mélange. Un processus X est mélangeant si la dépendance entre X_n et X_{n+k} tend vers 0 quand k tend vers l'infini. De très nombreux types de mélanges existent : α-mélange, β-mélange, ρ-mélange, ϕ-mélange, etc.. On se réfère à Doukhan (1994) pour la définition et l'utilisation de ces différents mélanges.

Pour estimer la densité de la loi stationnaire, Rosenblatt (1970) et Basu et Sahoo (1998) emploient une hypothèse notée (G_2) équivalente à du ρ-mélange. Un processus markovien mélangeant a d'ailleurs la propriété d'être automatiquement géométriquement mélangeant, c'est-à-dire que si les coefficients de mélange tendent vers 0, ils le font à vitesse géométrique.

Bosq (1973) impose une condition de ϕ-mélange. Dans le cas d'un processus de Markov stationnaire, les coefficients de ϕ-mélange vérifient

$$\frac{1}{2}\phi_n \leq \sup \operatorname{ess}_\mu \|P^n(x,.) - \mu\|_{TV} \leq \phi_n$$

Le ϕ-mélange correspond donc à l'ergodicité uniforme. Un processus ϕ-mélangeant est ainsi géométriquement ϕ-mélangeant, c'est-à-dire géométriquement ergodique. C'est la condition requise au chapitre 1, mais aussi dans Clémençon (1999), Doukhan et Ghindès (1983).

Dans ce travail, à partir du chapitre 2, nous avons choisi d'utiliser le β-mélange. En effet, celui-ci ne permet pas seulement d'avoir des inégalité de covariance (voir le lemme 4.6 section 4.7.8), il fournit aussi des variables d'approximation qui permettent de se ramener au cas indépendant. Suivant les travaux de Viennet (1996), on construit des variables aléatoires X_i^* qui sont indépendantes par blocs et ont même loi que les X_i. La différence entre X_i et X_i^* est contrôlée par les coefficients de β-mélange. De nombreux travaux statistiques utilisent le β-mélange pour gérer la dépendance, on peut citer entre autres Baraud *et al.* (2001), Comte et Genon-Catalot (2006), Tribouley et Viennet (1998). Dans

le cas d'une chaîne de Markov stationnaire, les coefficients de β-mélange ont une expression particulièrement simple

$$\beta_n = \int \|P^n(x,.) - \mu\|_{TV}\mu(dx).$$

Nummelin et Tuominen (1982) ont démontré que si la chaîne est géométriquement ergodique alors les quantités ρ et V dans (4) peuvent être choisies de telle sorte que V soit μ-intégrable. La chaîne est alors géométriquement β-mélangeante. On peut retrouver ce résultat en remarquant que l'on a toujours $\beta_n \leq \phi_n$ et donc le ϕ-mélange implique le β-mélange. Ainsi la condition de β-mélange est plus faible que l'ergodicité géométrique.

e/ Hypothèses

Dans cette étude, on suppose que le noyau de transition de la chaîne de Markov considérée admet une densité Π par rapport à la mesure de Lebesgue. On a alors

$$\forall x \in E \quad \forall B \in \mathcal{E} \quad P(x,B) = \int_B \Pi(x,y)dy.$$

On suppose également que la loi stationnaire admet une densité f par rapport à la mesure de Lebesgue. Les hypothèses requises dans ce travail sont les suivantes :

H1 La chaîne de Markov $(X_i)_{i\geq 0}$ est irréductible, récurrente positive et stationnaire.

H2 La chaîne de Markov est ergodique. Plus précisément, on utilisera l'une ou l'autre des hypothèses suivantes :

H2a La chaîne de Markov est géométriquement ergodique et fortement apériodique.

H2b La chaîne de Markov est géométriquement ou arithmétiquement β-mélangeante.

Dans le cadre de l'ergodicité géométrique, l'hypothèse supplémentaire d'apériodicité forte est nécessaire pour utiliser la technique de scission de Nummelin (détaillée au chapitre 1). L'apériodicité forte signifie qu'il existe une fonction $h : E \mapsto [0,1]$ avec $\int h d\mu > 0$ et une distribution strictement positive ν telle que, pour tout évènement B de $\mathcal{E}$ et pour tout x de E,

$$P(x,B) \geq h(x)\nu(B).$$

Pour l'établissement des théorèmes limites pour les chaînes de Markov, on peut s'affranchir de cette condition de minoration car il suffit qu'elle soit vérifiée pour un itéré P^k, ce qui est le cas de toutes les chaînes de Markov irréductibles. Mais ce n'est pas le cas ici.

On estime la densité stationnaire sur $\mathbb{R}$ ou sur un intervalle compact A_1 de $\mathbb{R}$, et la densité de transition sur un pavé compact $A = A_1 \times A_2$ de $\mathbb{R}^2$. On supposera en général que ces fonctions sont bornées et de carré intégrable sur ces compacts.

H3 La densité stationnaire et la densité de transition vérifient l'une ou l'autre des conditions suivantes :

H3a $f \in L^\infty(A_1)$

H3b $\Pi \in L^\infty(A)$

H3c $f \in (L^2 \cap L^\infty)(\mathbb{R})$

H3d $F = f\Pi \in (L^2 \cap L^\infty)(\mathbb{R}^2)$

H3e $\Pi \in L^2(A)$

Il est à noter que, si f (resp. Π) appartient à un espace de Besov de régularité $\alpha > 1/2$ (resp. $\alpha > 1$), alors elle est continue. Dans ce cas, l'hypothèse H3a (resp. H3b) est automatiquement satisfaite.

Une dernière hypothèse est nécessaire pour mettre en oeuvre l'estimation de la densité de transition Π (mais inutile pour l'estimation de f), que ce soit par une méthode quotient (chapitres 1 et 4) ou par une méthode de régression (chapitres 2 et 5) :

H4 Il existe un réel f_0 strictement positif tel que $\forall x \in A_1 \quad f(x) \geq f_0$.

Cette hypothèse, bien que restrictive (c'est essentiellement pour cette raison que l'estimation de Π n'est assurée que sur un compact), est cruciale ici. Il s'agit cependant d'une condition classique dans un cadre de régression.

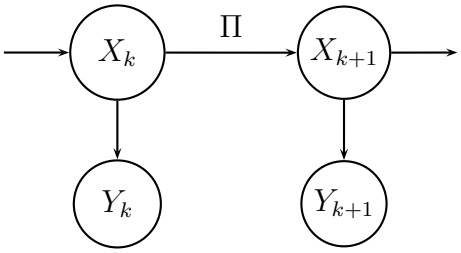
De nombreuses chaînes de Markov vérifient ces hypothèses, des exemples sont détaillés au chapitre 1 section 1.2.2.

f/ Chaîne de Markov cachée

Dans la partie B, nous traitons le cas des chaînes de Markov cachées.

Définition *Un processus bivarié $(X_i, Y_i)_{i\geq 0}$ est une chaîne de Markov cachée si $(X_i)_{i\geq 0}$ est une chaîne de Markov et, conditionnellement à $(X_i)_{i\geq 0}$, $(Y_i)_{i\geq 0}$ est une suite de variables aléatoires indépendantes et telle que pour tout n la loi conditionnelle de Y_n ne dépende que de X_n.*

Une chaîne de Markov cachée est traditionnellement représentée de la façon suivante



Du fait de ses multiples applications, en biologie, théorie de la communication, reconnaissance de la parole ou encore en finance, ce modèle a été amplement étudié. De nombreuses études se placent dans le cadre d'un espace d'états fini et supposent que la

distribution des Y_i appartient à une famille paramétrique. Nous ne développerons pas ici l'importance de ce modèle et pour plus de détails, nous renvoyons à Cappé *et al.* (2005). En effet, on ne considère dans cette thèse que les chaînes de Markov cachées par bruit additif. Dans ce cas, la chaîne de Markov cachée est décrite par

$$\forall i \geq 0 \quad Y_i = X_i + \varepsilon_i$$

où (X_i) est une chaîne de Markov et (ε_i) est une suite de variables aléatoires indépendante de (X_i). Ce modèle est une classe importante des modèles de Markov cachés car il rend compte de tous les phénomènes altérés par une erreur de mesure. Par passage au logarithme, il permet également de traiter des cas multiplicatifs et même des modèles à volatilité stochastique (voir les exemples donnés au chapitre 4 section 4.5).

Résumé de la thèse

Chapitre 1 [3]

Dans le premier chapitre, on cherche à estimer la densité de transition Π d'une chaîne de Markov (X_i) à partir des observations $X_1, \ldots, X_n$. Pour cela, on utilise la remarque suivante :

$$\Pi(x, y) = \frac{F(x, y)}{f(x)}$$

où f est la densité stationnaire de X_i et F la densité du couple (X_i, X_{i+1}). Si l'on parvient à estimer f et F par des estimateurs $\tilde{f}$ et $\tilde{F}$, un estimateur naturel de Π sera

$$\tilde{\Pi}(x, y) = \frac{\tilde{F}(x, y)}{\tilde{f}(x)}. \tag{5}$$

On procède donc dans ce chapitre en trois étapes : estimation de f, estimation de la densité jointe F et enfin estimation de la transition Π.

- *Estimation de f* : il s'agit d'estimer la densité de variables dépendantes. On procède par minimisation de contraste et sélection de modèle comme expliqué ci-dessus. Tribouley et Viennet (1998) ont traité le cas de variables β-mélangeantes. Dans ce cas, on observe l'apparition dans la pénalité d'un terme de mélange, plus précisément de la somme $\sum_k \beta_k$ des coefficients de β-mélange. Mais ce terme n'est *a priori* pas connu. Le but initial était, dans le cas d'une chaîne de Markov, de remplacer ce terme inconnu par une quantité connue dans la pénalité. Pour cela nous avons utilisé des méthodes propres aux chaînes de Markov : décomposition par temps d'entrée dans un atome pour se ramener à des blocs

[3] Ce chapitre est une version modifiée de l'article *Nonparametric estimation of the stationary density and the transition density of a Markov chain* accepté pour publication à Stochastic Processes and their Applications.

de variables i.i.d. On obtient alors dans la pénalité un terme différent, mais également inconnu : il s'écrit $\mathbb{E}(e^\tau)$ où τ est un temps d'arrêt théorique introduit par la méthode de scission de Nummelin utilisée. Bien qu'en pratique ce terme ne soit nullement gênant, il justifie notre choix d'utiliser des méthodes de mélange dans les chapitres suivants. En effet, ces méthodes permettent également de se ramener à des blocs de variables aléatoires indépendants, mais cette fois de taille fixée et non aléatoire, ce qui est plus simple à traiter.

On obtient dans la section 1.3 un très bon estimateur de f, puisqu'il est adaptatif et atteint la vitesse de convergence optimale $n^{-\frac{2\alpha}{2\alpha+1}}$ (théorème 1.1). C'est donc une amélioration des travaux précédents sur la question (Clémençon (1999) avait obtenu un estimateur adaptatif par seuillage d'ondelettes mais avec une perte logarithmique dans la vitesse). De plus, l'estimateur obtenu est implémentable et donne des résultats très satisfaisants (voir section 1.5) compte tenu de la dépendance des observations.

- *Estimation de* F : l'estimation de la densité jointe se fait de la même façon que celle de f. Le contraste est modifié en

$$\Gamma_n(T) = \|T\|_2^2 - \frac{2}{n-1}\sum_{i=1}^{n-1} T(X_i, X_{i+1}).$$

La sélection de modèles s'effectue sur les produits tensoriels $\mathbb{S}_m$ des modèles de dimension 1 utilisés pour l'estimation de f. En réalité, il n'y a aucun changement conceptuel au passage à la dimension 2. On obtient donc (théorème 1.2) le même type de résultat qu'en dimension 1. Si la pénalité est supérieure à une certaine constante multipliée par D_m^2/n, l'estimateur $\tilde{F}$ atteint la vitesse de convergence optimale.

- *Estimation de* Π : la formule (5) doit être légèrement remaniée car $\tilde{f}$ peut s'annuler. Pour éviter une explosion de l'estimateur $\tilde{\Pi}$, on définit en réalité

$$\tilde{\Pi}(x,y) = \begin{cases} \dfrac{\tilde{F}(x,y)}{\tilde{f}(x)} & \text{si } |\tilde{F}(x,y)| \leq n|\tilde{f}(x)| \\ 0 & \text{sinon .} \end{cases}$$

En restreignant légèrement le nombre de modèles utilisés, on obtient (théorème 1.3)

$$\mathbb{E}\|\Pi - \tilde{\Pi}\|_2^2 \leq C_1 \mathbb{E}\|F - \tilde{F}\|_2^2 + C_2 \mathbb{E}\|f - \tilde{f}\|_2^2 + o\left(\frac{1}{n}\right),$$

ce qui permet d'obtenir une vitesse de convergence en fonction de la régularité de f et F. Si la chaîne vit sur A_1 et que Π est de régularité α, alors f et F sont de régularité α également. On obtient dans ce cas la vitesse de convergence optimale $n^{-\frac{2\alpha}{2\alpha+2}}$ pour l'estimation de Π. Si le support de Π n'est pas inclus dans A, f et Π n'ont plus nécessairement la même régularité. L'obtention de la vitesse de convergence optimale dans ce cas fait l'objet du chapitre 2.

Puisque notre procédure est entièrement implémentable, on présente en section 1.5 des simulations pour divers processus de Markov. Ce chapitre se termine (section 1.6) par les démonstrations des résultats énoncés.

Chapitre 2 [4]

Le deuxième chapitre est consacré à l'estimation de la densité de transition d'une chaîne de Markov mais par une méthode différente de celle employée au chapitre 1. En effet, la méthode quotient comporte des inconvénients. L'erreur d'estimation est la somme de deux erreurs, celle faite en estimant la densité invariante f, et celle due à l'estimation de la densité jointe F. Il est alors naturel d'essayer d'estimer Π de façon plus directe pour ne pas cumuler les erreurs. On peut exprimer cet inconvénient en terme de vitesse de convergence. Dans la méthode quotient, cette vitesse dépend de la régularité de la fonction f. Or cette régularité peut être très faible ce qui entraîne une vitesse de convergence trop lente. En effet, la densité de transition peut être plus régulière que f et il est alors intéressant de disposer d'un estimateur convergeant à une vitesse ne dépendant pas de f. On pourrait penser que la densité stationnaire, en tant que limite des itérés de Π, est nécessairement au moins aussi régulière que Π mais, cette convergence étant de type L^1, ce n'est pas le cas, tout du moins localement. Plus précisément, si Π appartient à l'espace de Besov $B^{\alpha}_{2,\infty}(\mathbb{R}^2)$ alors f appartient à $B^{\alpha}_{2,\infty}(\mathbb{R})$, mais si $\Pi \in B^{\alpha}_{2,\infty}(A)$, on peut avoir f de régularité plus faible sur A_1. Stéphan Clémençon a ainsi exhibé dans sa thèse un exemple de chaîne de Markov de densité de transition constante sur $[0,1]$ et de densité stationnaire discontinue ! Une méthode plus directe peut également permettre de s'affranchir du caractère asymptotique du théorème 1.3.

Dans ce chapitre 2, nous avons donc construit un autre estimateur de Π. Cet estimateur est obtenu par minimisation d'un nouveau contraste

$$\gamma_n(T) = \frac{1}{n}\sum_{i=1}^{n} \int T^2(X_i, y)dy - 2T(X_i, X_{i+1}) \tag{6}$$

qui utilise l'aspect régressif du problème. En minimisant ce contraste sur différents espaces de projection, on construit une collection d'estimateurs $\hat{\Pi}_m$. On procède alors par sélection de modèles comme au chapitre 1. On minimise le critère pénalisé suivant

$$\gamma_n(\hat{\Pi}_m) + \mathrm{pen}(m)$$

pour obtenir le modèle $\hat{m}$, l'estimateur final étant $\tilde{\Pi} = \hat{\Pi}_{\hat{m}}$.

On souhaite estimer au mieux des transitions anisotropes, c'est-à-dire ayant des régularités différentes dans les deux directions du plan. Pour ce faire, on introduit des modèles anisotropes, produits tensoriels d'un modèle de dimension D_{m_1} et d'un modèle de dimension D_{m_2}. L'estimateur construit est adaptatif et converge à la vitesse $n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}}$ où $\bar{\alpha}$ est la moyenne harmonique des deux régularités α_1 et α_2 de Π.

L'erreur d'estimation est d'abord évaluée en une norme empirique apparaissant naturellement dans le problème. Dans la section 2.5, on passe à la norme L^2. Pour cela, on

[4] Ce chapitre est une version modifiée de l'article *Adaptive estimation of the transition density of a Markov chain* à paraître aux Annales de l'Institut Henri Poincaré, Probabilités et Statistiques.

tronque l'estimateur si sa norme L^2 dépasse un certain seuil dépendant de n. On obtient pour la norme L^2 le même résultat que pour la norme empirique.

On énonce dans ce chapitre un résultat de borne inférieure qui prouve que la vitesse $n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}}$ est optimale. Il nous a également paru intéressant d'explorer le cas de la borne inférieure en norme L^p pour $1 \leq p < \infty$ qui met en évidence un phénomène de coude dans la vitesse d'estimation. Ce résultat est prouvé en détail en annexe, section 2.8.

Des simulations pour ce nouvel estimateur sont exposées et comparées avec celles du chapitre 1. Il s'avère que les résultats obtenus avec cet estimateur ne sont pas seulement meilleurs du point de vue théorique mais aussi du point de vue pratique, ce qui est en particulier dû à l'utilisation de l'anisotropie.

Chapitre 3 [5]

Avant de se consacrer au problème plus complexe de l'estimation de la transition d'une chaîne de Markov cachée, nous passons en revue au chapitre 3 les vitesses de convergence que l'on peut obtenir dans le cas d'une estimation avec des observations bruitées. Le modèle étudié dans ce chapitre est celui dit de convolution. On observe des données $Y_1, \ldots, Y_n$ où pour chaque $i \in \{1, \ldots, n\}$, $Y_i = X_i + \varepsilon_i$ avec (X_i) et (ε_i) des suites indépendantes de variables i.i.d. Le signal (X_i) est observé à travers un bruit (ε_i) de loi entièrement connue. On cherche à estimer la densité g de X_i à partir des observations $Y_1, \ldots, Y_n$. On suppose que la densité q du bruit (ε_i) vérifie l'hypothèse suivante

H5 Il existe $s \geq 0, b \geq 0, \gamma \in \mathbb{R}$ ($\gamma > 0$ si $s = 0$) et $k_0, k_1 > 0$ tels que

$$k_0(x^2+1)^{-\gamma/2}\exp(-b|x|^s) \leq |q^*(x)| \leq k_1(x^2+1)^{-\gamma/2}\exp(-b|x|^s)$$

où $q^*(x) = \int e^{-itx}q(t)dt$ est la transformée de Fourier de q.

Le bruit est dit ordinairement régulier si $s = 0$ et super régulier sinon. De même, on considère deux types de régularité pour la fonction g à estimer. On suppose en effet que celle-ci appartient à l'espace suivant

$$\mathcal{A}_{\delta,r,a}(l) = \left\{ f \text{ densité sur } \mathbb{R} \text{ et } \int |f^*(x)|^2(x^2+1)^{\delta}\exp(2a|x|^r)dx \leq l \right\}$$

avec $r = 0$ (régularité ordinaire) ou bien $r > 0$ (super régularité). On a alors quatre cas distincts, selon la régularité (inconnue) de g et la régularité (connue) de q.

Dans la section 3.2, on présente les estimateurs classiques utilisés pour ce modèle. Les performances de l'estimateur à noyau établies par Butucea et Tsybakov (2006) sont données dans la proposition 3.1. La proposition 3.2 rappelle quant à elle les résultats obtenus pour l'estimateur par projection de Comte *et al.* (2006b). On récapitule alors dans

[5] Ce chapitre est une version modifiée de l'article *Rates of convergence for nonparametric deconvolution* paru aux Comptes Rendus de l'Académie des Sciences, vol. 342.

la section suivante les vitesses de convergence optimales pour le risque intégré comme pour le risque ponctuel. Parmi les quatre cas évoqués précédemment, trois sont bien connus. Le quatrième cas (fonctions g et q super régulières), pour lequel les vitesses de convergence explicites n'avaient jamais été données jusqu'à présent, fait l'objet du théorème 3.1. On y obtient des vitesses de convergence peu courantes et tout à fait intéressantes. Ce théorème est démontré dans la dernière section.

Chapitre 4 [6]

Le quatrième chapitre est consacré à l'estimation de la transition d'une chaîne de Markov lorsque celle-ci est observée avec un bruit additif. Il s'agit du modèle de Markov caché suivant :

$$Y_i = X_i + \varepsilon_i \qquad i = 1, \ldots, n+1 \tag{7}$$

où $(X_i)_{i\geq 1}$ est la chaîne de Markov et $(\varepsilon_i)_{i\geq 1}$ un bruit indépendant de $(X_i)_{i\geq 1}$. On suppose que les variables aléatoires $\varepsilon_1, \ldots, \varepsilon_n$ sont indépendantes et identiquement distribuées de densité connue q. Cette forme du modèle rapproche ce chapitre des problématiques de déconvolution, on y trouve en particulier les mêmes types de vitesse de convergence.

Comme dans le chapitre 1, nous avons choisi d'estimer la densité de transition par le quotient (5) d'un estimateur de F et d'un estimateur de f. L'estimation de f relève typiquement du modèle de convolution pour des variables dépendantes. En effet la densité f_Y des observations Y_i est égale au produit de convolution $f * q$. Pour estimer la densité stationnaire, on observe que pour toute fonction t,

$$\mathbb{E}[t(X_i)] = \int t(x)f(x)dx = \frac{1}{2\pi}\int t^*(x)\overline{f^*(x)}dx = \frac{1}{2\pi}\int \frac{t^*(x)}{\overline{q^*(x)}}\overline{f^*(x)q^*(x)}dx$$

où t^* désigne la transformée de Fourier de t. Ainsi, en posant v_t la transformée de Fourier inverse de $t^*(x)/\overline{q^*(x)}$, on peut écrire

$$\mathbb{E}[t(X_i)] = \frac{1}{2\pi}\int v_t^*(x)\overline{f_Y^*(x)}dx = \int v_t(x)f_Y(x)dx = \mathbb{E}[v_t(Y_i)].$$

La procédure d'estimation est largement fondée sur le calcul précédent. Dans le cas où les X_i étaient directement observées, on avait estimé f à l'aide du contraste $(1/n)\sum_{i=1}^n[\|t\|_2^2 - 2t(X_i)]$. Maintenant que seules les données bruitées Y_i sont observées, la remarque précédente nous amène à considérer le contraste

$$\gamma_n(t) = (1/n)\sum_{i=1}^n[\|t\|_2^2 - 2v_t(Y_i)]$$

[6] Ce chapitre est une version modifiée de l'article *Adaptive estimation of the transition density of a particular hidden Markov chain* accepté pour publication à Journal of Multivariate Analysis.

déjà utilisé par Comte *et al.* (2006b). Une collection d'estimateurs de f est alors calculée par minimisation de ce contraste sur les espaces

$$S_m = \text{Vect}\{x \mapsto \sqrt{m}\varphi(mx - j)\}_{j\in\mathbb{Z}} \text{ avec } \varphi(x) = \frac{\sin(x)}{x}.$$

Ces espaces ont des propriétés très intéressantes car $\varphi^* = \mathbb{1}_{[-\pi,\pi]}$, ce qui permet de localiser en fréquence. On estime finalement f par $\hat{f}_{\hat{m}}$ où $\hat{m}$ est le modèle qui minimise un certain critère pénalisé.

Dans un deuxième temps, il faut estimer F, la densité du couple (X_i, X_{i+1}). On procède de manière similaire. L'opérateur v_t est remplacé par V_t transformée de Fourier inverse de $T^*(x,y)/\overline{q^*(x)q^*(y)}$. Il vérifie la propriété $\mathbb{E}[t(X_i, X_{i+1})] = \mathbb{E}[V_t(Y_i, Y_{i+1})]$ pour toute fonction T. On en déduit le contraste

$$\Gamma_n(t) = (1/n)\sum_{i=1}^{n}[\|t\|_2^2 - 2V_t(Y_i, Y_{i+1})]$$

et l'estimateur $\tilde{F}$.

L'intérêt principal de ce chapitre réside dans les vitesses obtenues. Ces vitesses dépendent conjointement de la régularité de la transition estimée et de celle du bruit. Le bruit peut être ordinairement régulier (décroissance polynomiale de q^*) ou super régulier (décroissance exponentielle de q^*). On considère que la densité stationnaire appartient à l'espace $\mathcal{A}_{\delta,r,a}(l)$ évoqué au chapitre 3. La fonction f est dite super régulière si $r > 0$ et ordinairement régulière si $r = 0$. On peut alors considérer quatre cas, selon que le bruit est ordinairement ou super régulier et que la fonction estimée est ordinairement ou super régulière. Si la densité du bruit et f sont toutes deux ordinairement régulières, la vitesse de convergence est polynomiale en n. Si la densité f reste ordinairement régulière mais que l'erreur ε_i devient super régulière, la vitesse de convergence devient logarithmique. Cela s'explique par le fort lissage effectué par le bruit, qui rend très délicate la reconstruction du signal. Une déconvolution efficace est possible dans ce cas de bruit très régulier si la fonction à estimer est elle aussi super régulière. On obtient alors les vitesses singulières développées au chapitre 3. Si la fonction est super régulière et le bruit est ordinairement régulier, la déconvolution peut être réalisée facilement et la vitesse de convergence est très proche de la vitesse paramétrique $1/n$.

Le même phénomène se reproduit pour la densité jointe F. Que ce soit pour l'estimation de f ou celle de F, notre procédure permet d'atteindre la vitesse optimale de façon adaptative dans presque tous les cas.

Un estimateur de la densité de transition est finalement obtenu par le quotient

$$\tilde{\Pi}(x,y) = \begin{cases} \dfrac{\tilde{F}(x,y)}{\tilde{f}(x)} & \text{si } |\tilde{F}(x,y)| \le n|\tilde{f}(x)| \\ 0 & \text{sinon.} \end{cases}$$

Son risque est majoré par la somme des erreurs d'estimations $\|f - \tilde{f}\|_2^2$ et $\|F - \tilde{F}\|_2^2$. Des exemples de processus sont exposés en section 4.5 pour illustrer les vitesses de convergence variées qui peuvent apparaître dans ce problème.

Chapitre 5

Etant donné les inconvénients de la procédure quotient expliqués précédement, il était naturel de chercher une procédure semblable à celle du chapitre 2, mais dans le cadre d'une chaîne de Markov cachée. C'est l'objet de ce dernier chapitre. On modifie le contraste (6) en utilisant la transformation de Fourier pour l'adapter au cadre (7). Il s'écrit sous la forme

$$\gamma_n(T) = \frac{1}{n}\sum_{k=1}^{n}[Q_{T^2}(Y_k) - 2V_T(Y_k, Y_{k+1})]$$

où V_T est l'opérateur déjà utilisé au chapitre 4 et Q_T est la transformée de Fourier inverse de $T^*(.,0)/\overline{q^*}$. La construction précise du contraste $\gamma_n(T)$ est expliquée en section 5.3.2.

Le choix des modèles est délicat. L'estimation directe de la densité de transition requiert l'utilisation de bases à support compact à cause de l'hypothèse H4. De plus, les fonctions de bases doivent être plus régulières que la densité du bruit. Ces contraintes nous ont amené à choisir des bases d'ondelettes à support compact, décrites en section 5.3.1, et à se restreindre au cas d'un bruit ordinairement régulier.

On souhaite alors définir une collection d'estimateurs de Π par minimisation du contraste. Cependant la minimisation de $\gamma_n(T)$ pour T appartenant à un espace $\mathbb{S}_m$ équivaut à l'équation matricielle $G_m A_m = Z_m$ où A_m est le vecteur des coefficients d'un minimiseur dans la base $(\omega_\lambda)_\lambda$ de $\mathbb{S}_m$ et

$$G_m = \left[\frac{1}{n}\sum_{i=1}^{n} Q_{\omega_\lambda\omega_\mu}(Y_i)\right]_{\lambda,\mu}, \quad Z_m = \left[\frac{1}{n}\sum_{i=1}^{n} V_{\omega_\lambda}(Y_i, Y_{i+1})\right]_{\lambda}.$$

Comme la matrice G_m n'est pas nécessairement inversible, on se place sur l'ensemble

$$\Gamma = \left\{\min \mathrm{Sp}(G_m) \geq \frac{2}{3} f_0\right\}$$

où Sp désigne l'ensemble des valeurs propres. Pour chaque modèle $\mathbb{S}_m$, l'estimateur est alors défini par

$$\hat{\Pi}_m = \arg\min_{T\in\mathbb{S}_m} \gamma_n(T)\mathbb{1}_\Gamma.$$

On sélectionne ensuite le modèle

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n}\left\{\gamma_n(\hat{\Pi}_m) + K\frac{D_m^{4\gamma+2}}{n}\right\}$$

où K est une constante et γ désigne la régularité du bruit. Comme au chapitre 2, l'estimateur $\hat{\Pi}_{\hat{m}}$ est tronqué pour pouvoir contrôler la norme L^2.

L'estimateur $\tilde{\Pi}$ obtenu atteint la vitesse de convergence $n^{-\frac{2\alpha}{2\alpha+4\gamma+2}}$ lorsque la densité de transition appartient à l'espace de Besov $B^{\alpha}_{2,\infty}(A)$. Cet estimateur est adaptatif et sa vitesse de convergence ne dépend pas de la régularité de f. Un schéma de la preuve est donné en section 5.4.2, suivi d'une démonstration détaillée.

Notations

Dans toute la thèse, f désigne la densité de la loi stationnaire μ de la chaîne de Markov considérée et Π la densité de transition. La lettre F renvoie à la densité du couple (X_i, X_{i+1}). On note q la densité du bruit additionnel le cas échéant (chapitres 4 et 5). Le nombre d'observations sera toujours n ou $n+1$. Le compact d'estimation est noté $A = A_1 \times A_2$.

Par souci de clarté, on utilise en général des lettres minuscules pour la dimension 1 et des majuscules pour la dimension 2. Pour une fonction $t : \mathbb{R} \mapsto \mathbb{R}$, on note $\|t\|$ sa norme L^2 : $\|t\|^2 = \int_{\mathbb{R}} t^2(x)dx$ et

$$L^2(\mathbb{R}) = \{t : \mathbb{R} \mapsto \mathbb{R}, \|t\| < \infty\}.$$

Si S est un sous-espace de $L^2(\mathbb{R})$ et $g \in L^2(\mathbb{R})$ une fonction, on note

$$d(g, S) = \inf_{t \in S} \|g - t\|.$$

Pour un compact A_1 de $\mathbb{R}$, on définit

$$L^2(A_1) = \left\{ t : \mathbb{R} \mapsto \mathbb{R}, \|t\|_{A_1} = \left(\int_{A_1} |t(x)|^2 dx \right)^{1/2} < \infty \right\}$$

et

$$L^{\infty}(A_1) = \{t : \mathbb{R} \mapsto \mathbb{R}, \|t\|_{\infty, A_1} = \sup_{x \in A_1} |t(x)| < \infty\}.$$

La notation $\mathcal{D}(A_1)$ désigne l'ensemble des fonctions à support compact inclus dans A_1.

De la même manière, pour une fonction $T : \mathbb{R}^2 \mapsto \mathbb{R}$, $\|T\|^2 = \iint_{\mathbb{R}^2} T^2(x, y)dxdy$ et

$$L^2(\mathbb{R}^2) = \{T : \mathbb{R}^2 \mapsto \mathbb{R}, \|T\| < \infty\}.$$

Pour un compact A de $\mathbb{R}^2$, on définit

$$L^2(A) = \left\{ T : \mathbb{R}^2 \mapsto \mathbb{R}, \|T\|_A = \left(\iint_A |T(x, y)|^2 dxdy \right)^{1/2} < \infty \right\}$$

et

$$L^\infty(A) = \{T : \mathbb{R}^2 \mapsto \mathbb{R}, \|T\|_{\infty,A} = \sup_{(x,y)\in A} |T(x,y)| < \infty\}.$$

On note $t \otimes s$ la fonction : $(x,y) \mapsto (t \otimes s)(x,y) = t(x)s(y)$ et $\mathbb{1}_B$ la fonction indicatrice d'un ensemble B

$$\mathbb{1}_B(x) = \begin{cases} 1 \text{ si } x \in B \\ 0 \text{ sinon.} \end{cases}$$

Enfin, la notation $[x]_+$ désigne $\max(x, 0)$, la partie positive de x.

Première partie

Estimation pour les chaînes de Markov

Chapitre 1

Estimation de la densité de transition par quotient

Version modifiée de l'article *Nonparametric estimation of the stationary density and the transition density of a Markov chain* accepté pour publication à Stochastic Processes and their Applications.

1.1 Introduction

Nonparametric estimation is now a very rich branch of statistical theory. The case of i.i.d. observations is the most detailed but many authors are also interested in the case of Markov processes. Early results are stated by Roussas (1969), who studies nonparametric estimators of the stationary density and the transition density of a Markov chain. He considers kernel estimators and assumes that the chain satisfies the strong Doeblin's condition (D_0) (see Doob (1953) p.221). He shows consistency and asymptotic normality of his estimator. Several authors tried to consider weaker assumptions than the Doeblin's condition. Rosenblatt (1970) introduces another condition, denoted by (G_2), and he gives results on the bias and the variance of the kernel estimator of the invariant density in this weaker framework. Yakowitz (1989) improves also the result of asymptotic normality by considering a Harris-condition. The study of kernel estimators is completed by Masry and Györfi (1987) who find sharp rates for this kind of estimators of the stationary density and by Basu and Sahoo (1998) who prove a Berry-Esseen inequality under the condition (G_2) of Rosenblatt. Other authors are interested in the estimation of the invariant distribution and the transition density in the non-stationary case: Doukhan and Ghindès (1983) bound the integrated risks for any initial distribution. In Hernández-Lerma *et al.* (1988), recursive estimators for a non-stationary Markov chain are described. Liebscher (1992) gives results for the invariant density in this non-stationary framework using a condition denoted by (D_1) derived from the Doeblin's condition but weaker than (D_0). All the above papers deal with kernel estimators. Among those who are not interested in such estimators, let us mention Bosq (1973) who studies an estimator of the stationary density by projection on a Fourier basis, Prakasa Rao (1978) who outlines a new estimator for the stationary density by using delta-sequences and Gillert and Wartenberg (1984) who present estimators based on Hermite bases or trigonometric bases.

The recent work of Clémençon (1999) allows to measure the performance of all these estimators since he proves lower bounds for the minimax rates and thus gives the optimal convergence rates for the estimation of the stationary density and the transition density. Clémençon also provides another kind of estimator for the stationary density and for the transition density, that he obtains by projection on wavelet bases. He presents an adaptive procedure which is "quasi-optimal" in the sense that the procedure reaches almost the optimal rate but with a logarithmic loss. He needs other conditions than those we cited above and in particular a minoration condition derived from Nummelin's (1984) works. In this chapter, we will use the same condition.

The aim of this chapter is to estimate the stationary density of a discrete-time Markov chain and its transition density. We consider an irreducible positive recurrent Markov chain (X_n) with a stationary density denoted by f. We suppose that the initial density is f (hence the process is stationary) and we construct an estimator $\tilde{f}$ from the data $X_1, \ldots, X_n$. Then, we study the mean integrated squared error $\mathbb{E}\|\tilde{f} - f\|^2$ and its convergence rate. The same technique enables to estimate the density F of (X_i, X_{i+1}) and so to

provide an estimator of the transition density $\Pi = F/f$, called the quotient estimator.

An adaptive procedure is proposed for the two estimations and it is proved that both resulting estimators reach the optimal minimax rates without additive logarithmic factor.

We will use here some technical methods known as the Nummelin splitting technique (see Nummelin (1984), Meyn and Tweedie (1993) or Höpfner and Löcherbach (2003)). This method allows to reduce the general state space Markov chain theory to the countable space theory. Actually, the splitting of the original chain creates an artificial accessible atom and we will use the hitting times to this atom to decompose the chain, as we would have done for a countable space chain.

To build our estimator of f, we use model selection via penalization as described in Barron *et al.* (1999). First, estimators by projection denoted by $\hat{f}_m$ are considered. The index m denotes the model, i.e. the subspace to which the estimator belongs. Then the model selection technique allows to select automatically an estimator $\hat{f}_{\hat{m}}$ from the collection of estimators $(\hat{f}_m)$. The estimator of F is built in the same way. The collections of models that we consider here include wavelets but also trigonometric polynomials and piecewise polynomials.

This chapter is organized as follows. In Section 1.2, we present our assumptions on the Markov chain and on the collections of models. We also give examples of chains and models. Section 1.3 is devoted to estimation of the stationary density and in Section 1.4 the estimation of the transition density is explained. Some simulations are presented in Section 1.5. The proofs are gathered in the last section, which contains also a presentation of the Nummelin splitting technique.

1.2 The framework

1.2.1 Assumptions on the Markov chain

We consider an irreducible Markov chain (X_n) taking its values in the real line $\mathbb{R}$. We suppose that (X_n) is positive recurrent, i.e. it admits a stationary probability measure μ. We assume that the distribution μ has a density f with respect to the Lebesgue measure and it is this quantity that we want to estimate. The function f is estimated on a compact set A_1 only. More precisely, the Markov process is supposed to satisfy the following assumptions:

H1 The chain (X_n) is irreducible, positive recurrent and stationary.

H2a (i) The chain is strongly aperiodic, i.e. it satisfies the following minorization condition: there is some function $h : \mathbb{R} \mapsto [0, 1]$ with $\int h d\mu > 0$ and a positive distribution ν such that, for all event B and for all x,

$$P(x, B) \geq h(x)\nu(B)$$

where P is the transition kernel of (X_n).

(ii) The chain is geometrically ergodic, i.e. there exists a function $V > 0$ finite and a constant $\rho \in (0,1)$ such that, for all $n \geq 1$

$$\|P^n(x,.) - \mu\|_{TV} \leq V(x)\rho^n$$

where $\|.\|_{TV}$ is the total variation norm.

H3a The stationary density f belongs to $L^\infty(A_1)$ i.e. $\sup_{x\in A_1} |f(x)| < \infty$.

We can remark that condition H3a implies that f belongs to $L^2(A_1)$.

Notice that, since the chain is irreducible, condition H2a(i) holds for some m-skeleton (i.e. a chain with transition probability P^m) (see Theorem 5.2.3 in Meyn and Tweedie (1993)). This minorization condition is used in the Nummelin splitting technique and is also required in Clémençon (1999).

The Assumption H2a(ii), which is called geometric regularity by Clémençon (2000), means that the convergence of the chain to the invariant distribution is geometrically fast. In Meyn and Tweedie (1993), we find a slightly different condition (replacing the total variation norm by the V-norm). This condition, which is sufficient for H2a(ii), is widely used in Monte Carlo Markov Chain literature because it guarantees central limit theorems and enables to simulate laws via a Markov chain (see for example Jarner and Hansen (2000), Roberts and Rosenthal (1998) or Meyn and Tweedie (1994)).

For the estimation of the joint density F on the compact $A = A_1 \times A_1$, we need the additional assumption:

H3b Π belongs to $L^\infty(A)$.

A last assumption is required for the estimation of the transition Π:

H4 There exists a positive constant f_0 such that $\forall x \in A_1, \quad f(x) \geq f_0$.

The following subsection gives some examples of Markov chains satisfying hypotheses H1–H4.

1.2.2 Examples of chains

Many processes verify the previous assumptions, as (classical or more general) autoregressive processes, or diffusions. Here we give a nonexhaustive list of such chains.

Diffusion processes

We consider the process $(X_{i\Delta})_{1\leq i\leq n}$ where $\Delta > 0$ is the observation step and $(X_t)_{t\geq 0}$ is defined by the equation

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t,\ t \geq 0, \quad X_0 \sim \mu \tag{1.1}$$

where W is the standard Brownian motion, b is a locally bounded Borel function and σ an uniformly continuous function. We suppose that the drift function b and the diffusion coefficient σ satisfy the following conditions, given in Pardoux and Veretennikov (2001)(Proposition 1):

1. there exists λ_-, λ_+ such that $\quad \forall x \neq 0, \quad 0 < \lambda_- < \sigma^2(x) < \lambda_+,$

2. there exists $M_0 \geq 0, \alpha \geq 0$ and $r > 0$ such that

$$\forall |x| \geq M_0, \quad xb(x) \leq -r|x|^{\alpha+1}.$$

Then, Equation (1.1) has a unique strong solution. Moreover the discretized process $(X_{i\Delta})_{1\leq i\leq n}$ satisfies Assumptions H1 and H2a. The continuity of the transition density ensures that Assumption H3b holds. Moreover, we can write

$$f(x) = \frac{1}{M\sigma^2(x)} \exp\left[2\int_0^x \frac{b(u)}{\sigma^2(u)}du\right]$$

with M such that $\int f = 1$. Consequently Assumption H3a and H4 are verified with $\|f\|_{\infty,A_1} \leq \frac{1}{M\lambda_-} \exp\left[\frac{2}{\lambda_-}\sup_{x\in A_1}\int_0^x |b(u)|du\right]$ and $f_0 \geq \frac{1}{M\lambda_+} \exp\left[\frac{2}{\lambda_+}\inf_{x\in A_1}\int_0^x |b(u)|du\right]$.

Nonlinear AR(1) processes

Let us consider the following process

$$X_n = \varphi(X_{n-1}) + \varepsilon_{X_{n-1},n}$$

where $\varepsilon_{x,n}$ has a positive density l_x with respect to the Lebesgue measure, which does not depend on n. We suppose that the following conditions are verified:

1. There exist $M > 0$ and $\rho < 1$ such that, for all $|x| > M$, $|\varphi(x)| < \rho|x|$ and $\sup_{|x|\leq M} |\varphi(x)| < \infty$.

2. There exist $l_0 > 0$, $l_1 > 0$ such that $\forall x, y \quad l_0 \leq l_x(y) \leq l_1$.

Then Mokkadem (1987) proves that the chain is Harris recurrent and geometrically ergodic. It implies that Assumptions H1 and H2a are satisfied. Moreover $\Pi(x, y) = l_x(y - \varphi(x))$ and $f(y) = \int f(x)\Pi(x, y)dx$ and then Assumptions H3a,b-H4 hold with $f_0 \geq l_0$ and $\|f\|_{\infty,A_1} \leq \|\Pi\|_\infty \leq l_1$.

ARX(1,1) models

The nonlinear process ARX(1,1) is defined by

$$X_n = H(X_{n-1}, Z_n) + \xi_n$$

where H is bounded and (ξ_n), (Z_n) are independent sequences of i.i.d. random variables with $\mathbb{E}|\xi_n| < \infty$. We suppose that the distribution of Z_n has a positive density l with respect to the Lebesgue measure. Assume that there exist $\rho < 1$, a locally bounded and mesurable function $h : \mathbb{R} \mapsto \mathbb{R}^+$ such that $\mathbb{E}h(Z_n) < \infty$ and positive constants M, c such that

$$\forall |(u,v)| > M \quad |H(u,v)| < \rho|u| + h(v) - c \text{ and } \sup_{|x|\leq M} |H(x)| < \infty.$$

Then Doukhan (1994) proves (p.102) that (X_n) satisfies H1 and H2a. We can write

$$\Pi(x,y) = \int l(z) f_\xi(y - F(x,z)) dz$$

where f_ξ is the density of ξ_n. So, if we assume furthermore that there exist $a_0, a_1 > 0$ such that $a_0 \leq f_\xi \leq a_1$, then Assumptions H3a,b-H4 are verified with $f_0 \geq a_0$ and $\|f\|_{\infty,A_1} \leq \|\Pi\|_\infty \leq a_1$.

ARCH processes

The model is

$$X_{n+1} = H(X_n) + G(X_n)\varepsilon_{n+1}$$

where H and G are continuous functions and for all x, $G(x) \neq 0$. We suppose that the distribution of ε_n has a positive density l with respect to the Lebesgue measure and that there exists $s \geq 1$ such that $\mathbb{E}|\varepsilon_n|^s < \infty$. The chain (X_n) satisfies Assumptions H1 and H2a if (see Ango Nzé (1992)):

$$\limsup_{|x|\to\infty} \frac{|H(x)| + |G(x)|(\mathbb{E}|\varepsilon_n|^s)^{1/s}}{|x|} < 1. \tag{1.2}$$

In addition, we assume that $\forall x \quad l_0 \leq l(x) \leq l_1$. Then Assumption H3b is verified with $\|\Pi\|_{\infty,A} \leq l_1 / \inf_{x\in A_1} G(x)$. And Assumption H3a-H4 hold with $f_0 \geq l_0 \int fG^{-1}$ and $\|f\|_{\infty,A_1} \leq l_1 \int fG^{-1}$.

1.2.3 Assumptions on the models

In order to estimate f, we need to introduce some collections of models. We recall that $\mathcal{D}(A_1)$ denotes the set of functions with support included in A_1. The assumptions on the models are the following:

M1. Each S_m is a linear subspace of $L^\infty(A_1) \cap \mathcal{D}(A_1)$ with dimension $D_m \leq \sqrt{n}$.

M2. Let

$$\phi_m = \frac{1}{\sqrt{D_m}} \sup_{t \in S_m \setminus \{0\}} \frac{\|t\|_\infty}{\|t\|}.$$

There exists a real ϕ_0 such that for all m, $\quad \phi_m \leq \phi_0$.

This assumption (L^2-L^∞ connexion) is introduced by Barron *et al.* (1999) and can be written:

$$\forall t \in S_m \qquad \|t\|_\infty \leq \phi_0 \sqrt{D_m} \|t\|. \tag{1.3}$$

We get then a set of models $(S_m)_{m \in \mathcal{M}_n}$ where $\mathcal{M}_n = \{m, \quad D_m \leq \sqrt{n}\}$. Now we need a last assumption regarding the whole collection, which ensures that, for m and m' in $\mathcal{M}_n$, $S_m + S'_m$ belongs to the collection of models.

M3. The models are nested, that is for all $m, m' \in \mathcal{M}_n$, $D_m \leq D_{m'} \Rightarrow S_m \subset S_{m'}$.

1.2.4 Examples of models

We show here that the assumptions M1-M3 are not too restrictive. Indeed, for $A_1 = [0, 1]$, they are verified for the models spanned by the following bases (see Barron *et al.* (1999)):

- Histogram basis: $S_m =< \varphi_1, \ldots, \varphi_{2^m} >$ with $\varphi_j = 2^{m/2} \mathbb{1}_{[\frac{j-1}{2^m}, \frac{j}{2^m}[}$ for $j = 1, \ldots, 2^m$. Here $D_m = 2^m$, $\phi_0 = 1$ and $\mathcal{M}_n = \{1, \ldots, \lfloor \log n / 2 \log 2 \rfloor\}$ where $\lfloor x \rfloor$ denotes the floor of x, i.e. the largest integer less than or equal to x.

- Trigonometric basis: $S_m =< \varphi_0, \ldots, \varphi_{m-1} >$ with $\varphi_0(x) = \mathbb{1}_{[0,1]}(x)$, $\varphi_{2j} = \sqrt{2} \cos(2\pi j x) \mathbb{1}_{[0,1]}(x)$, $\varphi_{2j-1} = \sqrt{2} \sin(2\pi j x) \mathbb{1}_{[0,1]}(x)$ for $j \geq 1$. For this model $D_m = m$ and $\phi_0 = \sqrt{2}$ hold.

- Regular piecewise polynomial basis: S_m is spanned by polynomials of degree $0, \ldots, r$ (where r is fixed) on each interval $[(j-1)/2^D, j/2^D[, j = 1, \ldots, 2^D$. In this case, $m = (D, r)$, $D_m = (r+1)2^D$ and $\mathcal{M}_n = \{(D, r),\ D = 1, \ldots, \lfloor \log_2(\sqrt{n}/(r+1)) \rfloor\}$.We can put $\phi_0 = \sqrt{r+1}$.

- Regular wavelet basis: $S_m =< \psi_{jk}, j = -1, \ldots, m, k \in \Lambda(j) >$ where $\psi_{-1,k}$ points out the translates of the father wavelet and $\psi_{jk}(x) = 2^{j/2} \psi(2^j x - k)$ where ψ is the mother wavelet. We assume that the support of the wavelets is included in $[0, 1]$ and that $\psi_{-1} = \varphi$ belongs to the Sobolev space W_2^r. In this framework $\Lambda(j) = \{0, \ldots, K2^j - 1\}$ (for $j \geq 0$) where K is a constant which depends on the supports of

φ and ψ: for example for the Haar basis $K = 1$. We have then $D_m = \sum_{j=-1}^{m} |\Lambda(j)| = |\Lambda(-1)| + K(2^{m+1} - 1)$. Moreover

$$\begin{aligned}\phi_m &\leq \frac{\sum_k |\psi_{-1,k}| + \sum_{j=0}^{m} \sum_k |\psi_{j,k}|}{\sqrt{D_m}} \\ &\leq \frac{\|\varphi\|_\infty \vee \|\psi\|_\infty (1 + \sum_{j=0}^{m} 2^{j/2})}{\sqrt{(K \wedge |\Lambda(-1)|) 2^{m+1}}} \leq \frac{\|\varphi\|_\infty \vee \|\psi\|_\infty}{K \wedge |\Lambda(-1)|} =: \phi_0\end{aligned}$$

1.3 Estimation of the stationary density

1.3.1 Decomposition of the risk for the projection estimator

Let us define the contrast

$$\gamma_n(t) = \frac{1}{n} \sum_{i=1}^{n} [\|t\|^2 - 2t(X_i)]. \tag{1.4}$$

Notice that $\mathbb{E}(\gamma_n(t)) = \|t - f\|^2 - \|f\|^2$ and therefore $\gamma_n(t)$ is the empirical version of the L^2 distance between t and f. Thus, $\hat{f}_m$ is defined by

$$\hat{f}_m = \arg\min_{t \in S_m} \gamma_n(t) \tag{1.5}$$

where S_m is a subspace of $(L^\infty \cap \mathcal{D})(A_1)$ which satisfies M2. Although this estimator depends on n, no index n is mentioned in order to simplify the notations . It is also the case for all the estimators in this chapter and thereafter.

A more explicit formula for $\hat{f}_m$ is easy to derive:

$$\hat{f}_m = \sum_{\lambda \in \Lambda_m} \hat{\beta}_\lambda \varphi_\lambda, \qquad \hat{\beta}_\lambda = \frac{1}{n} \sum_{i=1}^{n} \varphi_\lambda(X_i) \tag{1.6}$$

where $(\varphi_\lambda)_{\lambda \in \Lambda_m}$ is an orthonormal basis of S_m. Note that

$$\mathbb{E}(\hat{f}_m) = \sum_{\lambda \in \Lambda_m} < f, \varphi_\lambda > \varphi_\lambda,$$

which is the projection of f on S_m.

In order to evaluate the quality of this estimator, we now compute the mean integrated squared error $\mathbb{E}\|f - \hat{f}_m\|_{A_1}^2$ (often denoted by MISE).

Proposition 1.1 *Let X_n be a Markov chain which satisfies Assumptions H1-H2a-H3a and S_m be a subspace of $(L^\infty \cap \mathcal{D})(A_1)$ with dimension $D_m \leq n$. If S_m satisfies condition M2, then the estimator $\hat{f}_m$ defined by (1.5) satisfies*

$$\mathbb{E}\|f - \hat{f}_m\|_{A_1}^2 \leq d^2(f \mathbb{1}_{A_1}, S_m) + C D_m / n$$

where C is a constant which does not depend on n.

To compute the bias term $d(f\mathbb{1}_{A_1}, S_m)$, we assume that f (actually the restriction of f to A_1) belongs to the Besov space $B^\alpha_{2,\infty}(A_1)$. We refer to DeVore and Lorentz (1993) p.54 for the definition of $B^\alpha_{2,\infty}(A_1)$. Notice that when α is an integer, the Besov space $B^\alpha_{2,\infty}(A_1)$ contains the Sobolev space W^α_2 (see DeVore and Lorentz (1993) p.51–55).

Hence, we have the following corollary.

Corollary 1.1 *Let X_n be a Markov chain which satisfies Assumptions H1-H2a-H3a. Assume that the stationary density f belongs to $B^\alpha_{2,\infty}(A_1)$ and that S_m is one of the spaces mentioned in Section 1.2.4 (with the regularity of polynomials and wavelets larger than $\alpha-1$). If we choose $D_m = \lfloor n^{\frac{1}{2\alpha+1}} \rfloor$, then the estimator defined by* (1.5) *satisfies*

$$\mathbb{E}\|f - \hat{f}_m\|^2_{A_1} = O(n^{-\frac{2\alpha}{2\alpha+1}}).$$

We can notice that we obtain the same rate as in the i.i.d. case (see Donoho *et al.* (1996)). Actually, Clémençon (1999) proves that $n^{-\frac{2\alpha}{2\alpha+1}}$ is the optimal rate in the minimax sense in the Markovian framework. With very different theoretical tools, Tribouley and Viennet (1998) show that this rate is also reached in the case of the univariate density estimation of β-mixing random variables by using a wavelet estimator.

However, the choice $D_m = \lfloor n^{\frac{1}{2\alpha+1}} \rfloor$ is possible only if we know the regularity α of the unknown f. But generally, it is not the case. It is the reason why we construct an adaptive estimator, i.e. an estimator which achieves the optimal rate without requiring the knowledge of α.

1.3.2 Adaptive estimation

Let $(S_m)_{m\in\mathcal{M}_n}$ be a collection of models as described in Section 1.2.3. For each S_m, $\hat{f}_m$ is defined as above by (1.5). Next, we choose $\hat{m}$ among the family $\mathcal{M}_n$ such that

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n}[\gamma_n(\hat{f}_m) + \mathrm{pen}(m)]$$

where pen is a penalty function to be specified later. We denote $\tilde{f} = \hat{f}_{\hat{m}}$ and we bound the L^2-risk $\mathbb{E}\|f - \tilde{f}\|_{A_1}$ as follows.

Theorem 1.1 *Let X_n be a Markov chain which satisfies Assumptions H1-H2a-H3a and $(S_m)_{m\in\mathcal{M}_n}$ be a collection of models satisfying Assumptions M1–M3. Then the estimator defined by*

$$\tilde{f} = \hat{f}_{\hat{m}} \quad \textit{where} \quad \hat{m} = \arg\min_{m\in\mathcal{M}_n}[\gamma_n(\hat{f}_m) + \mathrm{pen}(m)], \tag{1.7}$$

with

$$\mathrm{pen}(m) = K\frac{D_m}{n} \quad \textit{for some } K > K_0 \tag{1.8}$$

(where K_0 is a constant which depends on ϕ_0, $\|f\|_{\infty,A_1}$, h and ν) satisfies

$$\mathbb{E}\|\tilde{f} - f\|_{A_1}^2 \leq 3 \inf_{m\in\mathcal{M}_n} \{d^2(f\mathbb{1}_{A_1}, S_m) + \mathrm{pen}(m)\} + \frac{C}{n}$$

where C does not depend on n.

The constant K_0 in the penalty only depends on the distribution of the chain and can be chosen equal to $\max(\phi_0^2, 1)(C_1 + C_2\|f\|_{\infty,A_1})$ where C_1 and C_2 are constants depending on the quantities h and ν introduced in Assumption H2a. These are theoretical constants provided by the Nummelin splitting technique. The number ϕ_0 is known and depends on the chosen base (see Section 1.2.3). The mention of $\|f\|_{\infty,A_1}$ in the penalty term seems to be a problem, seeing that f is unknown. Actually, we could replace $\|f\|_{\infty,A_1}$ by $\|\hat{f}\|_{\infty,A_1}$ with $\hat{f}$ an estimator of f. This method of random penalty is successfully applied in Birgé and Massart (1997) or Comte (2001) for example. But we choose not to use this method here, since the constants C_1 and C_2 in K_0 are not computable either. Notice that Clémençon (2000) handle with the same kind of unknown quantities in the threshold of his nonlinear wavelet estimator. Actually it is the price to pay for dealing with dependent variables (see also the mixing constant in the threshold in Tribouley and Viennet (1998)). But this annoyance can be circumvented for practical purposes. Indeed, for the simulations the computation of the penalty is hand-adjusted. Some techniques of calibration can be found in Lebarbier (2005) in the context of multiple change point detection. In a Gaussian framework the practical choice of the penalty for implementation is also discussed in Section 4 of Birgé and Massart (2007).

Corollary 1.2 *Let X_n be a Markov chain which satisfies Assumptions H1-H2a-H3a and $(S_m)_{m\in\mathcal{M}_n}$ be a collection of models mentioned in Section 1.2.4 (with the regularity of polynomials and wavelets larger than $\alpha - 1$). If f belongs to $B^{\alpha}_{2,\infty}(A_1)$, with $\alpha > 1/2$, then the estimator defined by* (1.7) *and* (1.8) *satisfies*

$$\mathbb{E}\|\tilde{f} - f\|_{A_1}^2 = O(n^{-\frac{2\alpha}{2\alpha+1}}).$$

Remark 1.1 *When $\alpha > \frac{1}{2}$, $B^{\alpha}_{2,\infty}(A_1) \subset C(A_1)$ (where $C(A_1)$ is the set of the continuous functions with support in A_1) and then the assumption H3a $\|f\|_{\infty,A_1} < \infty$ is superfluous.*

We have already noticed that it is the optimal rate in the minimax sense (see the lower bound in Clémençon (1999)). Note that here the procedure reaches this rate whatever the regularity of f, without needing to know α. This result is thus a improvement of the one of Clémençon (1999), whose adaptive procedure only achieves the rate $(\log(n)/n)^{\frac{2\alpha}{2\alpha+1}}$. Moreover, our procedure allows to use more bases (not only wavelets) and is easy to implement.

1.4 Estimation of the transition density

We now suppose that the transition kernel P has a density Π. In order to estimate Π, we remark that Π can be written F/f where F is the density of (X_i, X_{i+1}). Thus we begin with the estimation of F. As previously, F and Π are estimated on a compact set only, this compact is $A = A_1 \times A_2$ and here we consider that $A_2 = A_1$.

1.4.1 Estimation of the joint density F

We consider now the following subspaces.

$$\mathbb{S}_m = \{T : \mathbb{R}^2 \mapsto \mathbb{R}, \quad T(x,y) = \sum_{\lambda,\mu \in \Lambda_m} \alpha_{\lambda,\mu} \varphi_\lambda(x) \varphi_\mu(y)\}$$

where $(\varphi_\lambda)_{\lambda \in \Lambda_m}$ is an orthonormal basis of S_m. Notice that, if we set

$$\Phi_m \quad = \quad \frac{1}{D_m} \sup_{T \in \mathbb{S}_m \setminus \{0\}} \frac{\|T\|_\infty}{\|T\|},$$

hypothesis M2 implies that Φ_m is bounded by ϕ_0^2. The condition M1 must be replaced by the following condition:

M1'. Each $\mathbb{S}_m$ is a linear subspace of $(L^\infty \cap \mathcal{D})(A)$ with dimension $D_m^2 \leq \sqrt{n}$.

Let now

$$\Gamma_n(T) = \frac{1}{n-1} \sum_{i=1}^{n-1} \{\|T\|^2 - 2T(X_i, X_{i+1})\}$$

the contrast to estimate F. We define as above

$$\hat{F}_m = \arg\min_{T \in \mathbb{S}_m} \Gamma_n(T)$$

and $\hat{M} = \arg\min_{m \in \mathcal{M}_n} [\Gamma_n(\hat{F}_m) + \mathrm{Pen}(m)]$ where $\mathrm{Pen}(m)$ is a penalty function which would be specified later. Lastly, we set $\tilde{F} = \hat{F}_{\hat{M}}$.

Theorem 1.2 *Let X_n be a Markov chain which satisfies Assumptions H1-H2a-H3a,b and $(S_m)_{m \in \mathcal{M}_n}$ be a collection of models satisfying Assumptions M1'-M2-M3. Then the estimator defined by*

$$\tilde{F} = \hat{F}_{\hat{M}} \quad \textit{where} \quad \hat{M} = \arg\min_{m \in \mathcal{M}_n} [\Gamma_n(\hat{F}_m) + \mathrm{Pen}(m)], \tag{1.9}$$

with

$$\mathrm{Pen}(m) = K^{(2)} \frac{D_m^2}{n} \quad \textit{for some } K^{(2)} > K_0^{(2)} \tag{1.10}$$

(where $K_0^{(2)}$ is a constant which depends on ϕ_0, $\|F\|_{\infty,A}$, h and ν) satisfies

$$\mathbb{E}\|\tilde{F} - F\|_A^2 \leq 3 \inf_{m \in \mathcal{M}_n} \{d^2(F\mathbb{1}_A, \mathbb{S}_m) + \mathrm{Pen}(m)\} + \frac{C}{n}$$

where C does not depend on n.

The constant $K_0^{(2)}$ in the penalty is similar to the constant K_0 in Theorem 1.1 (replacing ϕ_0 by ϕ_0^2 and $\|f\|_{\infty,A_1}$ by $\|F\|_{\infty,A}$). We refer the reader to the remark following Theorem 1.1 on page 34 for considerations related to these constants.

Corollary 1.3 *Let X_n be a Markov chain which satisfies Assumptions H1-H2a-H3a,b and $(S_m)_{m\in\mathcal{M}_n}$ be a collection of models mentioned in Section 1.2.4 (with the regularity of polynomials and wavelets larger than $\beta - 1$). If (the restriction to A of) F belongs to $B^{\beta}_{2,\infty}(A)$, with $\beta > 1$, then*

$$\mathbb{E}\|\tilde{F} - F\|_A^2 = O(n^{-\frac{2\beta}{2\beta+2}}).$$

This rate of convergence is the minimax rate for density estimation in dimension 2 in the case of i.i.d. random variables (see for instance Ibragimov and Has′minskiĭ (1980)). Let us now proceed to the estimation of the transition density.

1.4.2 Estimation of Π

The estimator of Π is defined in the following way. Let

$$\tilde{\Pi}(x,y) = \begin{cases} \dfrac{\tilde{F}(x,y)}{\tilde{f}(x)} & \text{if } |\tilde{F}(x,y)| \leq k_n |\tilde{f}(x)| \\ 0 & \text{else} \end{cases}$$

with $k_n = n$.

Theorem 1.3 *Let X_n be a Markov chain which satisfies Assumptions H1-H2a-H3a,b-H4 and $(S_m)_{m\in\mathcal{M}_n}$ be a collection of models mentioned in Section 1.2.4 (with the regularity of polynomials and wavelets larger than $\alpha - 1$). We suppose that the dimension D_m of the models is such that, $\forall m \in \mathcal{M}_n$,*

$$\begin{array}{ll} \textit{for the estimation of } f & \log n \leq D_m \leq \sqrt{n}/(\log n), \\ \textit{for the estimation of } F & D_m^2 \leq \sqrt{n}. \end{array}$$

If f belongs to $B^{\alpha}_{2,\infty}(A_1)$, with $\alpha > 1/2$, then for n large enough

- *there exists C_1 and C_2 such that*

$$\mathbb{E}\|\Pi - \tilde{\Pi}\|_A^2 \leq C_1 \mathbb{E}\|F - \tilde{F}\|_A^2 + C_2 \mathbb{E}\|f - \tilde{f}\|_{A_1}^2 + o\left(\frac{1}{n}\right)$$

- *if furthermore F belongs to $B_{2,\infty}^{\beta}(A)$ with $\beta > 1$, then*

$$\mathbb{E}\|\Pi - \tilde{\Pi}\|_A^2 = O(\sup(n^{-\frac{2\beta}{2\beta+2}}, n^{-\frac{2\alpha}{2\alpha+1}})).$$

Clémençon (2000) proved that $n^{-2\beta/(2\beta+2)}$ is the minimax rate for f and F of same regularity β. Notice that in this case the procedure is adaptive and there is no logarithmic loss in the estimation rate contrary to the result of Clémençon (2000).

But it should be remembered that we consider only the restriction of f or Π since the observations are in a compact set. And the restriction of the stationary density to A_1 may be less regular than the restriction of the transition density. The previous procedure has thus the disadvantage that the resulting rate does not depend only on the regularity of Π but also on the one of f.

However, if the chain lives on A_1 and if F belongs to $B_{2,\infty}^{\beta}(A)$ (that is to say that we consider the regularity of F on its whole support and not only on the compact of the observations) then equality $f(y) = \int F(x,y)dx$ yields that f belongs to $B_{2,\infty}^{\beta}(A_1)$ and then $\mathbb{E}\|\Pi - \tilde{\Pi}\|^2 = O(n^{-\frac{2\beta}{2\beta+2}})$. Moreover, if Π belongs to $B_{2,\infty}^{\beta}(A)$, formula $f(y) = \int f(x)\Pi(x,y)dx$ implies that f belongs to $B_{2,\infty}^{\beta}(A_1)$. Then, by using properties of Besov spaces (see Runst and Sickel (1996) p.192), $F = f\Pi$ belongs to $B_{2,\infty}^{\beta}(A)$. So in this case of a chain with compact support the minimax rate is achieved as soon as Π belongs to $B_{2,\infty}^{\beta}(A)$ with $\beta > 1$.

1.5 Simulations

The computation of the previous estimator is very simple. We use the following procedure in 3 steps:

First step:

- For each m, compute $\gamma_n(\hat{f}_m) + \text{pen}(m)$. Notice that $\gamma_n(\hat{f}_m) = -\sum_{\lambda \in \Lambda_m} \hat{\beta}_\lambda^2$ where $\hat{\beta}_\lambda$ is defined by (1.6) and is quickly computed.
- Select the argmin $\hat{m}$ of $\gamma_n(\hat{f}_m) + \text{pen}(m)$.
- Choose $\tilde{f} = \sum_{\lambda \in \Lambda_{\hat{m}}} \hat{\beta}_\lambda \varphi_\lambda$.

Second step:

- For each m such that $D_m^2 \leq \sqrt{n}$ compute $\Gamma_n(\hat{F}_m) + \mathrm{Pen}(m)$, with $\Gamma_n(\hat{F}_m) = -\sum_{\lambda,\mu\in\Lambda_m} \hat{a}_{\lambda,\mu}^2$ where $\hat{a}_{\lambda,\mu} = (1/(n-1))\sum_{i=1}^{n-1} \varphi_\lambda(X_i)\varphi_\mu(X_{i+1})$.
- Select the argmin $\hat{M}$ of $\Gamma_n(\hat{F}_m) + \mathrm{Pen}(m)$.
- Choose $\tilde{F}(x,y) = \sum_{\lambda,\mu\in\Lambda_{\hat{M}}} \hat{a}_{\lambda,\mu}\varphi_\lambda(x)\varphi_\mu(y)$.

Third step: Compute $\tilde{\Pi}(x,y) = \tilde{F}(x,y)/\tilde{f}(x)$ if $|\tilde{F}(x,y)| \leq k_n|\tilde{f}(x)|$ and 0 otherwise.

We consider 2 different bases (see Section 1.2.4): trigonometric basis and histogram basis[(1)]. The bases are here adjusted with an affin transform in order to be defined on the estimation interval $[c,d]$ instead of $[0,1]$.

We found that a good choice for the penalty functions is $\mathrm{pen}(m) = 5D_m/n$ and $\mathrm{Pen}(m) = 0.02D_m^2/n$.

We consider several kinds of Markov chains :

- An autoregressive process denoted by AR and defined by:

$$X_{n+1} = aX_n + b + \varepsilon_{n+1}$$

  where the ε_{n+1} are independent and identical distributed random variables, with centered Gaussian distribution with variance σ^2. For this process, the stationary distribution is a Gaussian with mean $b/(1-a)$ and variance $\sigma^2/(1-a^2)$. By denoting by $\varphi(z) = 1/(\sigma\sqrt{2\pi})\exp(-z^2/2\sigma^2)$ the Gaussian density, the transition density can be written $\Pi(x,y) = \varphi(y-ax-b)$. We consider the following parameter values :

  (i) $a = 2/3$, $b = 0$, $\sigma^2 = 5/9$, estimated on $[-2,2]^2$. The stationary density of this chain is the standard Gaussian distribution.

  (ii) $a = 0.5$, $b = 3$, $\sigma^2 = 1$, and then the process is estimated on $[4,8]^2$.

- A radial Ornstein-Uhlenbeck process (in its discrete version). For $j = 1,\ldots,\delta$, we define the processes: $\xi_{n+1}^j = a\xi_n^j + \beta\varepsilon_n^j$ where the ε_n^j are i.i.d. standard Gaussian. The chain is then defined by $X_n = \sqrt{\sum_{i=1}^{\delta}(\xi_n^i)^2}$. The transition density is given in Chaleyat-Maurel and Genon-Catalot (2006) where this process is studied in detail:

$$\Pi(x,y) = \mathbb{1}_{y>0}\exp\left(-\frac{y^2+a^2x^2}{2\beta^2}\right) I_{\delta/2-1}\left(\frac{axy}{\beta^2}\right)\frac{ax}{\beta^2}\left(\frac{y}{ax}\right)^{\delta/2}$$

  and $I_{\delta/2-1}$ is the Bessel function with index $\delta/2-1$. The invariant density is $f(x) = C\mathbb{1}_{x>0}\exp(-x^2/2\rho^2)x^{\delta-1}$ with $\rho^2 = \beta^2/(1-a^2)$ and C such that $\int f = 1$. This

(1) Although the rates of convergence given in this chapter are valid only if the bases are regular enough, we can consider the estimators when the basis is less regular. Actually it is known that a very regular basis does not necessarily improve the performance of the estimator for practical purposes and an histogram basis can be better than very smooth bases for some given processes.

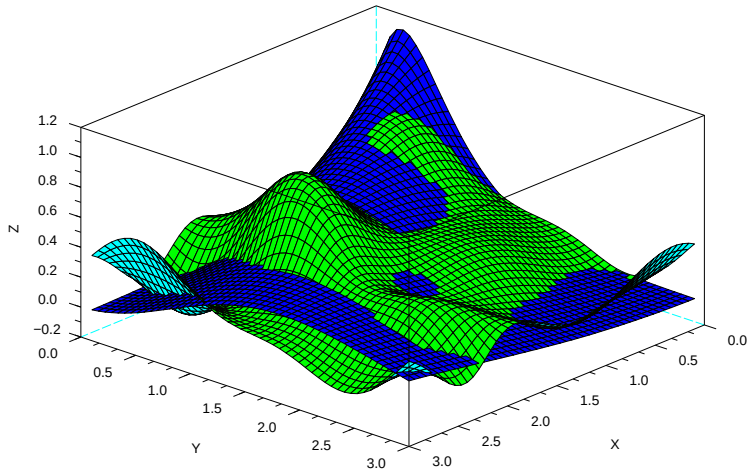


Figure 1.1: Estimator (light surface) and true transition (dark surface) for the process CIR(iii) estimated with a trigonometric basis, n=1000.

process (with here $a = 0.5$, $\beta = 3$, $\delta = 3$) is denoted by $\sqrt{\text{CIR}}$ since its square is actually a Cox-Ingersoll-Ross process. The estimation domain for this process is $[2, 10]^2$.

- A Cox-Ingersoll-Ross process, which is exactly the square of the previous process. It follows a Gamma density for invariant distribution with scale parameter $l = 1/2\rho^2$ and shape parameter $a = \delta/2$. The transition density is

$$\Pi(x, y) = \frac{1}{2\beta^2} \exp\left(-\frac{y + a^2 x}{2\beta^2}\right) I_{\delta/2-1}\left(\frac{a\sqrt{xy}}{\beta^2}\right) \left(\frac{y}{a^2 x}\right)^{\delta/4-1/2}$$

  The used parameters are the following:

  (iii) $a = 3/4$, $\beta = \sqrt{7/48}$ (so that $l = 3/2$) and $\delta = 4$, estimated on $[0.1, 3]^2$.

  (iv) $a = 1/3$, $\beta = 3/4$ and $\delta = 2$. This chain is estimated on $[0, 2]^2$.

- An ARCH process defined by $X_{n+1} = \sin(X_n) + (\cos(X_n) + 3)\varepsilon_{n+1}$ where the ε_{n+1} are i.i.d. standard Gaussian. The transition density of this chain is

$$\Pi(x, y) = \varphi\left(\frac{y - \sin(x)}{\cos(x) + 3}\right) \frac{1}{\cos(x) + 3}$$

  and we estimate this process on $[-5, 5]^2$.

n	50	100	250	500	1000	basis
AR(i)	0.728	0.544	0.277	0.187	0.177	H
	0.526	0.468	0.222	0.180	0.148	T
AR(ii)	0.480	0.325	0.225	0.116	0.084	H
	0.287	0.239	0.177	0.134	0.108	T
$\sqrt{\text{CIR}}$	0.305	0.232	0.172	0.152	0.128	H
	0.216	0.194	0.145	0.128	0.082	T
CIR(iii)	0.509	0.308	0.211	0.176	0.148	H
	0.417	0.396	0.284	0.257	0.227	T
CIR(iv)	0.338	0.210	0.121	0.076	0.046	H
	0.227	0.221	0.172	0.134	0.133	T
ARCH	0.317	0.301	0.242	0.212	0.161	H
	0.255	0.254	0.208	0.188	0.169	T

Table 1.1: MISE $\mathbb{E}\|\Pi - \tilde{\Pi}\|^2$ averaged over $N = 200$ samples. H: histogram basis, T: trigonometric basis.

For this last chain, the stationary density is not explicit. So we simulate $n + 500$ variables and we estimate only from the last n to ensure the stationarity of the process. For the other chains, it is sufficient to simulate an initial variable X_0 with density f.

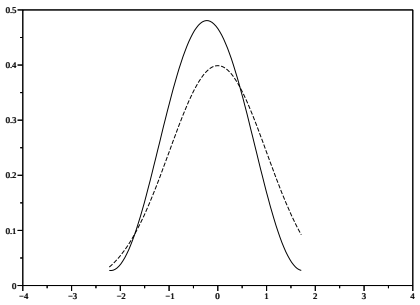
n	50	100	250	500	1000	basis
AR(i)	0.066	0.060	0.033	0.014	0.012	H
	0.057	0.054	0.025	0.004	0.003	T
AR(ii)	0.039	0.035	0.031	0.015	0.008	H
	0.034	0.034	0.033	0.020	0.005	T
$\sqrt{\text{CIR}}$	0.013	0.012	0.011	0.010	0.010	H
	0.017	0.017	0.017	0.016	0.011	T
CIR(iii)	0.034	0.027	0.023	0.022	0.021	H
	0.063	0.039	0.022	0.021	0.019	T
CIR(iv)	0.032	0.025	0.022	0.019	0.010	H
	0.087	0.073	0.057	0.052	0.046	T

Table 1.2: MISE $\mathbb{E}\|f - \tilde{f}\|^2$ averaged over $N = 200$ samples. H: histogram basis, T: trigonometric basis.

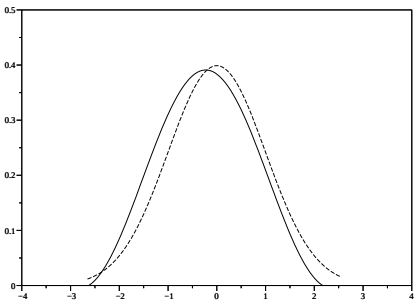
Figure 1.1 illustrates the performance of the method and Table 1.1 shows the L^2-risk for different values of n.

The results in Table 1.1 are roughly good and illustrate that we can not pretend that a basis among the others gives better results. We can then imagine a mixed strategy, i.e. a procedure which uses several kinds of bases and which can choose the best basis or, for instance, the best degree for a polynomial basis. These techniques are successfully used in regression frameworks by Comte and Rozenholc (2002, 2004).

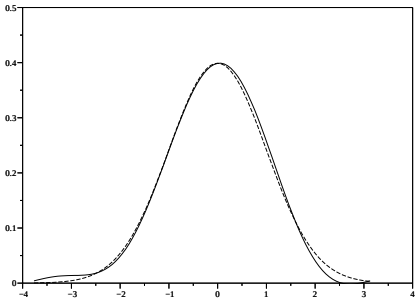
The results for the stationary density are given in Table 1.2 and illustrated in Figure 1.2.



$n = 100$



$n = 250$



$n = 1000$

Figure 1.2: Estimator (solid line) and true function (dotted line) for a Gaussian distribution estimated with a trigonometric basis.

We can compare results of Table 1.2 with those of Dalelane (2005) who gives results of simulations for i.i.d. random variables. For density estimation, she uses three types of kernel: Gauss kernel, sinc-kernel (where $\mathrm{sinc}(x) = \sin(x)/x$) and her Cross Validation optimal kernel (denoted by Dal). Table 1.3 gives her results for the Gaussian density and the Gamma distribution with the same parameters that we used (2 and 3/2). If we compare the results that she obtains with her optimal kernel and our results with the trigonometric basis, we observe that her risks are about 5 times less than ours. However

n	50	100	500	1000	kernel
	0.012	0.007	0.001	0.001	Dal
Gaussian	0.017	0.013	0.003	0.002	Gauss
(=AR(i))	0.022	0.011	0.003	0.001	sinc
	0.027	0.015	0.005	0.003	Dal
Gamma	0.028	0.021	0.006	0.003	Gauss
(=CIR(iii))	0.062	0.040	0.017	0.004	sinc

Table 1.3: MISE obtained by Dalelane (2005) for i.i.d. data, averaged over 50 samples

this kernel is particularly effective and if we consider the classical kernels, we notice that the results are almost comparable, with a reasonable price for dependency.

1.6 Proofs

1.6.1 The Nummelin splitting technique

This whole subsection is summarized from Höpfner and Löcherbach (2003) p.60–63 and is detailed for the sake of completeness.

The interest of the Nummelin splitting technique is to create a two-dimensional chain (the "split chain"), which contains automatically an atom. Let us recall the definition of an atom. Let $\mathfrak{A}$ be a set such that $\psi(\mathfrak{A}) > 0$ where ψ is an irreducibility measure. The set $\mathfrak{A}$ is called an atom for the chain (X_n) with transition kernel P if there exists a measure ν such that $P(x, B) = \nu(B)$, for all x in $\mathfrak{A}$ and for all event B.

Let us now describe the splitting method. Let $E = \mathbb{R}$ the state space and $\mathcal{E}$ the associated σ-field. Each point x in E is splitted in $x_0 = (x, 0) \in E_0 = E \times \{0\}$ and $x_1 = (x, 1) \in E_1 = E \times \{1\}$. Each set B in $\mathcal{E}$ is splitted in $B_0 = B \times \{0\}$ and $B_1 = B \times \{1\}$. Thus, we have defined a new probability space $(E^*, \mathcal{E}^*)$ where $E^* = E_0 \cup E_1$ and $\mathcal{E}^* = \sigma(B_0, B_1 : B \in \mathcal{E})$. Using h defined in H2a(i), a measure λ on $(E, \mathcal{E})$ splits according to

$$\begin{cases} \lambda^*(B_1) &= \int \mathbb{1}_B(x) h(x) \lambda(dx) \\ \lambda^*(B_0) &= \int \mathbb{1}_B(x)(1 - h)(x) \lambda(dx) \end{cases}$$

Notice that $\lambda^*(B_0 \cup B_1) = \lambda(B)$. Now the aim is to define a new transition probability $P^*(.,.)$ on $(E^*, \mathcal{E}^*)$ to replace the transition kernel P of (X_n). Let

$$P^*(x_i, .) = \begin{cases} \dfrac{1}{1 - h(x)} (P - h \otimes \nu)^*(x, .) & \text{if } i = 0 \text{ and } h(x) > 1 \\ \nu^* & \text{else} \end{cases}$$

where ν is the measure introduced in H2a(i) and $h \otimes \nu$ is a kernel defined by

$$h \otimes \nu(x, dy) = h(x)\nu(dy).$$

Consider now a chain (X_n^*) on $(E^*, \mathcal{E}^*)$ with one-step transition P^* and with starting law μ^*. The split chain (X_n^*) has the following properties:

P1. For all $(B_p)_{0 \leq p \leq N} \in \mathcal{E}^N$ and for all measure λ

$$P_\lambda(X_p \in B_p, 0 \leq p \leq N) = P_{\lambda^*}(X_p^* \in B_p \times \{0,1\}, 0 \leq p \leq N).$$

P2. The split chain is irreducible positive recurrent with stationary distribution μ^*.

P3. The set E_1 is an atom for (X_n^*).

We can also extend functions $g : E \mapsto \mathbb{R}$ to E^* via $g^*(x_0) = g(x) = g^*(x_1)$. Then, the property P1 can be written: for all function $\mathcal{E}$-measurable $g : E^N \mapsto \mathbb{R}$

$$\mathbb{E}_\lambda(g(X_1, .., X_N)) = \mathbb{E}_{\lambda^*}(g^*(X_1^*, .., X_N^*)).$$

We can say that (X_n) is a marginal chain of (X_n^*). When necessary, the following proofs are decomposed in two steps: first, we assume that the Markov chain has an atom, next we extend the result to the general chain by introducing the artificial atom E_1.

1.6.2 Proof of Proposition 1.1

First step: We suppose that (X_n) has an atom $\mathfrak{A}$.

Let f_m be the orthogonal projection of f on S_m. Pythagoras theorem gives us:

$$\mathbb{E}\|f - \hat{f}_m\|_{A_1}^2 = d^2(f\mathbb{1}_{A_1}, S_m) + \mathbb{E}\|f_m - \hat{f}_m\|_{A_1}^2.$$

We recognize in the right member a bias term and a variance term. According to the expresssion (1.6) of $\hat{f}_m$ the variance term can be written:

$$\mathbb{E}\|f_m - \hat{f}_m\|_{A_1}^2 = \sum_{\lambda \in \Lambda_m} \mathrm{Var}(\hat{\beta}_\lambda) = \sum_{\lambda \in \Lambda_m} \mathbb{E}(\nu_n^2(\varphi_\lambda)) \tag{1.11}$$

where $\nu_n(t) = (1/n)\sum_{i=1}^n [t(X_i) - < t, f >]$. By denoting $\tau = \tau(1) = \inf\{n \geq 1, X_n \in \mathfrak{A}\}$ and $\tau(j) = \inf\{n > \tau(j-1), X_n \in \mathfrak{A}\}$ for $j \geq 2$, we can decompose $\nu_n(t)$ in the classic following way:

$$\nu_n(t) = \nu_n^{(1)}(t) + \nu_n^{(2)}(t) + \nu_n^{(3)}(t) + \nu_n^{(4)}(t) \tag{1.12}$$

$$\begin{aligned}
\text{with} \quad \nu_n^{(1)}(t) &= \nu_n(t)\mathbb{1}_{\tau>n}, \\
\nu_n^{(2)}(t) &= \frac{1}{n}\sum_{i=1}^{\tau}[t(X_i) - < t, f >]\mathbb{1}_{\tau\le n}, \\
\nu_n^{(3)}(t) &= \frac{1}{n}\sum_{i=1+\tau(1)}^{\tau(l_n)}[t(X_i) - < t, f >]\mathbb{1}_{\tau\le n}, \\
\nu_n^{(4)}(t) &= \frac{1}{n}\sum_{i=\tau(l_n)+1}^{n}[t(X_i) - < t, f >]\mathbb{1}_{\tau\le n},
\end{aligned}$$

and $l_n = \sum_{i=1}^n \mathbb{1}_{\mathfrak{A}}(X_i)$ (number of visits to the atom $\mathfrak{A}$). Hence,

$$\nu_n(t)^2 \le 4\{{\nu_n}^{(1)}(t)^2 + {\nu_n}^{(2)}(t)^2 + {\nu_n}^{(3)}(t)^2 + {\nu_n}^{(4)}(t)^2\}.$$

We set $B_m = \{t \in S_m,\ \|t\| = 1\}$.

• To bound $\nu_n^{(1)}(t)^2$, notice that $|\nu_n(t)| \le 2\|t\|_\infty$. And then, by using M2 and (1.3), $|\nu_n^{(1)}(t)| \le 2\phi_0\sqrt{D_m}\|t\|\mathbb{1}_{\tau>n}$.Thus,

$$\mathbb{E}(\sup_{t\in B_m}\nu_n^{(1)}(t)^2) \le 4\phi_0^2 D_m P(\tau > n) \le 4\phi_0^2\mathbb{E}(\tau^2)\frac{D_m}{n^2}.$$

• We bound the second term in the same way. Since $|\nu_n^{(2)}(t)| \le 2(\tau/n)\|t\|_\infty$, we obtain $|\nu_n^{(2)}(t)| \le 2\|t\|\phi_0\tau\sqrt{D_m}/n$ and then

$$\mathbb{E}(\sup_{t\in B_m}\nu_n^{(2)}(t)^2) \le 4\phi_0^2\mathbb{E}(\tau^2)\frac{D_m}{n^2}.$$

• Let us study now the fourth term. As

$$|\nu_n^{(4)}(t)| \le 2\frac{n-\tau(l_n)}{n}\|t\|_\infty\mathbb{1}_{\tau\le n} \le 2(n-\tau(l_n))\frac{\sqrt{D_m}}{n}\phi_0\|t\|\mathbb{1}_{\tau\le n},$$

we get $\mathbb{E}(\sup_{t\in B_m}\nu_n^{(4)}(t)^2) \le 4\phi_0^2\frac{D_m}{n^2}\mathbb{E}((n-\tau(l_n))^2\mathbb{1}_{\tau\le n})$.

It remains to bound $\mathbb{E}((n-\tau(l_n))^2\mathbb{1}_{\tau\le n})$:

$$\begin{aligned}
\mathbb{E}_\mu((n-\tau(l_n))^2\mathbb{1}_{\tau\le n}) &= \sum_{k=1}^n \mathbb{E}_\mu((n-k)^2\mathbb{1}_{\tau(l_n)=k}\mathbb{1}_{\tau\le n}) \\
&= \sum_{k=1}^n (n-k)^2 P_\mu(X_{k+1}\notin\mathfrak{A},..,X_n\notin\mathfrak{A}|X_k\in\mathfrak{A})P_\mu(X_k\in\mathfrak{A}) \\
&= \sum_{k=1}^n (n-k)^2 P_{\mathfrak{A}}(X_1\notin\mathfrak{A},..,X_{n-k}\notin\mathfrak{A})\mu(\mathfrak{A})
\end{aligned}$$

by using the stationarity of X and the Markov property. Hence

$$\begin{aligned}\mathbb{E}_\mu((n-\tau(l_n))^2\mathbb{1}_{\tau\le n}) &= \sum_{k=1}^{n}(n-k)^2 P_{\mathfrak{A}}(\tau > n-k)\mu(\mathfrak{A}) \\ &\le \sum_{k=1}^{n-1}\frac{\mathbb{E}_{\mathfrak{A}}(\tau^4)}{(n-k)^2}\mu(\mathfrak{A}).\end{aligned}$$

Therefore $\mathbb{E}_\mu((n-\tau(l_n))^2\mathbb{1}_{\tau\le n}) \le 2\mathbb{E}_{\mathfrak{A}}(\tau^4)\mu(\mathfrak{A})$. Finally

$$\mathbb{E}(\sup_{t\in B_m}\nu_n^{(4)}(t)^2) \le 8\phi_0^2\mu(\mathfrak{A})\mathbb{E}_{\mathfrak{A}}(\tau^4)\frac{D_m}{n^2}$$

and we can summarize the last three results by

$$\mathbb{E}\left(\sup_{t\in B_m}\nu_n^{(1)}(t)^2+\nu_n^{(2)}(t)^2+\nu_n^{(4)}(t)^2\right) \le 8\phi_0^2[\mathbb{E}_\mu(\tau^2)+\mu(\mathfrak{A})\mathbb{E}_{\mathfrak{A}}(\tau^4)]\frac{D_m}{n^2}. \tag{1.13}$$

In particular, using that $D_m \le n$,

$$\mathbb{E}\left(\nu_n^{(1)}(\varphi_\lambda)^2+\nu_n^{(2)}(\varphi_\lambda)^2+\nu_n^{(4)}(\varphi_\lambda)^2\right) \le 8\phi_0^2\frac{\mathbb{E}_\mu(\tau^2)+\mu(\mathfrak{A})\mathbb{E}_{\mathfrak{A}}(\tau^4)}{n}.$$

• Last we can write $\nu_n^{(3)}(t) = (1/n)\sum_{j=1}^{l_n-1} S_j(t)\mathbb{1}_{\tau\le n}$ where

$$S_j(t) = \sum_{i=1+\tau(j)}^{\tau(j+1)}(t(X_i) - < t, f >). \tag{1.14}$$

We remark that, according to the Markov property, the $S_j(t)$ are independent identically distributed and centered. Thus,

$$\mathbb{E}(\nu_n^{(3)}(\varphi_\lambda)^2) \le \frac{1}{n^2}\sum_{j=1}^{l_n-1}\mathbb{E}|S_j(\varphi_\lambda)|^2.$$

Then, we use Lemma 1.1 below (proved on page 46) to bound the expectation of $\nu_n^{(3)}(\varphi_\lambda)^2$:

Lemma 1.1 *For all $m \ge 2$, $\mathbb{E}_\mu|S_j(t)|^m \le (2\|t\|_\infty)^{m-2}\|f\|_{\infty,A_1}\|t\|^2\mathbb{E}_{\mathfrak{A}}(\tau^m)$.*

We can then give the bound

$$\mathbb{E}(\nu_n^{(3)}(\varphi_\lambda)^2) \le \frac{1}{n^2}\sum_{j=1}^{n}\|f\|_{\infty,A_1}\|\varphi_\lambda\|^2\mathbb{E}_{\mathfrak{A}}(\tau^2) \le \frac{\|f\|_{\infty,A_1}\mathbb{E}_{\mathfrak{A}}(\tau^2)}{n}.$$

Finally

$$\mathbb{E}(\nu_n^2(\varphi_\lambda)) \leq \frac{4}{n}[8\phi_0^2(\mathbb{E}_\mu(\tau^2) + \mu(\mathfrak{A})\mathbb{E}_{\mathfrak{A}}(\tau^4)) + \|f\|_{\infty,A_1}\mathbb{E}_{\mathfrak{A}}(\tau^2)].$$

Let $C = 4[8\phi_0^2(\mathbb{E}_\mu(\tau^2) + \mu(\mathfrak{A})\mathbb{E}_{\mathfrak{A}}(\tau^4)) + \|f\|_{\infty,A_1}\mathbb{E}_{\mathfrak{A}}(\tau^2)]$. We obtain with (1.11)

$$\mathbb{E}\|f_m - \hat{f}_m\|_{A_1}^2 \leq C\frac{D_m}{n}.$$

Second step: We do not suppose any more that (X_n) has an atom.

Let us apply the Nummelin splitting technique to the chain (X_n) and let

$$\gamma_n^*(t) = \frac{1}{n}\sum_{i=1}^{n}[\|t\|^2 - 2t^*(X_i^*)]. \tag{1.15}$$

We define also

$$\hat{f}_m^* = \arg\min_{t\in S_m}\gamma_n^*(t). \tag{1.16}$$

Then the property P1 in Section 1.6.1 yields $\mathbb{E}\|f - \hat{f}_m^*\|_{A_1}^2 = \mathbb{E}\|f - \hat{f}_m\|_{A_1}^2$. The split chain having an atom (property P3), we can use the first step to deduce $\mathbb{E}\|f - \hat{f}_m^*\|_{A_1}^2 \leq d^2(f\mathbb{1}_{A_1}, S_m) + CD_m/n$. It follows that

$$\mathbb{E}\|f - \hat{f}_m\|_{A_1}^2 \leq d^2(f\mathbb{1}_{A_1}, S_m) + CD_m/n.$$

□

Proof of Lemma 1.1: For all j, $\mathbb{E}_\mu|S_j(t)|^m = \mathbb{E}_\mu|S_1(t)|^m = \mathbb{E}_\mu|\sum_{i=\tau+1}^{\tau(2)}\bar{t}(X_i)|^m$ where $\bar{t} = t - < t, f >$. Thus

$$\mathbb{E}_\mu|S_j(t)|^m = \sum_{k<l}\mathbb{E}\Big(\Big|\sum_{i=k+1}^{l}\bar{t}(X_i)\Big|^m|\tau = k, \tau(2) = l\Big)P(\tau = k, \tau(2) = l)$$

$$\leq \sum_{k<l}(2\|t\|_\infty(l-k))^{m-2}\mathbb{E}\Big(\Big|\sum_{i=k+1}^{l}\bar{t}(X_i)\Big|^2|\tau = k, \tau(2) = l\Big)P(\tau = k, \tau(2) = l)$$

$$\leq \sum_{k<l}(2\|t\|_\infty)^{m-2}(l-k)^{m-1}\sum_{i=k+1}^{l}\mathbb{E}\Big(|\bar{t}(X_i)|^2|\tau = k, \tau(2) = l\Big)P(\tau = k, \tau(2) = l)$$

using the Schwarz inequality. Then, since the X_i have the same distribution under μ.

$$\begin{aligned}\mathbb{E}_\mu|S_j(t)|^m &\leq \sum_{k<l}(2\|t\|_\infty)^{m-2}(l-k)^m\mathbb{E}(t^2(X_1))P(\tau = k, \tau(2) = l)\\ &\leq \sum_{k<l}(2\|t\|_\infty)^{m-2}(l-k)^m\|f\|_{\infty,A_1}\|t\|^2P(\tau = k, \tau(2) = l)\\ &\leq (2\|t\|_\infty)^{m-2}\mathbb{E}(|\tau(2) - \tau|^m)\|f\|_{\infty,A_1}\|t\|^2.\end{aligned}$$

We conclude by using the Markov property. □

1.6.3 Proof of Corollary 1.1

According to Proposition 1.1 $\mathbb{E}\|f-\hat{f}_m\|_{A_1}^2 \leq d^2(f\mathbb{1}_{A_1}, S_m)+CD_m/n$. Then we use Lemma 12 in Barron *et al.* (1999) which ensures that (for piecewise polynomials or wavelets having a regularity larger than $\alpha-1$ and for trigonometric polynomials) $d^2(f\mathbb{1}_{A_1}, S_m) = O(D_m^{-2\alpha})$. Thus,

$$\mathbb{E}\|f-\hat{f}_m\|_{A_1}^2 = O(D_m^{-2\alpha} + \frac{D_m}{n})$$

In particular, if $D_m = \lfloor n^{\frac{1}{1+2\alpha}} \rfloor$, then $\mathbb{E}\|f-\hat{f}_m\|_{A_1}^2 = O(n^{-\frac{2\alpha}{1+2\alpha}})$. □

1.6.4 Proof of Theorem 1.1

First step: We suppose that (X_n) has an atom $\mathfrak{A}$.

Let m in $\mathcal{M}_n$. The definition of $\hat{m}$ yields that $\gamma_n(\hat{f}_{\hat{m}}) + \text{pen}(\hat{m}) \leq \gamma_n(f_m) + \text{pen}(m)$. But for all t, s,

$$\begin{aligned}\gamma_n(t) - \gamma_n(s) &= \|t\|^2 - \|s\|^2 - \frac{2}{n}\sum_{i=1}^n (t-s)(X_i) \\ &= \|t-f\|^2 - \|s-f\|^2 - 2\nu_n(t-s) \\ &= \|t-f\mathbb{1}_{A_1}\|^2 - \|s-f\mathbb{1}_{A_1}\|^2 - 2\nu_n(t-s)\end{aligned}$$

where $\nu_n(t) = (1/n)\sum_{i=1}^n [t(X_i) - < t, f >]$. This leads to

$$\|\hat{f}_{\hat{m}} - f\|_{A_1}^2 \leq \|f_m - f\|_{A_1}^2 + 2\nu_n(\hat{f}_{\hat{m}} - f_m) + \text{pen}(m) - \text{pen}(\hat{m}). \tag{1.17}$$

Remark 1.2 *If t is deterministic, $\nu_n(t)$ can actually be written $\nu_n(t) = (1/n)\sum_{i=1}^n [t(X_i) - \mathbb{E}(t(X_i))]$.*

We set $B(m,m') = \{t \in S_m + S_{m'}, \|t\| = 1\}$. Let us write now

$$2\nu_n(\hat{f}_{\hat{m}} - f_m) = 2\|\hat{f}_{\hat{m}} - f_m\| \nu_n\left(\frac{\hat{f}_{\hat{m}} - f_m}{\|\hat{f}_{\hat{m}} - f_m\|}\right)$$
$$\leq 2\|\hat{f}_{\hat{m}} - f_m\| \sup_{t \in B(m,\hat{m})} \nu_n(t) \leq \frac{1}{5}\|\hat{f}_{\hat{m}} - f_m\|^2 + 5 \sup_{t \in B(m,\hat{m})} \nu_n(t)^2$$

by using inequality $2xy \leq \frac{1}{5}x^2 + 5y^2$. Thus,

$$2\mathbb{E}|\nu_n(\hat{f}_{\hat{m}} - f_m)| \leq \frac{1}{5}\mathbb{E}\|\hat{f}_{\hat{m}} - f_m\|^2 + 5\mathbb{E}(\sup_{t \in B(m,\hat{m})} \nu_n(t)^2). \tag{1.18}$$

Consider decomposition (1.12) of $\nu_n(t)$ again and let

$$Z_n(t) = \frac{1}{n} \sum_{j=1+\tau(1)}^{\tau(l_n)} [t(X_i) - < t, f >]. \tag{1.19}$$

Since $|\nu_n^{(3)}(t)| \leq |Z_n(t)|$, we can write

$$\sup_{t \in B(m,\hat{m})} \nu_n^{(3)}(t)^2 \;\leq\; p(m,\hat{m}) + \sum_{m' \in \mathcal{M}_n} [\sup_{t \in B(m,m')} Z_n(t)^2 - p(m,m')]_+$$

where $p(.,.)$ is a function to be specified later. Then, the bound (1.13) combined with M1, (1.17) and (1.18) give

$$\begin{aligned} \mathbb{E}\|\hat{f}_{\hat{m}} - f\|_{A_1}^2 \;\leq\; & \|f_m - f\|_{A_1}^2 + \frac{1}{5}\mathbb{E}\|\hat{f}_{\hat{m}} - f_m\|^2 + 160\phi_0^2 \frac{\mathbb{E}(\tau^2) + \mu(\mathfrak{A})\mathbb{E}_{\mathfrak{A}}(\tau^4)}{n} \\ & + 20 \sum_{m' \in \mathcal{M}_n} \mathbb{E}[\sup_{t \in B(m,m')} Z_n(t)^2 - p(m,m')]_+ \\ & + \mathbb{E}(20p(m,\hat{m}) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m})). \end{aligned}$$

We choose $p(m,m')$ such that $20p(m,m') \leq \mathrm{pen}(m) + \mathrm{pen}(m')$. Thus $20p(m,\hat{m}) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) \leq 2\mathrm{pen}(m)$. Let

$$W(m,m') = [\sup_{t \in B(m,m')} Z_n^2(t) - p(m,m')]_+. \tag{1.20}$$

We use now the inequality $\frac{1}{5}(x+y)^2 \leq \frac{1}{3}x^2 + \frac{1}{2}y^2$ to deduce

$$\mathbb{E}\|\hat{f}_{\hat{m}} - f\|_{A_1}^2 \leq \frac{1}{3}\mathbb{E}\|\hat{f}_{\hat{m}} - f\|_{A_1}^2 + \frac{3}{2}\|f_m - f\|_{A_1}^2 + 20 \sum_{m' \in \mathcal{M}_n} \mathbb{E}W(m,m') + 2\mathrm{pen}(m) + \frac{C}{n}$$

and thus

$$\mathbb{E}\|\hat{f}_{\hat{m}} - f\|_{A_1}^2 \leq \frac{9}{4}\|f_m - f\|_{A_1}^2 + 30 \sum_{m' \in \mathcal{M}_n} \mathbb{E}W(m,m') + 3\mathrm{pen}(m) + \frac{3C}{2n}.$$

We need now to bound $\mathbb{E}W(m,m')$ to complete the proof. Proposition 1.2 below implies

$$\mathbb{E}W(m,m') \leq K' e^{-D_{m'}} (\phi_0 \vee 1)^2 K_3 \frac{1 + K_2\|f\|_{\infty,A_1}}{n}$$

where K' is a numerical constant and K_2, K_3 depend on the chain and with

$$p(m,m') \;=\; K \frac{\dim(S_m + S_{m'})}{n} (\phi_0 \vee 1)^2 K_3 (1 + K_2\|f\|_{\infty,A_1}). \tag{1.21}$$

The notation $a \vee b$ means $\max(a, b)$.

Assumption M3 yields $\sum_{m' \in \mathcal{M}_n} e^{-D_{m'}} \leq \sum_{k \geq 1} e^{-k} = 1/(e-1)$. Thus, by summation on m' in $\mathcal{M}_n$

$$\sum_{m' \in \mathcal{M}_n} \mathbb{E}W(m, m') \leq K' \frac{1}{e-1} (\phi_0 \vee 1)^2 K_3 \frac{1 + K_2 \|f\|_{\infty, A_1}}{n}.$$

It remains to specify the penalty, which has to satisfy $20p(m, m') \leq \text{pen}(m) + \text{pen}(m')$. The value of $p(m, m')$ is given by (1.21), so we set

$$\text{pen}(m) \geq 20K \frac{D_m}{n} (\phi_0 \vee 1)^2 K_3 (1 + K_2 \|f\|_{\infty, A_1}).$$

Finally

$$\forall m \in \mathcal{M}_n \qquad \mathbb{E}\|\hat{f}_{\hat{m}} - f\|_{A_1}^2 \leq 3\|f_m - f\|_{A_1}^2 + 3\text{pen}(m) + \frac{C_1}{n}$$

where C_1 depends on $\phi_0, \|f\|_{\infty, A_1}, \mu(\mathfrak{A}), \mathbb{E}_\mu(\tau^2), \mathbb{E}_{\mathfrak{A}}(\tau^4), K_2, K_3$. Since it is true for all m, we obtain the result.

Second step: We do not suppose any more that (X_n) has an atom.

The Nummelin splitting technique allows us to create the chain (X_n^*) and to define $\gamma_n^*(t)$ and $\hat{f}_m^*$ as above by (1.15),(1.16). Set now

$$\hat{m}^* = \arg\min_{m \in \mathcal{M}_n} [\gamma_n^*(\hat{f}_m^*) + \text{pen}(m)]$$

and $\tilde{f}^* = \hat{f}_{\hat{m}^*}^*$. The property P1 in Section 1.6.1 gives $\mathbb{E}\|f - \tilde{f}\|_{A_1}^2 = \mathbb{E}\|f - \tilde{f}^*\|_{A_1}^2$. The split chain having an atom, we can use the first step to deduce

$$\mathbb{E}\|f - \tilde{f}^*\|_{A_1}^2 \leq 3 \inf_{m \in \mathcal{M}_n} \{d^2(f\mathbb{1}_{A_1}, S_m) + \text{pen}(m)\} + \frac{C_1}{n}.$$

And then the result is valid when replacing $\tilde{f}^*$ by $\tilde{f}$.

□

Proposition 1.2 *Let (X_n) be a Markov chain which satisfies H1-H2a-H3a and $(S_m)_{m \in \mathcal{M}_n}$ be a collection of models satisfying M1–M3. We suppose that (X_n) has an atom $\mathfrak{A}$. Let $Z_n(t)$ and $W(m, m')$ defined by (1.19) and (1.20) with*

$$p(m, m') = K \frac{\dim(S_m + S_{m'})}{n} (\phi_0 \vee 1)^2 \frac{1 + \|f\|_{\infty, A_1} \mathbb{E}_{\mathfrak{A}}(s^\tau)}{(\log s)^2}$$

(where K is a numerical constant and s is a real depending on the chain). Then

$$\mathbb{E}W(m, m') \leq K' e^{-D_{m'}} (\phi_0 \vee 1)^2 \frac{1 + \|f\|_{\infty, A_1} \mathbb{E}_{\mathfrak{A}}(s^\tau)}{(\log s)^2 n}.$$

Proof of Proposition 1.2: We can write $Z_n(t) = (1/n)\sum_{j=1}^{l_n-1} S_j(t)$ where $S_j(t)$ is defined by (1.14). According to Lemma 1.1:

$$\mathbb{E}_\mu|S_j(t)|^m \leq (2\|t\|_\infty)^{m-2}\|f\|_{\infty,A_1}\|t\|^2\mathbb{E}_{\mathfrak{A}}(\tau^m).$$

Now, we use condition H2a(ii) of geometric ergodicity. The proof of Theorem 15.4.2 in Meyn and Tweedie (1993) shows that $\mathfrak{A}$ is a Kendall set, i.e. there exists $s > 1$ (depending on $\mathfrak{A}$) such that $\sup_{x\in\mathfrak{A}}\mathbb{E}_x(s^\tau) < \infty$. Then $\mathbb{E}_{\mathfrak{A}}(\tau^m) \leq [m!/(\log s)^m]\mathbb{E}_{\mathfrak{A}}(s^\tau)$. Indeed

$$\begin{aligned}\mathbb{E}_{\mathfrak{A}}(\tau^m) &= \int_0^\infty mx^{m-1}P_{\mathfrak{A}}(\tau > x)dx \\ &\leq \int_0^\infty mx^{m-1}s^{-x}\mathbb{E}_{\mathfrak{A}}(s^\tau)dx = \frac{m!}{(\log s)^m}\mathbb{E}_{\mathfrak{A}}(s^\tau).\end{aligned}$$

Thus

$$\forall m \geq 2 \qquad \mathbb{E}_\mu|S_j(t)|^m \leq m!\left(\frac{2\|t\|_\infty}{\log s}\right)^{m-2}\frac{\|f\|_{\infty,A_1}\|t\|^2}{(\log s)^2}\mathbb{E}_{\mathfrak{A}}(s^\tau). \tag{1.22}$$

We use now the following inequality (see Petrov (1975) p.49):

$$P\big(\max_{1\leq l\leq n}\sum_{j=1}^{l} S_j(t) \geq y\big) \leq 2P\big(\sum_{j=1}^{n} S_j(t) \geq y - \sqrt{2B_n}\big)$$

where $B_n \geq \sum_{j=1}^n \mathbb{E}S_j(t)^2$. The inequality (1.22) gives us $B_n = 2n\dfrac{\|f\|_{\infty,A_1}\|t\|^2}{(\log s)^2}\mathbb{E}_{\mathfrak{A}}(s^\tau)$ and

$$P\big(\sum_{j=1}^{l_n-1} S_j(t) \geq y\big) \leq P\big(\max_{1\leq l\leq n}\sum_{j=1}^{l} S_j(t) \geq y\big) \leq 2P\big(\sum_{j=1}^{n} S_j(t) \geq y - 2\sqrt{n}\|t\|M/\log s\big)$$

where $M^2 = \|f\|_{\infty,A_1}\mathbb{E}_{\mathfrak{A}}(s^\tau)$. We use then the Bernstein inequality given by Birgé and Massart (1998):

$$P(\sum_{j=1}^{n} S_j(t) \geq n\varepsilon) \leq e^{-nx}$$

with $\varepsilon = \dfrac{2\|t\|_\infty}{\log s}x + \dfrac{2\|t\|M}{\log s}\sqrt{x}$. Indeed, according to (1.22),

$$\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}|S_j(t)|^m \leq \frac{m!}{2}\left(\frac{2\|t\|_\infty}{\log s}\right)^{m-2}\left(\frac{\sqrt{2}\|t\|M}{\log s}\right)^2.$$

Finally

$$P\left(Z_n(t) \geq \frac{2}{\log s}\left[\|t\|_\infty x + M\|t\|\sqrt{x} + M\|t\|/\sqrt{n}\right]\right) \leq 2e^{-nx}. \tag{1.23}$$

We will now use a chaining technique used in Barron *et al.* (1999). Let us recall first the following lemma (Lemma 9 p.400 in Barron *et al.* (1999), see also Proposition 1 in Birgé and Massart (1998)).

Lemma 1.2 *Let $\bar{S}$ a subspace of L^2 with dimension D spanned by an orthonormal basis $(\varphi_\lambda)_{\lambda\in\Lambda}$. Let*

$$r = \frac{1}{\sqrt{D}}\sup_{\beta\neq 0}\frac{\|\sum_{\lambda\in\Lambda}\beta_\lambda\varphi_\lambda\|_\infty}{\sup_{\lambda\in\Lambda}|\beta_\lambda|}.$$

Then, for all $\delta > 0$, we can find a countable set $T \subset \bar{S}$ and a mapping π from $\bar{S}$ to T such that :

- *for all ball $\mathcal{B}$ with radius $\sigma \geq 5\delta$*

$$|T \cap \mathcal{B}| \leq (5\sigma/\delta)^D \tag{1.24}$$

- *$\|u - \pi(u)\| \leq \delta$, $\forall u \in \bar{S}$ and $\sup_{u\in\pi^{-1}(t)}\|u - t\|_\infty \leq r\delta, \quad \forall t \in T$.*

We apply this lemma to the subpace $S_m + S_{m'}$ with dimension $D_m \vee D_{m'}$ denoted by $D(m')$ and $r = r(m')$ defined by

$$r(m') = \frac{1}{\sqrt{D(m')}}\sup_{\beta\neq 0}\frac{\|\sum_{\lambda\in\Lambda(m')}\beta_\lambda\varphi_\lambda\|_\infty}{\sup_{\lambda\in\Lambda(m')}|\beta_\lambda|}$$

where $(\varphi_\lambda)_{\lambda\in\Lambda(m')}$ is an orthonormal basis of $S_m + S_{m'}$. Notice that this quantity satisfy $\phi_{m"} \leq r(m') \leq \sqrt{D(m')}\phi_{m"}$ where $m"$ is such that $S_m + S_{m'} = S_{m"}$ and then, using M2,

$$r(m') \leq \phi_0\sqrt{D(m')}.$$

We consider $\delta_0 \leq 1/5$, $\delta_k = \delta_0 2^{-k}$, and the $T_k = T \cap B(m, m')$ where T is defined by Lemma 1.2 with $\delta = \delta_k$ and $B(m, m')$ is the unit ball of $S_m + S_{m'}$. Inequality (1.24) gives us $|T \cap B(m, m')| \leq (5/\delta_k)^{D(m')}$. By letting $H_k = \log(|T_k|)$, we obtain

$$H_k \leq D(m')[\log(\frac{5}{\delta_0}) + k\log 2]. \tag{1.25}$$

Thus, for all u in $B(m, m')$, we can find a sequence $\{u_k\}_{k\geq 0}$ with $u_k \in T_k$ such that $\|u - u_k\| \leq \delta_k$ and $\|u - u_k\|_\infty \leq r(m')\delta_k$. Hence, we have the following decomposition:

$$u = u_0 + \sum_{k=1}^{\infty}(u_k - u_{k-1})$$

with $\|u_0\| \leq 1$ and $\|u_0\|_\infty \leq \phi_0\sqrt{D(m')}\|u_0\| \leq \phi_0\sqrt{D(m')}$ and for all $k \geq 1$,

$$\begin{aligned}\|u_k - u_{k-1}\| &\leq \delta_k + \delta_{k-1} = 3\delta_{k-1}/2,\\ \|u_k - u_{k-1}\|_\infty &\leq 3r(m')\delta_{k-1}/2 \leq 3\phi_0\sqrt{D(m')}\delta_{k-1}/2.\end{aligned}$$

Then

$$P(\sup_{u\in B(m,m')} Z_n(u) > \eta) = P(\exists (u_k)_{k\geq 0} \in \prod_{k\geq 0} T_k, Z_n(u_0) + \sum_{k=1}^{\infty} Z_n(u_k - u_{k-1}) > \eta_0 + \sum_{k=1}^{\infty} \eta_k)$$
$$\leq \sum_{u_0\in T_0} P(Z_n(u_0) > \eta_0) + \sum_{k=1}^{\infty} \sum_{\substack{u_k\in T_k \\ u_{k-1}\in T_{k-1}}} P(Z_n(u_k - u_{k-1}) > \eta_k)$$

with $\eta_0 + \sum_{k=1}^{\infty} \eta_k \leq \eta$. We use the exponential inequality (1.23) to obtain

$$\sum_{u_0\in T_0} P(Z_n(u_0) > \eta_0) \leq 2e^{H_0 - nx_0}$$
$$\sum_{\substack{u_k\in T_k \\ u_{k-1}\in T_{k-1}}} P(Z_n(u_k - u_{k-1}) > \eta_k) \leq 2e^{H_k + H_{k-1} - nx_k}$$

by choosing $\begin{cases} \eta_0 = \dfrac{2}{\log s}\left(\phi_0\sqrt{D(m')}x_0 + M\sqrt{x_0} + \dfrac{M}{\sqrt{n}}\right) \\ \eta_k = \dfrac{3}{\log s}\left(\phi_0\sqrt{D(m')}\delta_{k-1}x_k + M\delta_{k-1}\sqrt{x_k} + \dfrac{M\delta_{k-1}}{\sqrt{n}}\right). \end{cases}$

Let us choose now the $(x_k)_{k\geq 0}$ such that $nx_0 = H_0 + D_{m'} + v$ and for $k \geq 1$,

$$nx_k = H_{k-1} + H_k + kD_{m'} + D_{m'} + v.$$

Thus

$$P(\sup_{u\in B(m,m')} Z_n(u) > \eta) \leq 2e^{-D_{m'}-v}(1 + \sum_{k\geq 1} e^{-kD_{m'}}) \leq 3.2e^{-D_{m'}-v}.$$

It remains to bound $\sum_{k=0}^{\infty} \eta_k$:

$$\sum_{k=0}^{\infty} \eta_k \leq \frac{1}{(\log s)}(A_1 + A_2 + A_3).$$

where $\begin{cases} A_1 = \phi_0\sqrt{D(m')}(2x_0 + 3\sum_{k=1}^{\infty} \delta_{k-1}x_k), \\ A_2 = 2M\sqrt{x_0} + 3M\sum_{k=1}^{\infty} \delta_{k-1}\sqrt{x_k}, \\ A_3 = 2\dfrac{M}{\sqrt{n}} + \sum_{k=1}^{\infty} \dfrac{3M\delta_{k-1}}{\sqrt{n}}. \end{cases}$

• Regarding the third term, just write

$$A_3 \;=\; \frac{M}{\sqrt{n}}\Big(2+3\sum_{k=1}^{\infty}\delta_{k-1}\Big)=\frac{M}{\sqrt{n}}(6\delta_0+2)\le c_1(\delta_0)\frac{M}{\sqrt{n}}$$

with $c_1(\delta_0) \;=\; 6\delta_0+2$.

• Let us bound the first term. First, recall that $D(m')\le\sqrt{n}$ and then

$$A_1 \;\le\; \phi_0\sqrt{\frac{n}{D(m')}}\left(2\frac{H_0+D_{m'}+v}{n}+3\sum_{k=1}^{\infty}\delta_{k-1}\frac{H_{k-1}+H_k+kD_{m'}+D_{m'}+v}{n}\right).$$

Observing that $\sum_{k=1}^{\infty}\delta_{k-1}=2\delta_0$ and $\sum_{k=1}^{\infty}k\delta_{k-1}=4\delta_0$ and using (1.25), we get

$$A_1 \;\le\; c_1(\delta_0)\phi_0\frac{v}{\sqrt{nD(m')}}+c_2(\delta_0)\phi_0\sqrt{\frac{D(m')}{n}}$$

with $c_2(\delta_0)=c_1(\delta_0)+\log(5/\delta_0)(2+12\delta_0)+6\delta_0(2+3\log 2)$.

• To bound the second term, we use the Schwarz inequality and the inequality $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$. We obtain

$$A_2 \;\le\; c_1(\delta_0)M\sqrt{\frac{v}{n}}+c_3(\delta_0)M\sqrt{\frac{D(m')}{n}}$$

with $c_3(\delta_0)=2\sqrt{1+\log(5/\delta_0)}+3\sqrt{2\delta_0}\sqrt{6\delta_0(1+\log 2)+4\delta_0\log(5/\delta_0)}$.

We get so

$$\begin{aligned}(\sum_{k=0}^{\infty}\eta_k) \;&\le\; \left(\frac{\phi_0\vee 1}{\log s}\right)c_1\left(\frac{v}{\sqrt{nD(m')}}+M\sqrt{\frac{v}{n}}\right)\\ &\quad+\sqrt{\frac{D(m')}{n}}\left(\frac{\phi_0\vee 1}{\log s}\right)[c_2+c_3M+c_1M]\\ (\sum_{k=0}^{\infty}\eta_k)^2 \;&\le\; c_4(\delta_0)\left(\frac{\phi_0\vee 1}{\log s}\right)^2[\frac{v^2}{nD(m')}\vee M^2\frac{v}{n}]+c_5(\delta_0)\frac{D(m')}{n}\left(\frac{\phi_0\vee 1}{\log s}\right)^2(1+M)^2\end{aligned}$$

where $\begin{cases}c_4(\delta_0)= & 6c_1^2\\ c_5(\delta_0)= & (6/5)\sup(c_2,c_3+c_1)^2.\end{cases}$

Let us choose now $\delta_0=0.024$ and then $c_4=28$, $c_5=268$. Let $K_1=c_4(\phi_0\vee 1/\log s)^2$. Then

$$\eta^2=K_1[\frac{v^2}{nD(m')}\vee M^2\frac{v}{n}]+p(m,m')$$

where

$$p(m,m') = c_5(\phi_0 \vee 1)^2 \frac{D(m')}{n} \frac{1 + \|f\|_{\infty,A_1} \mathbb{E}_{\mathfrak{A}}(s^\tau)}{(\log s)^2}.$$

$$\begin{aligned} \text{We get} \qquad & P(\sup_{u \in B(m,m')} Z_n^2(u) > K_1[\frac{v^2}{nD(m')} \vee M^2 \frac{v}{n}] + p(m,m')) \\ = \; & P(\sup_{u \in B(m,m')} Z_n^2(u) > \eta^2) \\ \leq \; & P(\sup_{u \in B(m,m')} Z_n(u) > \eta) + P(\sup_{u \in B(m,m')} Z_n(u) < -\eta) \end{aligned}$$

Now

$$\begin{aligned} P(\sup_{u \in B(m,m')} Z_n(u) < -\eta) &\leq \sum_{u_0 \in T_0} P(Z_n(u_0) < -\eta_0) + \sum_{k=1}^{\infty} \sum_{\substack{u_k \in T_k \\ u_{k-1} \in T_{k-1}}} P(Z_n(u_k - u_{k-1}) < -\eta_k) \\ &\leq \sum_{u_0 \in T_0} P(Z_n(-u_0) > \eta_0) + \sum_{k=1}^{\infty} \sum_{\substack{u_k \in T_k \\ u_{k-1} \in T_{k-1}}} P(Z_n(-u_k + u_{k-1}) > \eta_k) \\ &\leq 3.2 e^{-D_{m'} - v}. \end{aligned}$$

Hence

$$P(\sup_{u \in B(m,m')} Z_n^2(u) > K_1[\frac{v^2}{nD(m')} \vee M^2 \frac{v}{n}] + p(m,m')) \leq 6.4 e^{-D_{m'} - v}. \tag{1.26}$$

We obtain then

$$\begin{aligned} &\mathbb{E}[\sup_{t \in B(m,m')} Z_n^2(t) - p(m,m')]_+ \leq \int_0^\infty P(\sup_{u \in B(m,m')} Z_n^2(u) > p(m,m') + z) dz \\ &\leq \int_0^{M^2 D(m')} P(\sup_{u \in B(m,m')} Z_n^2(u) > p(m,m') + K_1 M^2 \frac{v}{n}) K_1 \frac{M^2}{n} dv \\ &+ \int_{M^2 D(m')}^\infty P(\sup_{u \in B(m,m')} Z_n^2(u) > p(m,m') + K_1 \frac{v^2}{nD(m')}) K_1 \frac{2v}{nD(m')} dv \\ &\leq \frac{K_1}{n} \left[M^2 \int_0^\infty 6.4 e^{-D_{m'} - v} dv + \frac{2}{D(m')} \int_0^\infty 6.4 e^{-D_{m'} - v} v dv \right] \\ &\leq \frac{6.4 K_1}{n} e^{-D_{m'}} (M^2 + \frac{2}{D(m')}) \leq 12.8 K_1 e^{-D_{m'}} \frac{1 + M^2}{n}. \end{aligned}$$

By replacing M^2 by its value, we get so

$$\mathbb{E} W(m,m') \leq K' (\frac{\phi_0 \vee 1}{\log s})^2 e^{-D_{m'}} \frac{1 + \|f\|_{\infty,A_1} \mathbb{E}_{\mathfrak{A}}(s^\tau)}{n}$$

where K' is a numerical constant □

1.6.5 Proof of Corollary 1.2

According to Theorem 1.1, $\mathbb{E}\|\tilde{f}-f\|_{A_1}^2 \leq C_2 \inf_{m\in\mathcal{M}_n}\{d^2(f\mathbb{1}_{A_1}, S_m)+D_m/n\}$. Since $d^2(f\mathbb{1}_{A_1}, S_m)$ $= O(D_m^{-2\alpha})$ (see Lemma 12 in Barron *et al.* (1999)),

$$\mathbb{E}\|\tilde{f}-f\|_{A_1}^2 \leq C_3 \inf_{m\in\mathcal{M}_n}\left\{D_m^{-2\alpha}+\frac{D_m}{n}\right\}.$$

In particular, if m_0 is such that $D_{m_0} = \lfloor n^{\frac{1}{1+2\alpha}} \rfloor$, then

$$\mathbb{E}\|\tilde{f}-f\|_{A_1}^2 \leq C_3\left\{D_{m_0}^{-2\alpha}+\frac{D_{m_0}}{n}\right\} \leq C_4 n^{-\frac{2\alpha}{1+2\alpha}}.$$

The condition $D_m \leq \sqrt{n}$ allows this choice of m only if $\alpha > 1/2$. □

1.6.6 Proof of Theorem 1.2

The proof is identical to the one of Theorem 1.1. □

1.6.7 Proof of Corollary 1.3

It is sufficient to prove that $d_A(F, \mathbb{S}_m) = \inf_{T\in\mathbb{S}_m}\|F-T\|_A \leq CD_m^{-\alpha}$ if F belongs to $B_{2,\infty}^{\alpha}(A)$. It is done in the following lemma. □

Lemma 1.3 *Let F in the Besov space $B_{2,\infty}^{\alpha}(A)$. We consider the following spaces of dimension D^2 :*

- *S_1 is a space of piecewiwe polynomials of degree bounded by $s > \alpha - 1$ based on the regular partition with D^2 squares,*
- *S_2 is a space of of orthonormal wavelets of regularity $s > \alpha - 1$,*
- *S_3 is the space of trigonometric polynomials.*

Then, there exist positive constants C_i such that

$$d_A(F, S_i) \leq C_i D^{-\alpha} \quad \textit{for } i = 1, 2, 3.$$

Proof of Lemme 1.3: Let us recall the definition of $B_{2,\infty}^{\alpha}(A)$. Let

$$\Delta_h^r F(x,y) = \sum_{k=0}^{r}(-1)^{r-k}\binom{r}{k}F(x+kh_1, y+kh_2)$$

the rth difference operateur with step h and $\omega_r(F,t) = \sup_{|h|\leq t} \|\Delta_h^r F\|_A$ the rth modulus of smoothness of F. We say F is in the Besov space $B^\alpha_{2,\infty}(A)$ if $\sup_{t>0} t^{-\alpha}\omega_r(F,t) < \infty$ for $r = \lfloor \alpha \rfloor + 1$, or equivalently, for r an integer larger than α.

DeVore (1998) proved that $d_A(F, S_1) \leq C\omega_{s+1}(F, D^{-1})$, so

$$d_A(F, S_1) \leq CD^{-\alpha}.$$

For the wavelets case, we use the fact that f belongs to $B^\alpha_{2,\infty}(A)$ if and only if $\sup_{j\geq -1} 2^{j\alpha}\|\beta_j\| < \infty$ (see Meyer (1990) chapter 6, section 10). If F_D is the orthogonal projection of F on S_2,

$$\|F - F_D\|_A^2 = \sum_{j>m}\sum_{k,l} |\beta_{jkl}|^2 \leq C\sum_{j>m} 2^{-2j\alpha} \leq C'D^{-j\alpha}$$

where m is such that $2^m = D$.

For the trigonometric case, it is proved in Nikol′skiĭ (1975) (p. 191 and 200) that $d_A(F, S_3) \leq C\omega_{s+1}(F, D^{-1})$ so that $d_A(F, S_3) \leq C'D^{-\alpha}$. □

1.6.8 Proof of Theorem 1.3

Let us prove first the first item. Let $E_n = \{\|f - \tilde{f}\|_{\infty,A_1} \leq f_0/2\}$ and E_n^c its complement. On E_n and for $x \in A_1$, $\tilde{f}(x) = \tilde{f}(x) - f(x) + f(x) \geq f_0/2$. Moreover

$$\|\tilde{F}\|^2 = \sum_{\lambda,\mu}\left(\frac{1}{n-1}\sum_{i=1}^{n-1}\varphi_\lambda(X_i)\varphi_\mu(X_{i+1})\right)^2 \leq \|\sum_\lambda \varphi_\lambda^2\|_\infty^2 \leq \phi_0^4 D_{\hat{M}}^2$$

so that $\|\tilde{F}\|_\infty \leq \phi_0^2 D_{\hat{M}}\|\tilde{F}\| \leq \phi_0^4 D_{\hat{M}}^2$. Then, for n large enough, for all x, y,

$$\left|\frac{\tilde{F}(x,y)}{\tilde{f}(x)}\right| \leq \frac{\phi_0^4}{f_0/2}\sqrt{n} \leq k_n$$

and $\tilde{\Pi}(x,y) = \dfrac{\tilde{F}(x,y)}{\tilde{f}(x)}$. For all $(x,y) \in A$,

$$\begin{aligned}
|\tilde{\Pi}(x,y) - \Pi(x,y)|^2 &\leq \left|\frac{\tilde{F}(x,y) - \tilde{f}(x)\Pi(x,y)}{\tilde{f}(x)}\right|^2 \mathbb{1}_{E_n} + (\|\tilde{\Pi}\|_\infty + \|\Pi\|_{\infty,A})^2\mathbb{1}_{E_n^c} \\
&\leq \frac{|\tilde{F}(x,y) - F(x,y) + \Pi(x,y)(f(x) - \tilde{f}(x))|^2}{f_0^2/4} \\
&\quad +(k_n + \|\Pi\|_{\infty,A})^2\mathbb{1}_{E_n^c} \\
\mathbb{E}\|\Pi - \tilde{\Pi}\|_A^2 &\leq \frac{8}{f_0^2}[\mathbb{E}\|F - \tilde{F}\|_A^2 + \|\Pi\|_{\infty,A}^2\mathbb{E}\|f - \tilde{f}\|_{A_1}^2] + (k_n + \|\Pi\|_{\infty,A})^2|A|P(E_n^c).
\end{aligned}$$

It remains to bound $P(E_n^c)$. To do this, we observe that

$$\|f - \tilde{f}\|_{\infty,A_1} \leq \|f - f_{\hat{m}}\|_{\infty,A_1} + \|f_{\hat{m}} - \hat{f}_{\hat{m}}\|_\infty.$$

Let $\gamma = \alpha - \frac{1}{2}$, then $B^\alpha_{2,\infty}(A_1) \subset B^\gamma_{\infty,\infty}(A_1)$ (see DeVore and Lorentz (1993) p.182). Thus f belongs to $B^\gamma_{\infty,\infty}(A_1)$ and Lemma 12 in Barron *et al.* (1999) gives

$$\|f - f_{\hat{m}}\|_{\infty,A_1} \leq C D_{\hat{m}}^{-\gamma} \leq C(\log n)^{-\gamma}.$$

Thus $\|f - f_{\hat{m}}\|_{\infty,A_1}$ decreases to 0 and $\|f - f_{\hat{m}}\|_{\infty,A_1} \leq f_0/4$ for n large enough. So

$$P(E_n^c) \leq P(\|f_{\hat{m}} - \hat{f}_{\hat{m}}\|_\infty > \frac{f_0}{4}).$$

But $\|f_{\hat{m}} - \hat{f}_{\hat{m}}\|_\infty \leq \phi_0\sqrt{D_{\hat{m}}}\|f_{\hat{m}} - \hat{f}_{\hat{m}}\|$and $\|f_{\hat{m}} - \hat{f}_{\hat{m}}\|^2 = \sum_{\lambda\in\Lambda_{\hat{m}}} \nu_n^2(\varphi_\lambda) = \sup_{t\in B_{\hat{m}}} |\nu_n(t)|^2$. Thus,

$$\begin{aligned}
P(E_n^c) &\leq P(\sup_{t\in B_{\hat{m}}} |\nu_n(t)|^2 > \frac{f_0^2}{16\phi_0^2 D_{\hat{m}}}) \\
&\leq \sup_{m\in\mathcal{M}_n} P(\sup_{t\in B_m} |\nu_n(t)|^2 > \frac{f_0^2}{16\phi_0^2 D_m}) \\
&\leq \sup_{m\in\mathcal{M}_n} P(\sup_{t\in B_m} |\nu_n^{(1)}(t) + \nu_n^{(2)}(t) + \nu_n^{(4)}(t)|^2 > \frac{f_0^2}{32\phi_0^2 D_m}) \\
&\quad + \sup_{m\in\mathcal{M}_n} P(\sup_{t\in B_m} |Z_n(t)|^2 > \frac{f_0^2}{32\phi_0^2 D_m})
\end{aligned}$$

We need then to bound two terms. For the first term, we use a Markov inequality:

$$\begin{aligned}
&P(\sup_{t\in B_m} |\nu_n^{(1)}(t) + \nu_n^{(2)}(t) + \nu_n^{(4)}(t)|^2 > \frac{f_0^2}{32\phi_0^2 D_m}) \\
&\leq \left(\frac{32\phi_0^2 D_m}{f_0^2}\right)^3 \mathbb{E}\left(\left(\sup_{t\in B_m} |\nu_n^{(1)}(t) + \nu_n^{(2)}(t) + \nu_n^{(4)}(t)|^2\right)^3\right) \\
&\leq \left(\frac{32\phi_0^2 D_m}{f_0^2}\right)^3 3^5 \mathbb{E}\left(\sup_{t\in B_m} |\nu_n^{(1)}(t)|^6 + |\nu_n^{(2)}(t)|^6 + |\nu_n^{(4)}(t)|^6\right)
\end{aligned}$$

Reasoning as for inequality (1.12), we obtain

$$\mathbb{E}\left(\sup_{t\in B_m} |\nu_n^{(1)}(t)|^6 + |\nu_n^{(2)}(t)|^6 + |\nu_n^{(4)}(t)|^6\right) \leq 2^7\phi_0^6[\mathbb{E}_\mu(\tau^6) + \mu(\mathfrak{A})\mathbb{E}_\mathfrak{A}(\tau^8)]\frac{D_m^3}{n^6}$$

and then

$$\begin{aligned}
&\sup_{m\in\mathcal{M}_n} P(\sup_{t\in B_m} |\nu_n^{(1)}(t) + \nu_n^{(2)}(t) + \nu_n^{(4)}(t)|^2 > \frac{f_0^2}{32\phi_0^2 D_m}) \\
&\leq \sup_{m\in\mathcal{M}_n} \left(\frac{32\phi_0^2 D_m}{f_0^2}\right)^3 3^5 2^7 \phi_0^6[\mathbb{E}_\mu(\tau^6) + \mu(\mathfrak{A})\mathbb{E}_\mathfrak{A}(\tau^8)]\frac{D_m^3}{n^6} \\
&\leq \sup_{m\in\mathcal{M}_n} C\frac{D_m^6}{n^6} \leq \frac{C}{n^3(\log n)^6}.
\end{aligned}$$

Besides, for all v, using (1.26),

$$P(\sup_{t\in B_m}|Z_n(t)|^2 > K_1[\frac{v^2}{nD_m}\vee M^2\frac{v}{n}]+p(m,m))\leq 6.4e^{-D_m-v}$$

with $p(m,m)=c_5(\phi_0\vee 1)^2(D_m/n)(1+\|f\|_{\infty,A_1}\mathbb{E}_{\mathfrak{A}}(s^\tau))/(\log s)^2$. Thus, letting $c_6=c_5(\phi_0\vee 1)^2(1+\|f\|_{\infty,A_1}\mathbb{E}_{\mathfrak{A}}(s^\tau))/(\log s)^2$,

$$P(\sup_{t\in B_m}|Z_n(t)|^2 > K_1[\frac{v^2}{nD_m}\vee M^2\frac{v}{n}]+c_6\frac{D_m}{n})\leq 6.4e^{-D_m-v}$$

Let now $v=n^{1/4}$, v verifies (for n large enough)

$$K_1[\frac{v^2}{nD_m}\vee M^2\frac{v}{n}]+c_6\frac{D_m}{n}\leq\frac{f_0^2}{32\phi_0^2D_m}$$

since $D_m\leq\sqrt{n}/(\log n)$. The previous inequality gives then

$$\sup_{m\in\mathcal{M}_n}P(\sup_{t\in B_m}|Z_n(t)|^2>\frac{f_0^2}{32\phi_0^2D_m})\leq 6.4\sup_{m\in\mathcal{M}_n}e^{-D_m-v}\leq C'e^{-n^{1/4}}$$

Finally

$$P(E_n^c)\leq\frac{C}{n^3(\log n)^6}+C'e^{-n^{1/4}}=o(n^{-3}).$$

And then, for n large enough, $(k_n+\|\Pi\|_{\infty,A})^2P(E_n^c)=o(k_n^2n^{-3})$. So, since $k_n=n$, $(k_n+\|\Pi\|_{\infty,A})^2P(E_n^c)=o(n^{-1})$.

Following result in Theorem 1.3 is provided by using Corollary 1.2 and 1.3. □

Chapitre 2

Estimation de la densité de transition par contraste moindres carrés

Version modifiée de l'article *Adaptive estimation of the transition density of a Markov chain* paru aux Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, vol. 43 (5).

2.1 Introduction

We consider (X_i) a homogeneous Markov chain. The purpose of this chapter is to estimate the transition density of such a chain. This quantity allows to comprehend the form of dependence between variables and is defined by $\Pi(x,y)dy = P(X_{i+1} \in dy | X_i = x)$. It enables also to compute other quantities, like $\mathbb{E}[g(X_{i+1})|X_i = x]$ for example. As many authors, we choose for this a nonparametric approach. Roussas (1969) first studies an estimator of the transition density of a Markov chain. He proves the consistency and the asymptotic normality of a kernel estimator for chains satisfying a strong condition known as Doeblin's hypothesis. In Bosq (1973), an estimator by projection is studied in a mixing framework and the consistence is also proved. Basu and Sahoo (1998) establish a Berry-Essen inequality for a kernel estimator under an assumption introduced by Rosenblatt, weaker than the Doeblin's hypothesis. Athreya and Atuncar (1998) improve the result of Roussas since they only need the Harris recurrence of the Markov chain. Other authors are interested in the estimation of the transition density in the non-stationary case: Doukhan and Ghindès (1983) bound the integrated risk for any initial distribution. In Hernández-Lerma *et al.* (1988), recursive estimators for a non-stationary Markov chain are described. More recently, Clémençon (2000) computes the lower bound of the minimax L^p risk and describes a quotient estimator using wavelets. In the first chapter of this thesis, we found an estimator by projection with model selection that reaches the optimal rate of convergence.

All these authors have estimated Π by observing that $\Pi = F/f$ where F is the density of (X_i, X_{i+1}) and f the stationary density. If $\hat{F}$ and $\hat{f}$ are estimators of F and f, then an estimator of Π can be obtained by writing $\hat{\Pi} = \hat{F}/\hat{f}$. But this method has the drawback that the resulting rate of convergence depends on the regularity of f. And the stationary density f can be less regular than the transition density.

The aim here is to find an estimator $\tilde{\Pi}$ of Π from the observations $X_1, \ldots, X_{n+1}$ such that the order of the L^2 risk depends only on the regularity of Π and is optimal.

Clémençon (2000) introduces an estimation procedure based on an analogy with the regression framework using the thresholding of wavelets coefficients for regular Markov chains. We propose in this chapter an other method based on regression, which improves the rate and has the advantage to be really computable. Indeed, this method allows to reach the optimal rate of convergence, without the logarithmic loss obtained by Clémençon (2000) and can be applied to β-mixing Markov chains (the notion of "regular" Markov chains in Clémençon (2000) is equivalent to ϕ-mixing and is then a stronger assumption). We use model selection via penalization as described in Barron *et al.* (1999) with a new contrast inspired by the classical regression contrast. To deal with the dependence we use auxiliary variables X_i^* as in Viennet (1997). But contrary to most cases in such estimation procedure, our penalty does not contain any mixing term and is entirely computable.

In addition, we consider transition densities belonging to anisotropic Besov spaces, i.e. with different regularities with respect to the two directions. Our projection spaces

(piecewise polynomials, trigonometric polynomials or wavelets) have different dimensions in the two directions and the procedure selects automatically both well fitted dimensions. A lower bound for the rate of convergence on anisotropic Besov balls is proved, which shows that our estimation procedure is optimal in a minimax sense.

The chapter is organized as follows. First, we present the assumptions on the Markov chain and on the collections of models. We also give examples of chains and models. Section 2.3 is devoted to estimation procedure and the link with classical regression. The bound on the empirical risk is established in Section 2.4 and the L^2 control is studied in Section 2.5. We compute both upper bound and lower bound for the mean integrated squared error. In Section 2.6, some simulation results are given. The proofs are gathered in Section 2.7.

2.2 Assumptions

2.2.1 Assumptions on the Markov chain

We consider an irreducible Markov chain (X_n) taking its values in the real line $\mathbb{R}$. We suppose that (X_n) is positive recurrent, i.e. it admits a stationary probability measure μ. We assume that the distribution μ has a density f with respect to the Lebesgue measure and that the transition kernel $P(x, A) = P(X_{i+1} \in A | X_i = x)$ has also a density, denoted by Π. This transition density Π is estimated on a compact set $A = A_1 \times A_2$ only. More precisely, the Markov process is supposed to satisfy the following assumptions:

H1 (X_n) is irreducible and positive recurrent and stationary.

H2b The chain is geometrically β-mixing ($\beta_q \leq e^{-\theta q}$), or arithmetically β-mixing ($\beta_q \leq q^{-\theta}$).

H3a The stationary density f verifies $\|f\|_{\infty,A_1} := \sup_{x \in A_1} |f(x)| < \infty$.

H3b The transition density Π is bounded on A, i.e. $\|\Pi\|_{\infty,A} := \sup_{(x,y)\in A} |\Pi(x, y)| < \infty$.

H4 There exists a positive real f_0 such that, for all x in A_1, $f(x) \geq f_0$.

Since (X_i) is a stationary Markov chain, the β-mixing is very explicit, the mixing coefficients can be written:

$$\beta_q = \int \|P^q(x, .) - \mu\|_{TV} f(x) dx \tag{2.1}$$

where $\|.\|_{TV}$ is the total variation norm (see Doukhan (1994)).

Notice that we distinguish the sets A_1 and A_2 in this work because the two directions x and y in $\Pi(x, y)$ do not play the same role, but in practice A_1 and A_2 will be equal and identical or close to the value domain of the chain.

Many processes verify the previous assumptions, as autoregressive processes or diffusions. A list of such chains is given in Chapter 1 (it is valid since the Assumption H2a is stronger than H2b).

2.2.2 Assumptions on the models

In order to estimate Π, we need to introduce a collection $\{S_m, m \in \mathcal{M}_n\}$ of spaces, that we call models. For each $m = (m_1, m_2)$, S_m is a space of functions with support in A defined from two spaces: F_{m_1} and H_{m_2}. F_{m_1} is a subspace of $(L^2 \cap L^\infty)(\mathbb{R})$ spanned by an orthonormal basis $(\varphi_j^m)_{j\in J_m}$ with $|J_m| = D_{m_1}$ such that, for all j, the support of φ_j^m is included in A_1. In the same way H_{m_2} is a subspace of $(L^2 \cap L^\infty)(\mathbb{R})$ spanned by an orthonormal basis $(\psi_k^m)_{k\in K_m}$ with $|K_m| = D_{m_2}$ such that, for all k, the support of ψ_k^m is included in A_2. Here j and k are not necessarily integers, it can be couples of integers as in the case of a piecewise polynomial space. Then, we define

$$S_m = F_{m_1} \otimes H_{m_2} = \{T, \quad T(x,y) = \sum_{j\in J_m} \sum_{k\in K_m} a_{j,k}\varphi_j^m(x)\psi_k^m(y)\}.$$

The assumptions on the models are the following:

M1. For all m_2, $D_{m_2} \leq n^{1/3}$ and $\mathcal{D}_n := \max_{m\in\mathcal{M}_n} D_{m_1} \leq n^{1/3}$.

M2. There exist positive reals ϕ_1, ϕ_2 such that, for all u in F_{m_1}, $\|u\|^2_{\infty,A_1} \leq \phi_1 D_{m_1} \int u^2$, and for all v in H_{m_2}, $\|v\|^2_{\infty,A_2} \leq \phi_2 D_{m_2} \int v^2$. By letting $\phi_0 = \sqrt{\phi_1\phi_2}$, that leads to

$$\forall T \in S_m \qquad \|T\|_\infty \leq \phi_0 \sqrt{D_{m_1} D_{m_2}} \|T\| \tag{2.2}$$

where $\|T\|^2 = \int_{\mathbb{R}^2} T^2(x,y)dxdy$.

M3. $D_{m_1} \leq D_{m_1'} \Rightarrow F_{m_1} \subset F_{m_1'}$ and $D_{m_2} \leq D_{m_2'} \Rightarrow H_{m_2} \subset H_{m_2'}$.

The first assumption guarantees that $\dim S_m = D_{m_1}D_{m_2} \leq n^{2/3} \leq n$ where n is the number of observations. The condition M2 implies a useful link between the L^2 norm and the infinite norm. The third assumption ensures that, for m and m' in $\mathcal{M}_n$, $S_m + S_{m'}$ is included in a model (since $S_m + S_{m'} \subset S_{m''}$ with $D_{m_1''} = \max(D_{m_1}, D_{m_1'})$ and $D_{m_2''} = \max(D_{m_2}, D_{m_2'})$). We denote by $\mathcal{S}$ the space with maximal dimension among the $(S_m)_{m\in\mathcal{M}_n}$. Thus for all m in $\mathcal{M}_n$, $S_m \subset \mathcal{S}$.

2.2.3 Examples of models

We show here that a lot of models are suitable. Indeed, Assumptions M1–M3 are verified for the spaces F_{m_1} (and H_{m_2}) spanned by the following bases (see Barron *et al.* (1999)):

- Trigonometric basis: for $A_1 = [0,1]$, $< \varphi_0, \ldots, \varphi_{m_1-1} >$ with $\varphi_0 = \mathbb{1}_{[0,1]}$, $\varphi_{2j}(x) = \sqrt{2}\cos(2\pi jx)\ \mathbb{1}_{[0,1]}(x)$, $\varphi_{2j-1}(x) = \sqrt{2}\sin(2\pi jx)\mathbb{1}_{[0,1]}(x)$ for $j \geq 1$. For this model $D_{m_1} = m_1$ and $\phi_1 = 2$ hold.

- Histogram basis: for $A_1 = [0,1]$, $< \varphi_1, \ldots, \varphi_{2^{m_1}} >$ with $\varphi_j = 2^{m_1/2}\mathbb{1}_{[(j-1)/2^{m_1}, j/2^{m_1}[}$ for $j = 1, \ldots, 2^{m_1}$. Here $D_{m_1} = 2^{m_1}$, $\phi_1 = 1$.

- Regular piecewise polynomial basis: for $A_1 = [0,1]$, polynomials of degree $0, \ldots, r$ (where r is fixed) on each interval $[(l-1)/2^D, l/2^D[, l = 1, \ldots, 2^D$. In this case, $m_1 = (D, r)$, $J_m = \{j = (l,d), \quad 1 \leq l \leq 2^D, 0 \leq d \leq r\}$, $D_{m_1} = (r+1)2^D$. We can put $\phi_1 = \sqrt{r+1}$.

- Regular wavelet basis: $< \Psi_{lk}, l = -1, \ldots, m_1, k \in \Lambda(l) >$ where $\Psi_{-1,k}$ points out the translates of the father wavelet and $\Psi_{lk}(x) = 2^{l/2}\Psi(2^l x - k)$ where Ψ is the mother wavelet. We assume that the support of the wavelets is included in A_1 and that Ψ_{-1} belongs to the Sobolev space W_2^r.

2.3 Estimation procedure

2.3.1 Definition of the contrast

To estimate the function Π, we define the contrast

$$\gamma_n(T) = \frac{1}{n}\sum_{i=1}^{n}[\int_{\mathbb{R}} T^2(X_i, y)dy - 2T(X_i, X_{i+1})]. \tag{2.3}$$

We choose this contrast because

$$\mathbb{E}\gamma_n(T) = \|T - \Pi\|_f^2 - \|\Pi\|_f^2$$

where

$$\|T\|_f^2 = \int_{\mathbb{R}^2} T^2(x,y)f(x)dxdy.$$

Therefore $\gamma_n(T)$ is the empirical counterpart of the $\|.\|_f$-distance between T and Π and the minimization of this contrast comes down to minimize $\|T - \Pi\|_f$. This contrast is new but is actually connected with the one used in regression problems, as we will see in the next subsection.

We want to estimate Π by minimizing this contrast on S_m. Let $T(x,y) = \sum_{j\in J_m}\sum_{k\in K_m} a_{j,k}\varphi_j^m(x)\psi_k^m(y)$ a function in S_m. Then, if A_m denotes the matrix $(a_{j,k})_{j\in J_m, k\in K_m}$,

$$\forall j_0 \forall k_0 \quad \frac{\partial\gamma_n(T)}{\partial a_{j_0,k_0}} = 0 \Leftrightarrow G_m A_m = Z_m,$$

$$\text{where } \begin{cases} G_m = \left(\frac{1}{n} \sum_{i=1}^{n} \varphi_j^m(X_i) \varphi_l^m(X_i) \right)_{j,l \in J_m} \\ Z_m = \left(\frac{1}{n} \sum_{i=1}^{n} \varphi_j^m(X_i) \psi_k^m(X_{i+1}) \right)_{j \in J_m, k \in K_m} \end{cases}$$

Indeed,

$$\frac{\partial \gamma_n(T)}{\partial a_{j_0,k_0}} = 0 \Leftrightarrow \sum_{j \in J_m} a_{j,k_0} \frac{1}{n} \sum_{i=1}^{n} \varphi_j^m(X_i) \varphi_{j_0}^m(X_i) = \frac{1}{n} \sum_{i=1}^{n} \varphi_{j_0}^m(X_i) \psi_{k_0}^m(X_{i+1}). \tag{2.4}$$

We cannot define a unique minimizer of the contrast $\gamma_n(T)$, since G_m is not necessarily invertible. For example, G_m is not invertible if there exists j_0 in J_m such that there is no observation in the support of φ_{j_0} (G_m has a null column). This phenomenon happens when localized bases (as histogram bases or piecewise polynomial bases) are used. However, the following proposition will enable us to define an estimator:

Proposition 2.1

$$\forall j_0 \forall k_0 \quad \frac{\partial \gamma_n(T)}{\partial a_{j_0,k_0}} = 0 \Leftrightarrow \forall y \quad (T(X_i, y))_{1 \le i \le n} = P_W \left(\left(\sum_k \psi_k^m(X_{i+1}) \psi_k^m(y) \right)_{1 \le i \le n} \right)$$

where P_W denotes the orthogonal projection on $W = \{(T(X_i, y))_{1 \le i \le n},\ T \in S_m\}$.

Thus the minimization of $\gamma_n(t)$ leads to a unique vector $(\hat{\Pi}_m(X_i, y))_{1 \le i \le n}$ defined as the projection of $(\sum_k \psi_k(X_{i+1}) \psi_k(y))_{1 \le i \le n}$ on W. The associated function $\hat{\Pi}_m(.,.)$ is not defined uniquely but we can choose a function $\hat{\Pi}_m$ in S_m whose values at (X_i, y) are fixed according to Proposition 2.1. For the sake of simplicity, we denote

$$\hat{\Pi}_m = \arg \min_{T \in S_m} \gamma_n(T).$$

This underlying function is more a theoretical tool and the estimator is actually the vector $(\hat{\Pi}_m(X_i, y))_{1 \le i \le n}$. This remark leads to consider the risk defined with the empirical norm

$$\|T\|_n = \left(\frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}} T^2(X_i, y) dy \right)^{1/2}. \tag{2.5}$$

This norm is the natural distance in this problem and we can notice that if T is deterministic with support included in $A_1 \times \mathbb{R}$

$$f_0 \|T\|^2 \le \mathbb{E}\|T\|_n^2 = \|T\|_f^2 \le \|f\|_{\infty, A_1} \|T\|^2$$

and then the mean of this empirical norm is equivalent to the L^2 norm $\|.\|$.

2.3.2 Link with classical regression

Let us fix k in K_m and let

$$Y_{i,k} = \psi_k^m(X_{i+1}) \qquad \text{for } i \in \{1, \dots, n\}$$

$$t_k(x) = \int T(x,y)\psi_k^m(y)dy \qquad \text{for all } T \text{ in } L^2(\mathbb{R}^2).$$

Actually, $Y_{i,k}$ and t_k depend on m but we do not mention this for the sake of simplicity. For the same reason, we denote in this subsection ψ_k^m by ψ_k and φ_j^m by φ_j. Then, if T belongs to S_m,

$$\begin{aligned} T(x,y) &= \sum_{j\in J_m}\sum_{k\in K_m}\left(\int T(x',y')\varphi_j(x')\psi_k(y')dx'dy'\right)\varphi_j(x)\psi_k(y) \\ &= \sum_{k\in K_m}\sum_{j\in J_m}\left(\int t_k(x')\varphi_j(x')dx'\right)\varphi_j(x)\psi_k(y) = \sum_{k\in K_m} t_k(x)\psi_k(y) \end{aligned}$$

and then, by replacing this expression of T in $\gamma_n(T)$, we obtain

$$\begin{aligned} \gamma_n(T) &= \frac{1}{n}\sum_{i=1}^n [\int \sum_{k,k'} t_k(X_i)t_{k'}(X_i)\psi_k(y)\psi_{k'}(y)dy - 2\sum_k t_k(X_i)\psi_k(X_{i+1})] \\ &= \frac{1}{n}\sum_{i=1}^n\sum_{k\in K_m}[t_k^2(X_i) - 2t_k(X_i)Y_{i,k}] = \frac{1}{n}\sum_{i=1}^n\sum_{k\in K_m}[t_k(X_i) - Y_{i,k}]^2 - Y_{i,k}^2. \end{aligned}$$

Consequently

$$\min_{T\in S_m}\gamma_n(T) = \sum_{k\in K_m}\min_{t_k\in F_{m_1}}\frac{1}{n}\sum_{i=1}^n [t_k(X_i) - Y_{i,k}]^2 - Y_{i,k}^2.$$

We recognize, for all k, the least squares contrast, which is used in regression problems. Here the regression function is $\pi_k = \int \Pi(.,y)\psi_k(y)dy$ which verifies

$$Y_{i,k} = \pi_k(X_i) + \varepsilon_{i,k} \tag{2.6}$$

where

$$\varepsilon_{i,k} = \psi_k(X_{i+1}) - \mathbb{E}[\psi_k(X_{i+1})|X_i]. \tag{2.7}$$

The estimator $\hat{\Pi}_m$ can be written as $\sum_{k\in K_m}\hat{\pi}_k(x)\psi_k(y)$ where $\hat{\pi}_k$ is the classical least squares estimator for the regression model (2.6) (as previously, only the vector $(\hat{\pi}_k(X_i))_{1\le i\le n}$ is uniquely defined).

This regression model is used in Clémençon (2000) to estimate the transition density. In the same manner, we could here use the contrast $\gamma_n^{(k)}(t) = \frac{1}{n}\sum_{i=1}^n[\psi_k(X_{i+1}) - t(X_i)]^2$ to take advantage of analogy with regression. This method allows to have a good estimation of the projection of Π on some S_m by estimating first each π_k, but does not provide an adaptive method. Model selection requires a more global contrast, as described in (2.3).

2.3.3 Definition of the estimator

We have then an estimator of Π for all S_m. Let now

$$\hat{m} = \arg \min_{m \in \mathcal{M}_n} \{\gamma_n(\hat{\Pi}_m) + \text{pen}(m)\}$$

where pen is a penalty function to be specified later. Then we can define $\tilde{\Pi} = \hat{\Pi}_{\hat{m}}$ and compute the empirical mean integrated squared error $\mathbb{E}\|\Pi - \tilde{\Pi}\|_n^2$ where $\|.\|_n$ is the empirical norm defined in (2.5).

2.4 Calculation of the risk

For a function h and a subspace S, we recall that

$$d(h, S) = \inf_{g \in S} \|h - g\| = \inf_{g \in S} \left(\iint |h(x,y) - g(x,y)|^2 dxdy \right)^{1/2}.$$

With an inequality of Talagrand (1996), we can prove the following result.

Theorem 2.1 *We consider a Markov chain satisfying Assumptions H1-H2b-H3a,b-H4 (with $\theta > 14$ in the case of an arithmetical mixing). We consider $\tilde{\Pi}$ the estimator of the transition density Π described in Section 2.3 with models verifying Assumptions M1–M3 and the following penalty:*

$$\text{pen}(m) = K_0 \|\Pi\|_{\infty,A} \frac{D_{m_1} D_{m_2}}{n} \tag{2.8}$$

where K_0 is a numerical constant. Then

$$\mathbb{E}\|\Pi \mathbb{1}_A - \tilde{\Pi}\|_n^2 \leq C \inf_{m \in \mathcal{M}_n} \{d^2(\Pi \mathbb{1}_A, S_m) + \text{pen}(m)\} + \frac{C'}{n}$$

where $C = \max(5\|f\|_{\infty,A_1}, 6)$ and C' is a constant depending on ϕ_1, ϕ_2, $\|\Pi\|_{\infty,A}$, f_0, $\|f\|_{\infty,A_1}$, θ.

The constant K_0 in the penalty is purely numerical (we can choose $K_0 = 45$). We observe that the term $\|\Pi\|_{\infty,A}$ appears in the penalty although it is unknown. Actually this term $\|\Pi\|_{\infty,A}$ in the penalty can be replaced by f_0^{-1} if we use an isotropic estimator (see Remark 2.2 in the proof 2.7.3). In the anisotropic case the infinite norm of Π can be replaced by any bound of $\|\Pi\|_{\infty,A}$. Moreover, it is possible to use $\|\hat{\Pi}\|_\infty$ where $\hat{\Pi}$ is some estimator of Π. This method of random penalty (specifically with infinite norm) is successfully used in Birgé and Massart (1997) and Comte (2001) for example, and can be applied here even if it means considering Π regular enough. More precisely, we can write the following theorem.

Theorem 2.2 *We consider the following penalty :*

$$\overline{\text{pen}}(m) = \overline{K_0}\|\hat{\Pi}\|_\infty \frac{D_{m_1} D_{m_2}}{n}$$

where $\overline{K_0}$ *is a numerical constant and* $\hat{\Pi} = \hat{\Pi}_{m*}$ *with* S_{m*} *a space of trigonometric polynomials such that*

$$\log n \leq D_{m_1*} = D_{m_2*} \leq n^{1/6}.$$

If the restriction of Π *to* A *belongs to* $B_{2,\infty}^{(\alpha_1,\alpha_2)}(A)$ *with* $\alpha_1 > 3/2$ *and* $\alpha_2 > \max(\frac{\alpha_1}{2\alpha_1-3}, \frac{3\alpha_1}{2\alpha_1-1})$*, then, under assumptions of Theorem 2.1, for* n *large enough,*

$$\mathbb{E}\|\Pi\mathbb{1}_A - \tilde{\Pi}\|_n^2 \leq C \inf_{m\in\mathcal{M}_n} \left\{ d^2(\Pi\mathbb{1}_A, S_m) + \frac{D_{m_1} D_{m_2}}{n} \right\} + \frac{C'}{n}.$$

Remark 2.1 *The condition on the regularity of* Π *is verified for example if* $\alpha_1 > 2$ *and* $\alpha_2 > 2$*. If* $\alpha_1 = \alpha_2 = \alpha$*, it is equivalent to* $\alpha > 2$*.*

It is relevant to notice that the penalty term does not contain any mixing term and is then entirely computable. It is in fact related to martingale properties of the underlying empirical processes. The constant K_0 is a fixed universal numerical constant; for practical purposes, it is adjusted by simulations.

We are now interested in the rate of convergence of the risk. We consider that Π restricted to A belongs to the anisotropic Besov space on A with regularity $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$. Note that if Π belongs to $B_{2,\infty}^{\boldsymbol{\alpha}}(\mathbb{R}^2)$, then Π restricted to A belongs to $B_{2,\infty}^{\boldsymbol{\alpha}}(A)$. Let us recall the definition of $B_{2,\infty}^{\boldsymbol{\alpha}}(A)$. Let e_1 and e_2 be the canonical basis vectors in $\mathbb{R}^2$ and for $i = 1, 2$, $A_{h,i}^r = \{x \in \mathbb{R}^2; x, x + he_i, \ldots, x + rhe_i \in A\}$. Next, for x in $A_{h,i}^r$, let

$$\Delta_{h,i}^r g(x) = \sum_{k=0}^{r} (-1)^{r-k} \binom{r}{k} g(x + khe_i)$$

the rth difference operator with step h. For $t > 0$, the directional moduli of smoothness are given by

$$\omega_{r_i,i}(g,t) = \sup_{|h|\leq t} \left(\int_{A_{h,i}^{r_i}} |\Delta_{h,i}^{r_i} g(x)|^2 dx \right)^{1/2}.$$

We say that g is in the Besov space $B_{2,\infty}^{\boldsymbol{\alpha}}(A)$ if

$$\sup_{t>0} \sum_{i=1}^{2} t^{-\alpha_i} \omega_{r_i,i}(g,t) < \infty$$

for r_i integers larger than α_i. The transition density Π can thus have different smoothness properties with respect to different directions. The procedure described here allows an adaptation of the approximation space to each directional regularity. More precisely, if $\alpha_2 > \alpha_1$ for example, the estimator chooses a space of dimension $D_{m_2} = D_{m_1}^{\alpha_1/\alpha_2} < D_{m_1}$ for the second direction, where Π is more regular. We can thus write the following corollary.

Corollary 2.1 *We suppose that* Π *restricted to* A *belongs to the anisotropic Besov space* $B^{\boldsymbol{\alpha}}_{2,\infty}(A)$ *with regularity* $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ *such that* $\alpha_1 - 2\alpha_2 + 2\alpha_1\alpha_2 > 0$ *and* $\alpha_2 - 2\alpha_1 + 2\alpha_1\alpha_2 > 0$. *We consider the spaces described in Subsection 2.2.3 (with the regularity* r *of the polynomials and the wavelets larger than* $\alpha_i - 1$*). Then, under the assumptions of Theorem 2.1,*

$$\mathbb{E}\|\Pi\mathbb{1}_A - \tilde{\Pi}\|_n^2 = O(n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}}).$$

where $\bar{\alpha}$ *is the harmonic mean of* α_1 *and* α_2.

The harmonic mean of α_1 and α_2 is the real $\bar{\alpha}$ such that $2/\bar{\alpha} = 1/\alpha_1 + 1/\alpha_2$. Note that the condition $\alpha_1 - 2\alpha_2 + 2\alpha_1\alpha_2 > 0$ is ensured as soon as $\alpha_1 \geq 1$ and the condition $\alpha_2 - 2\alpha_1 + 2\alpha_1\alpha_2 > 0$ as soon as $\alpha_2 \geq 1$.

Thus we obtain the rate of convergence $n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}}$, which is optimal in the minimax sense (see Section 5.3 for the lower bound).

2.5 L^2 control

2.5.1 Estimation procedure

Although the empirical norm is the more natural in this problem, we are interested in a L^2 control of the risk. For this, the estimation procedure must be modified. We truncate the previous estimator in the following way :

$$\tilde{\Pi}^* = \begin{cases} \tilde{\Pi} & \text{if } \|\tilde{\Pi}\| \leq k_n \\ 0 & \text{else} \end{cases} \tag{2.9}$$

with $k_n = n^{2/3}$.

2.5.2 Calculation of the L^2 risk

We obtain in this framework a result similar to Theorem 2.1.

Theorem 2.3 *We consider a Markov chain satisfying Assumptions H1-H2b-H3a,b-H4 (with* $\theta > 20$ *in the case of an arithmetical mixing). We consider* $\tilde{\Pi}^*$ *the estimator of the transition density* Π *described in Section 2.5.1. Then*

$$\mathbb{E}\|\tilde{\Pi}^* - \Pi\|_A^2 \leq C \inf_{m \in \mathcal{M}_n} \{d^2(\Pi\mathbb{1}_A, S_m) + \text{pen}(m)\} + \frac{C'}{n}.$$

where $C = \max(36 f_0^{-1}\|f\|_{\infty,A_1} + 2, 36 f_0^{-1})$ *and* C' *is a constant depending on* ϕ_1, ϕ_2, $\|\Pi\|_{\infty,A}$, f_0, $\|f\|_{\infty,A_1}$, θ.

If Π is regular, we can state the following corollary:

Corollary 2.2 *We suppose that the restriction of* Π *to* A *belongs to the anisotropic Besov space* $B^{\boldsymbol{\alpha}}_{2,\infty}(A)$ *with regularity* $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ *such that* $\alpha_1 - 2\alpha_2 + 2\alpha_1\alpha_2 > 0$ *and* $\alpha_2 - 2\alpha_1 + 2\alpha_1\alpha_2 > 0$. *We consider the spaces described in Subsection 2.2.3 (with the regularity* r *of the polynomials and the wavelets larger than* $\alpha_i - 1$*). Then, under the assumptions of Theorem 2.3,*

$$\mathbb{E}\|\Pi - \tilde{\Pi}^*\|_A^2 = O(n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}}).$$

where $\bar{\alpha}$ *is the harmonic mean of* α_1 *and* α_2.

The same rate of convergence is then achieved with the L^2 norm instead of the empirical norm. And the procedure allows to adapt automatically the two dimensions of the projection spaces to the regularities α_1 and α_2 of the transition density Π. If $\alpha_1 = 1$ we recognize the rate $n^{-\frac{\alpha_2}{3\alpha_2+1}}$ established by Birgé (1983) with metrical arguments. The optimality is proved in the following subsection.

If $\alpha_1 = \alpha_2 = \alpha$ ("classical" Besov space), then $\bar{\alpha} = \alpha$ and our result is thus an improvement of the one of Clémençon (2000), whose procedure achieves only the rate $(\log(n)/n)^{\frac{2\alpha}{2\alpha+2}}$ and allows to use only wavelets. We can observe that in this case, the condition $\alpha_1 - 2\alpha_2 + 2\alpha_1\alpha_2 > 0$ is equivalent to $\alpha > 1/2$ and so is verified if the function Π is regular enough.
Actually, in the case $\alpha_1 = \alpha_2$, an estimation with isotropic spaces ($D_{m_1} = D_{m_2}$) is preferable. Indeed, in this framework, the models are nested and so we can consider spaces with larger dimension ($D_m^2 \leq n$ instead of $D_m^2 \leq n^{2/3}$). Then Corollary 2.1 is valid whatever $\alpha > 0$. Moreover, for the arithmetic mixing, assumption $\theta > 7$ is sufficient.

2.5.3 Lower bound

We set

$$\mathcal{B} = \{\Pi \text{ transition density on } \mathbb{R} \text{ of a positive recurrent} \\ \text{Markov chain such that } \|\Pi\|_{B^{\boldsymbol{\alpha}}_{2,\infty}(A)} \leq L\}$$

and $\mathbb{E}_\Pi$ the expectation corresponding to the distribution of $X_1, \ldots, X_n$ if the true transition density of the Markov chain is Π and the initial distribution is the stationary distribution.

Theorem 2.4 *There exists a positive constant* C *such that, if* n *is large enough,*

$$\inf_{\hat{\Pi}_n} \sup_{\Pi \in \mathcal{B}} \mathbb{E}_\Pi \|\hat{\Pi}_n - \Pi\|_A^2 \geq C n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}}$$

where the infimum is taken over all estimators $\hat{\Pi}_n$ *of* Π *based on the observations* $X_1, \ldots, X_{n+1}$.

So the lower bound in Clémençon (2000) is generalized for the case $\alpha_1 \neq \alpha_2$. It shows that our procedure reaches the optimal minimax rate, whatever the regularity of Π, without needing to know $\boldsymbol{\alpha}$.

A more general theorem of lower bound for the L^p norm is proved in Annex 2.8.

2.5.4 Applications and prospects

• An easy application of our result is the computation of $\lambda(x) = \mathbb{E}[g(X_{i+1})|X_i = x]$ for x in A_1, where g is a borelian function belonging to $L^2(A_2)$ with support included in A_2. Since $\lambda(x) = \int_{\mathbb{R}} g(y)\Pi(x,y)dy = \int_{A_2} g(y)\Pi(x,y)dy$, a natural estimator is

$$\tilde{\lambda}^*(x) = \int g(y)\tilde{\Pi}^*(x,y)dy.$$

The Schwarz inequality leads to :

$$\mathbb{E}\|\lambda - \tilde{\lambda}^*\|_{A_1}^2 \leq \|g\|^2\mathbb{E}\|\Pi - \tilde{\Pi}^*\|_A^2 = O(n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}}).$$

If Π belongs to the anisotropic Besov space $B_{2,\infty}^{\boldsymbol{\alpha}}(A)$, then λ belongs to the Besov $B_{2,\infty}^{\alpha_1}(A_1)$. So the minimax rate for this regression function is $n^{-\frac{2\alpha_1}{2\alpha_1+1}}$. Here the convergence is a little slower but tends to the minimax rate if we make α_2 tend to infinity.

In the same way, we can estimate the random variables $\mathbb{E}[g(X_{i+1})|X_i] = \lambda(X_i)$ by $\tilde{\lambda}(X_i)$ where $\tilde{\lambda}(x) = \int_{A_2} g(y)\tilde{\Pi}(x,y)dy$. We have then

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}|\lambda - \tilde{\lambda}|^2(X_i)\right) \leq \|g\|^2\mathbb{E}\|\Pi\mathbb{1}_A - \tilde{\Pi}\|_n^2 = O(n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}}).$$

• The estimation of Π allows us to estimate the density of $P^q(x,.)$, the law of X_{i+q} conditionnally to $X_i = x$. We denote by $\Pi^{(q)}$ this density:

$$\Pi^{(q)}(x,y) = \int_{\mathbb{R}^{q-1}} \Pi(x,z_1)\Pi(z_1,z_2)...\Pi(z_{q-1},y)dz_1dz_2...dz_{q-1}.$$

We can define an estimator of $\Pi^{(q)}$ recursively by setting $\tilde{\Pi}^{(1)} = \tilde{\Pi}$ and

$$\tilde{\Pi}^{(q)}(x,y) = \int \tilde{\Pi}^{(q-1)}(x,z)\tilde{\Pi}(z,y)dz \qquad q \geq 2.$$

• With an estimator of the iterates of Π, it is possible to introduce an estimator of the mixing coefficients, since Formula (2.1) gives

$$\beta_q = \frac{1}{2}\iint |\Pi^{(q)}(x,y) - f(y)|f(x)dxdy.$$

2.6 Simulations

To evaluate the performance of our method, we simulate a Markov chain with a known transition density and then we estimate this density and compare the two functions for different values of n. The estimation procedure is easy, we can decompose it in some steps:

- find the coefficients matrix A_m for each $m = (m_1, m_2)$,
- compute $\gamma_n(\hat{\Pi}_m) = \mathrm{Tr}({}^tA_m G_m A_m - 2\,{}^tA_m Z_m) = -\mathrm{Tr}({}^tA_m Z_m)$,
- find $\hat{m}$ such that $\gamma_n(\hat{\Pi}_m) + \mathrm{pen}(m)$ is minimum,
- compute $\hat{\Pi}_{\hat{m}}$.

For the first step, we use two different kinds of bases : the histogram bases and the trigonometric bases, as described in subsection 2.2.3. We renormalize these bases so that they are defined on the estimation domain A instead of $[0,1]^2$. For the third step, we choose $\mathrm{pen}(m) = 0.5\dfrac{D_{m_1} D_{m_2}}{n}$.

We consider the same Markov chains as in Chapter 1 (see page 38 for a precise description):

- An autoregressive process denoted by AR and defined by:

  $$X_{n+1} = aX_n + b + \varepsilon_{n+1}$$

  where the ε_{n+1} are independent and identical distributed random variables, with centered Gaussian distribution with variance σ^2. We consider the following parameter values :

  (i) $a = 2/3$, $b = 0$, $\sigma^2 = 5/9$, estimated on $[-2,2]^2$.

  (ii) $a = 0.5$, $b = 3$, $\sigma^2 = 1$, and then the process is estimated on $[4,8]^2$.

- A radial Ornstein-Uhlenbeck process (in its discrete version). For $j = 1, \ldots, \delta$, we define the processes: $\xi^j_{n+1} = a\xi^j_n + \beta\varepsilon^j_n$ where the ε^j_n are i.i.d. standard Gaussian. The chain is then defined by $X_n = \sqrt{\sum_{i=1}^{\delta}(\xi^i_n)^2}$. This process (with here $a = 0.5$, $\beta = 3$, $\delta = 3$) is denoted by $\sqrt{\mathrm{CIR}}$. The estimation domain for this process is $[2,10]^2$.

- A Cox-Ingersoll-Ross process. The used parameters are the following:

  (iii) $a = 3/4$, $\beta = \sqrt{7/48}$ (so that $l = 3/2$) and $\delta = 4$, estimated on $[0.1, 3]^2$.

  (iv) $a = 1/3$, $\beta = 3/4$ and $\delta = 2$. This chain is estimated on $[0,2]^2$.

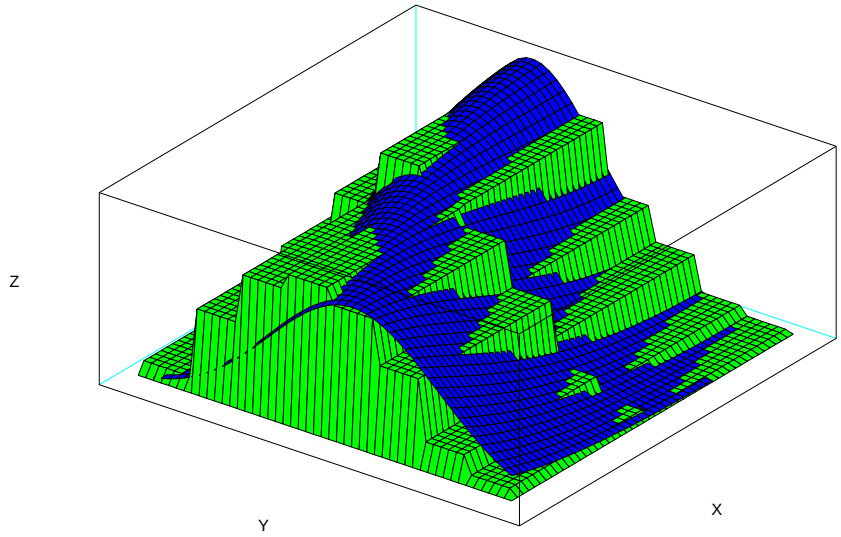
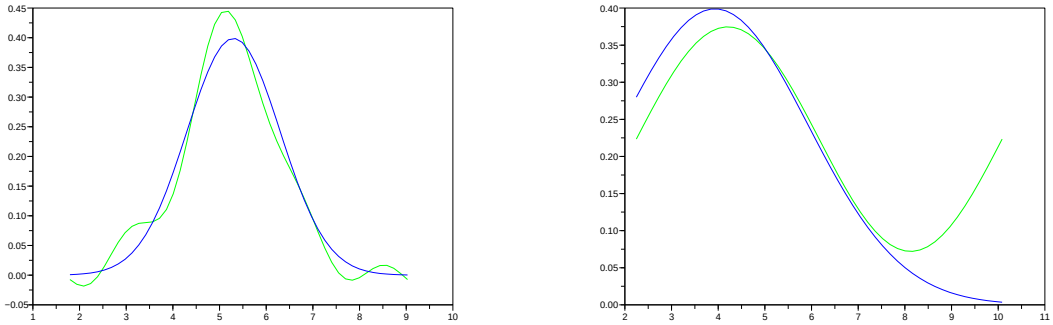


Figure 2.1: Estimator (light surface) and true fonction (dark surface) for a $\sqrt{\text{CIR}}$ process estimated with a histogram basis, $n = 1000$.



$x = 4.6$ $y = 5$

Figure 2.2: Sections for AR(ii) process estimated with a trigonometric basis, $n = 1000$, dark line: true function, light line: estimator.

- An ARCH process defined by $X_{n+1} = \sin(X_n) + (\cos(X_n) + 3)\varepsilon_{n+1}$ where the ε_{n+1} are i.i.d. standard Gaussian. We estimate this process on $[-5, 5]^2$.

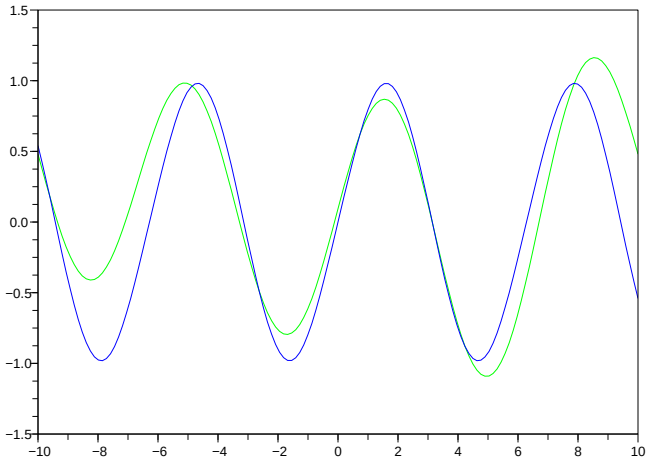


Figure 2.3: True function $\mathbb{E}[X_{i+1}|X_i = x]$ (dark line) and its estimator (light line) for ARCH process estimated with a trigonometric basis, $n = 1000$

We can illustrate the results by some figures. Figure 1 shows the surface $z = \Pi(x, y)$ and the estimated surface $z = \tilde{\Pi}(x, y)$. We use a histogram basis and we see that the procedure chooses different dimensions on the abscissa and on the ordinate since the estimator is constant on rectangles instead of squares. Figure 2.2 presents sections of this kind of surfaces for the AR(ii) process estimated with trigonometric bases. We can see the curves $z = \Pi(4.6, y)$ versus $z = \tilde{\Pi}(4.6, y)$ and the curves $z = \Pi(x, 5)$ versus $z = \tilde{\Pi}(x, 5)$. The second section shows that it may exist some edge effects due to the mixed control of the two directions. These edge effects are also observable on the last figure where we can see an estimator of $\mathbb{E}[X_{i+1}|X_i = x]$ (application presented in subsection 2.5.4 with $g(y) = y$) for the ARCH process. Here we have estimated Π on a larger domain (theoretically it should be $\mathbb{R}^2$ since the function g is defined on $\mathbb{R}$).

For more precise results, empirical risk and L^2 risk are respectively given in Table 2.1 and Table 2.2.

We observe that the results are better when we consider the empirical norm. It was expectable, given that this norm is adapted to the studied problem. Actually the better norm to evaluate the distance between Π and its estimator is the norm $\|.\|_f$. Table 2.3 shows that the errors in this case are very satisfactory.

We can compare these results with the one obtained by the quotient method in Chapter 1. Table 2.4 gives the ratio between the risk presented previously and the risk obtained in this chapter with the least-square type estimator. The errors are always lower with the regression method for an estimation by histograms. For the trigonometric basis, it depends more on the simulated processes. We can imagine that in some cases the two

law \ n	50	100	250	500	1000	basis
AR(i)	0.088	0.079	0.059	0.044	0.036	H
	0.096	0.083	0.063	0.055	0.047	T
AR(ii)	0.067	0.055	0.043	0.038	0.033	H
	0.096	0.081	0.063	0.054	0.045	T
$\sqrt{\text{CIR}}$	0.026	0.023	0.019	0.016	0.014	H
	0.019	0.015	0.009	0.007	0.006	T
CIR(iii)	0.097	0.091	0.067	0.057	0.047	H
	0.163	0.132	0.101	0.087	0.074	T
CIR(iv)	0.118	0.115	0.095	0.089	0.078	H
	0.344	0.272	0.185	0.149	0.118	T
ARCH	0.029	0.026	0.021	0.017	0.017	H
	0.019	0.017	0.010	0.009	0.008	T

Table 2.1: Empirical risk $\mathbb{E}\|\Pi - \tilde{\Pi}\|_n^2$ for simulated data with $\text{pen}(m) = 0.5 D_{m_1} D_{m_2}/n$, averaged over $N = 200$ samples. H: histogram basis, T: trigonometric basis.

law \ n	50	100	250	500	1000	basis
AR(i)	0.440	0.379	0.270	0.181	0.136	H
	0.537	0.440	0.315	0.276	0.239	T
AR(ii)	0.242	0.189	0.132	0.109	0.085	H
	0.438	0.357	0.253	0.213	0.180	T
$\sqrt{\text{CIR}}$	0.152	0.130	0.094	0.066	0.054	H
	0.152	0.123	0.072	0.052	0.046	T
CIR(iii)	0.297	0.273	0.185	0.148	0.108	H
	0.599	0.477	0.338	0.282	0.239	T
CIR(iv)	0.172	0.155	0.104	0.084	0.053	H
	0.844	0.635	0.410	0.323	0.250	T
ARCH	0.267	0.243	0.167	0.125	0.117	H
	0.223	0.187	0.115	0.094	0.092	T

Table 2.2: L^2 risk $\mathbb{E}\|\Pi - \tilde{\Pi}^*\|^2$ for simulated data with $\text{pen}(m) = 0.5 D_{m_1} D_{m_2}/n$, averaged over $N = 200$ samples. H: histogram basis, T: trigonometric basis.

errors that we made by quotient estimation balance each other instead of add up. In the other cases the regression method is 1.5 to 2 times better. The improvement is in particular due to the anisotropic nature of the estimation. It is especially the case for

law \ n	50	100	250	500	1000	basis
AR(i)	0.087	0.072	0.049	0.033	0.024	H
	0.091	0.073	0.048	0.041	0.033	T
AR(ii)	0.052	0.038	0.026	0.020	0.015	H
	0.081	0.069	0.046	0.037	0.031	T
$\sqrt{\text{CIR}}$	0.016	0.014	0.010	0.006	0.004	H
	0.018	0.012	0.008	0.006	0.004	T
CIR(iii)	0.086	0.077	0.051	0.040	0.029	H
	0.147	0.113	0.077	0.062	0.050	T
CIR(iv)	0.058	0.023	0.034	0.027	0.016	H
	0.297	0.224	0.146	0.115	0.089	T

Table 2.3: $L^2(f(x)dxdy)$ risk $\mathbb{E}\|\Pi - \tilde{\Pi}^*\|_f^2$ for simulated data with $\text{pen}(m) = 0.5D_{m_1}D_{m_2}/n$, averaged over $N = 200$ samples. H: histogram basis, T: trigonometric basis.

the histogram basis which divides the space in rectangles more or less large and which allows to adapt locally the estimator. Thus the least-square type estimator has not only theoretical advantages but is also preferable for practical purposes.

law \ n	50	100	250	500	1000	basis
AR(i)	1.65	1.44	1.03	1.03	1.30	H
	0.98	1.06	0.70	0.65	0.62	T
AR(ii)	1.98	1.72	1.70	1.06	0.99	H
	0.66	0.67	0.70	0.63	0.60	T
$\sqrt{CIR}$	2.01	1.78	1.83	2.30	2.37	H
	1.42	1.58	2.01	2.46	1.78	T
CIR(iii)	1.71	1.13	1.14	1.19	1.37	H
	0.70	0.83	0.84	0.91	0.95	T
CIR(iv)	1.97	1.35	1.16	0.90	0.87	H
	0.27	0.35	0.42	0.41	0.53	T
ARCH	1.19	1.24	1.45	1.70	1.38	H
	1.14	1.36	1.81	2.00	1.84	T

Table 2.4: ratio L^2 risk with quotient method/ L^2 risk with regression method.

2.7 Proofs

2.7.1 Proof of Proposition 2.1

Equality (2.4) yields, by multiplying by $\psi_{k_0}^m(y)$,

$$\sum_{j\in J_m} a_{j,k_0} \sum_{i=1}^n \varphi_j^m(X_i)\psi_{k_0}^m(y)\varphi_{j_0}^m(X_i) = \sum_{i=1}^n \varphi_{j_0}^m(X_i)\psi_{k_0}^m(X_{i+1})\psi_{k_0}^m(y).$$

Then, we sum over k_0 in K_m:

$$\sum_{i=1}^n T(X_i,y)\varphi_{j_0}^m(X_i) = \sum_{i=1}^n \sum_{k_0\in K_m} \psi_{k_0}^m(X_{i+1})\psi_{k_0}^m(y)\varphi_{j_0}^m(X_i).$$

If we multiply this equality by $a'_{j_0,k}\psi_k^m(y)$ and if we sum over $k \in K_m$ and $j_0 \in J_m$, we obtain

$$\sum_{i=1}^n [T(X_i,y) - \sum_{k_0\in K_m} \psi_{k_0}^m(X_{i+1})\psi_{k_0}^m(y)] \sum_{k\in K_m}\sum_{j_0\in J_m} a'_{j_0,k}\varphi_{j_0}^m(X_i)\psi_k^m(y) = 0$$

$$\text{i.e.} \qquad \sum_{i=1}^n [T(X_i,y) - \sum_{k_0\in K_m} \psi_{k_0}^m(X_{i+1})\psi_{k_0}^m(y)]u(X_i,y) = 0$$

for all u in S_m. So the vector $(T(X_i,y) - \sum_{k\in K_m}\psi_k^m(X_{i+1})\psi_k^m(y))_{1\le i\le n}$ is orthogonal to each vector in W. Since $T(X_i,y)$ belongs to W, the proposition is proved. □

2.7.2 Proof of Theorem 2.1

For ρ a real larger than 1, let

$$\Omega_\rho = \{\forall T \in \mathcal{S} \quad \|T\|_f^2 \le \rho\|T\|_n^2\}$$

In the case of an arithmetical mixing, since $\theta > 14$, there exists a real c such that

$$\begin{cases} 0 < c < \dfrac{1}{6} \\ \theta c > \dfrac{7}{3} \end{cases}$$

We set in this case $q_n = \frac{1}{2}\lfloor n^c \rfloor$. In the case of a geometrical mixing, we set $q_n = \frac{1}{2}\lfloor c\log(n)\rfloor$ where c is a real larger than $7/3\theta$.

For the sake of simplicity, we suppose that $n = 4p_nq_n$, with p_n an integer. Let for $i = 1,\dots,n/2$, $U_i = (X_{2i-1}, X_{2i})$.

Let $\begin{cases} A_l = & (U_{2lq_n+1}, ..., U_{(2l+1)q_n}) \qquad l = 0, \dots, p_n - 1, \\ B_l = & (U_{(2l+1)q_n+1}, ..., U_{(2l+2)q_n}) \qquad l = 0, \dots, p_n - 1. \end{cases}$

We use now the mixing assumption H2b. As in Viennet (1997) we can build a sequence (A_l^*) such that

$$\begin{cases} A_l \text{ and } A_l^* \text{ have the same distribution,} \\ A_l^* \text{ and } A_{l'}^* \text{ are independent if } l \neq l', \\ P(A_l \neq A_l^*) \leq \beta_{2q_n}. \end{cases}$$

In the same way, we build (B_l^*) and we define for any $l \in \{0, \dots, p_n - 1\}$, $A_l^* = (U_{2lq_n+1}^*, ..., U_{(2l+1)q_n}^*)$, $B_l^* = (U_{(2l+1)q_n+1}^*, ..., U_{(2l+2)q_n}^*)$ so that the sequence $(U_1^*, \dots, U_{n/2}^*)$ and then the sequence $(X_1^*, \dots, X_n^*)$ are well defined.
Let now $V_i = (X_{2i}, X_{2i+1})$ for $i = 1, \dots, n/2$ and

$\begin{cases} C_l = & (V_{2lq_n+1}, ..., V_{(2l+1)q_n}) \qquad l = 0, \dots, p_n - 1, \\ D_l = & (V_{(2l+1)q_n+1}, ..., V_{(2l+2)q_n}) \qquad l = 0, \dots, p_n - 1. \end{cases}$

We can build $(V_1^{**}, \dots, V_{n/2}^{**})$ and then $(X_2^{**}, \dots, X_{n+1}^{**})$ such that

$$\begin{cases} C_l \text{ and } C_l^{**} \text{ have the same distribution,} \\ C_l^{**} \text{ and } C_{l'}^{**} \text{ are independent if } l \neq l', \\ P(C_l \neq C_l^{**}) \leq \beta_{2q_n}. \end{cases}$$

We put $X_{n+1}^* = X_{n+1}$ and $X_1^{**} = X_1$. Now let

$$\Omega^* = \{\forall i \quad X_i = X_i^* = X_i^{**}\} \quad \text{and} \quad \Omega_\rho^* = \Omega_\rho \cap \Omega^*.$$

We denote by Π_m the orthogonal projection of Π on S_m. Now,

$$\mathbb{E}\|\tilde{\Pi} - \Pi \mathbb{1}_A\|_n^2 = \mathbb{E}\left(\|\tilde{\Pi} - \Pi \mathbb{1}_A\|_n^2 \mathbb{1}_{\Omega_\rho^*}\right) + \mathbb{E}\left(\|\tilde{\Pi} - \Pi \mathbb{1}_A\|_n^2 \mathbb{1}_{\Omega_\rho^{*c}}\right). \tag{2.10}$$

To bound the first term, we observe that for all S, T

$$\gamma_n(S) - \gamma_n(T) = \|T - \Pi\|_n^2 - \|S - \Pi\|_n^2 - 2Z_n(T - S)$$

where $Z_n(T) = \dfrac{1}{n}\sum_{i=1}^{n}\left\{T(X_i, X_{i+1}) - \displaystyle\int_{\mathbb{R}} T(X_i, y)\Pi(X_i, y)dy\right\}$.
Since $\|T - \Pi\|_n^2 = \|T - \Pi \mathbb{1}_A\|_n^2 + \|\Pi \mathbb{1}_{A^c}\|_n^2$, we can write

$$\gamma_n(T) - \gamma_n(S) = \|T - \Pi \mathbb{1}_A\|_n^2 - \|S - \Pi \mathbb{1}_A\|_n^2 - 2Z_n(T - S).$$

The definition of $\hat{m}$ gives, for some fixed $m \in \mathcal{M}_n$,

$$\gamma_n(\tilde{\Pi}) + \text{pen}(\hat{m}) \leq \gamma_n(\Pi_m) + \text{pen}(m)$$

And then

$$\begin{aligned}\|\tilde{\Pi}-\Pi\mathbb{1}_A\|_n^2 &\leq \|\Pi_m-\Pi\mathbb{1}_A\|_n^2+2Z_n(\tilde{\Pi}-\Pi_m)+\mathrm{pen}(m)-\mathrm{pen}(\hat{m})\\ &\leq \|\Pi_m-\Pi\mathbb{1}_A\|_n^2+2\|\tilde{\Pi}-\Pi_m\|_f\sup_{T\in B_f(m,\hat{m})}Z_n(T)+\mathrm{pen}(m)-\mathrm{pen}(\hat{m})\end{aligned}$$

where, for all m', $B_f(m,m')=\{T\in S_m+S_{m'},\quad \|T\|_f=1\}$. Let κ a real larger than 2ρ and $p(.,.)$ a function such that $\kappa p(m,m')\leq \mathrm{pen}(m)+\mathrm{pen}(m')$. Then

$$\begin{aligned}\|\tilde{\Pi}-\Pi\mathbb{1}_A\|_n^2\mathbb{1}_{\Omega_\rho^*} &\leq \|\Pi_m-\Pi\mathbb{1}_A\|_n^2+\frac{1}{\kappa}\|\tilde{\Pi}-\Pi_m\|_f^2\mathbb{1}_{\Omega_\rho^*}+2\mathrm{pen}(m)\\ &\quad+\kappa\sum_{m'\in\mathcal{M}_n}\left[\sup_{T\in B_f(m,m')}Z_n^2(T)-p(m,m')\right]_+\mathbb{1}_{\Omega^*} \qquad (2.11)\end{aligned}$$

But $\|\tilde{\Pi}-\Pi_m\|_f^2\mathbb{1}_{\Omega_\rho^*}\leq\rho\|\tilde{\Pi}-\Pi_m\|_n^2\mathbb{1}_{\Omega_\rho^*}\leq 2\rho\|\tilde{\Pi}-\Pi\mathbb{1}_A\|_n^2\mathbb{1}_{\Omega_\rho^*}+2\rho\|\Pi\mathbb{1}_A-\Pi_m\|_n^2$.

Then, inequality (2.11) becomes

$$\begin{aligned}\|\tilde{\Pi}-\Pi\mathbb{1}_A\|_n^2\mathbb{1}_{\Omega_\rho^*}\left(1-\frac{2\rho}{\kappa}\right) &\leq \left(1+\frac{2\rho}{\kappa}\right)\|\Pi_m-\Pi\mathbb{1}_A\|_n^2+2\mathrm{pen}(m)\\ &\quad+\kappa\sum_{m'\in\mathcal{M}_n}\left[\sup_{T\in B_f(m,m')}Z_n^2(T)-p(m,m')\right]_+\mathbb{1}_{\Omega^*}\\ \text{so}\quad \mathbb{E}\left(\|\tilde{\Pi}-\Pi\mathbb{1}_A\|_n^2\mathbb{1}_{\Omega_\rho^*}\right) &\leq \frac{\kappa+2\rho}{\kappa-2\rho}\mathbb{E}\|\Pi\mathbb{1}_A-\Pi_m\|_n^2+\frac{2\kappa}{\kappa-2\rho}\mathrm{pen}(m)\\ &\quad+\frac{\kappa^2}{\kappa-2\rho}\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left(\left[\sup_{T\in B_f(m,m')}Z_n^2(T)-p(m,m')\right]_+\mathbb{1}_{\Omega^*}\right) \qquad (2.12)\end{aligned}$$

We now use the following proposition:

Proposition 2.2 *Let $p(m,m')=10\|\Pi\|_{\infty,A}\dfrac{D(m,m')}{n}$ where $D(m,m')$ denotes the dimension of $S_m+S_{m'}$. Then, under the assumptions of Theorem 2.1, there exists a constant C_1 such that*

$$\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left(\left[\sup_{T\in B_f(m,m')}Z_n^2(T)-p(m,m')\right]_+\mathbb{1}_{\Omega^*}\right)\leq\frac{C_1}{n}. \qquad (2.13)$$

Then, with $\kappa=3\rho$, inequalities (2.12) and (2.13) yield

$$\mathbb{E}\left(\|\tilde{\Pi}-\Pi\mathbb{1}_A\|_n^2\mathbb{1}_{\Omega_\rho^*}\right)\leq 5\|f\|_{\infty,A_1}\|\Pi_m-\Pi\mathbb{1}_A\|^2+6\mathrm{pen}(m)+\frac{9\rho C_1}{n} \qquad (2.14)$$

The penalty term pen(m) has to verify $\text{pen}(m) + \text{pen}(m') \geq 30\rho\|\Pi\|_{\infty,A} \dfrac{D(m,m')}{n}$ i.e. $30\rho\|\Pi\|_{\infty,A}\dim(S_m + S_{m'}) \leq \text{pen}(m) + \text{pen}(m')$ We choose $\rho = 3/2$ and so $\text{pen}(m) = 45\|\Pi\|_{\infty,A}\dfrac{D_{m_1}D_{m_2}}{n}$.

To bound the second term in (2.10), we recall (see Section 2.3) that $(\hat{\Pi}_{\hat{m}}(X_i,y))_{1\leq i\leq n}$ is the orthogonal projection of $(\sum_k \psi_k(X_{i+1})\psi_k(y))_{1\leq i\leq n}$ on

$$W = \{(T(X_i,y))_{1\leq i\leq n}, \quad T \in S_{\hat{m}}\}$$

where $\psi_k = \psi_k^{\hat{m}}$. Thus, since P_W denotes the orthogonal projection on W, using (2.6),

$$\begin{aligned}
(\hat{\Pi}_{\hat{m}}(X_i,y))_{1\leq i\leq n} &= P_W((\sum_k \psi_k(X_{i+1})\psi_k(y))_{1\leq i\leq n}) \\
&= P_W((\sum_k \pi_k(X_i)\psi_k(y))_{1\leq i\leq n}) + P_W((\sum_k \varepsilon_{i,k}\psi_k(y))_{1\leq i\leq n}) \\
&= P_W(\Pi\mathbb{1}_A(X_i,y))_{1\leq i\leq n}) + P_W((\sum_k \varepsilon_{i,k}\psi_k(y))_{1\leq i\leq n})
\end{aligned}$$

We denote by $\|.\|_{\mathbb{R}^n}$ the Euclidean norm in $\mathbb{R}^n$, by X the vector $(X_i)_{1\leq i\leq n}$ and by ε_k the vector $(\varepsilon_{i,k})_{1\leq i\leq n}$. Thus

$$\begin{aligned}
\|\Pi\mathbb{1}_A - \hat{\Pi}_{\hat{m}}\|_n^2 &= \frac{1}{n}\int \|\Pi\mathbb{1}_A(X,y) - P_W(\Pi\mathbb{1}_A(X,y)) - P_W(\sum_k \varepsilon_k\psi_k(y))\|_{\mathbb{R}^n}^2 dy \\
&= \frac{1}{n}\int \|\Pi\mathbb{1}_A(X,y) - P_W(\Pi\mathbb{1}_A(X,y))\|_{\mathbb{R}^n}^2 dy + \frac{1}{n}\int \|P_W(\sum_k \varepsilon_k\psi_k(y))\|_{\mathbb{R}^n}^2 dy \\
&\leq \frac{1}{n}\int \|\Pi\mathbb{1}_A(X,y)\|_{\mathbb{R}^n}^2 dy + \frac{1}{n}\int \|\sum_k \varepsilon_k\psi_k(y)\|_{\mathbb{R}^n}^2 dy \\
&\leq \frac{1}{n}\sum_{i=1}^n \|\Pi\|_{\infty,A}\int \Pi(X_i,y)dy + \frac{1}{n}\sum_{i=1}^n \int [\sum_k \varepsilon_{i,k}\psi_k(y)]^2 dy \\
&\leq \|\Pi\|_{\infty,A} + \frac{1}{n}\sum_{i=1}^n \sum_k \varepsilon_{i,k}^2.
\end{aligned}$$

But Assumption M2 implies $\|\sum_{k\in K_{\hat{m}}} \psi_k^2\|_\infty \leq \phi_2 D_{\hat{m}_2}$. So, using (2.7),

$$\begin{aligned}
\varepsilon_{i,k}^2 &\leq 2\psi_k^2(X_{i+1}) + 2\mathbb{E}[\psi_k(X_{i+1})|X_i]^2 \\
\text{and} \quad \sum_k \varepsilon_{i,k}^2 &\leq 2\sum_k \psi_k^2(X_{i+1}) + 2\mathbb{E}[\sum_k \psi_k^2(X_{i+1})|X_i] \leq 4\phi_2 D_{\hat{m}_2}
\end{aligned}$$

Thus we obtain

$$\|\Pi\mathbb{1}_A - \hat{\Pi}_{\hat{m}}\|_n^2 \leq \|\Pi\|_{\infty,A} + 4\phi_2 D_{\hat{m}_2} \leq \|\Pi\|_{\infty,A} + 4\phi_2 n^{1/3} \tag{2.15}$$

and, by taking the expectation, $\mathbb{E}\left(\|\Pi\mathbb{1}_A - \hat{\Pi}_{\hat{m}}\|_n^2 \mathbb{1}_{\Omega_\rho^{*c}}\right) \leq (\|\Pi\|_{\infty,A} + 4\phi_2 n^{1/3}) P(\Omega_\rho^{*c})$.

We now remark that $P(\Omega_\rho^{*c}) = P(\Omega^{*c}) + P(\Omega_\rho^c \cap \Omega^*)$. In the geometric case $\beta_{2q_n} \leq e^{-\theta c \log(n)} \leq n^{-\theta c}$ and in the other case $\beta_{2q_n} \leq (2q_n)^{-\theta} \leq n^{-\theta c}$. Then

$$P(\Omega^{*c}) \leq 4p_n \beta_{2q_n} \leq n^{1-c\theta}.$$

But we have choosed c such that $c\theta > 7/3$ and so $P(\Omega^{*c}) \leq n^{-4/3}$. Now we will use the following proposition:

Proposition 2.3 *Let $\rho > 1$. Then, under the assumptions of Theorem 2.1 or Theorem 2.3, there exists $C_2 > 0$ such that $P(\Omega_\rho^c \cap \Omega^*) \leq \dfrac{C_2}{n^{7/3}}$.*

This proposition implies that $\mathbb{E}\left(\|\Pi\mathbb{1}_A - \hat{\Pi}_{\hat{m}}\|_n^2 \mathbb{1}_{\Omega_\rho^{*c}}\right) \leq \dfrac{C_3}{n}$. Now we use (2.14) and we observe that this inequality holds for all m in $\mathcal{M}_n$, so

$$\mathbb{E}\|\tilde{\Pi} - \Pi\mathbb{1}_A\|_n^2 \leq C \inf_{m \in \mathcal{M}_n} (\|\Pi\mathbb{1}_A - \Pi_m\|^2 + \mathrm{pen}(m)) + \frac{C_4}{n}$$

with $C = \max(5\|f\|_{\infty,A_1}, 6)$. □

2.7.3 Proof of Proposition 2.2

Let $\begin{cases} \Gamma_i(T) = T(X_i, X_{i+1}) - \int T(X_i, y)\Pi(X_i, y)dy, \\ \Gamma_i^*(T) = T(X_i^*, X_{i+1}^*) - \int T(X_i^*, y)\Pi(X_i^*, y)dy, \\ \Gamma_i^{**}(T) = T(X_i^{**}, X_{i+1}^{**}) - \int T(X_i^{**}, y)\Pi(X_i^{**}, y)dy. \end{cases}$

We now define $Z_n^*(T)$:

$$Z_n^*(T) = \frac{1}{n}\sum_{i \text{ odd}} \Gamma_i^*(T) + \frac{1}{n}\sum_{i \text{ even}} \Gamma_i^{**}(T).$$

Let us remark that $Z_n^*(T)\mathbb{1}_{\Omega^*} = Z_n(T)\mathbb{1}_{\Omega^*}$. Next we split each of these terms :

$$Z_{n,1}^*(T) = \frac{1}{n}\sum_{l=0}^{p_n-1} \sum_{i=4lq_n+1, i \text{ odd}}^{2(2l+1)q_n-1} \Gamma_i^*(T), \quad Z_{n,2}^*(T) = \frac{1}{n}\sum_{l=0}^{p_n-1} \sum_{i=2(2l+1)q_n+1, i \text{ odd}}^{2(2l+2)q_n-1} \Gamma_i^*(T),$$

$$Z_{n,3}^*(T) = \frac{1}{n}\sum_{l=0}^{p_n-1} \sum_{i=4lq_n+2, i \text{ even}}^{2(2l+1)q_n} \Gamma_i^{**}(T), \quad Z_{n,4}^*(T) = \frac{1}{n}\sum_{l=0}^{p_n-1} \sum_{i=2(2l+1)q_n+2, i \text{ even}}^{2(2l+2)q_n} \Gamma_i^{**}(T).$$

We use the following lemma, proved in Chapter 4 Section 4.7.8:

Lemma 2.1 *Let $U_1, \dots, U_N$ be independent random variables and $G(r) = (1/N)\sum_{i=1}^N [r(U_i) - \mathbb{E}(r(U_i))]$, for r belonging to a class $\mathcal{R}$ of measurable functions. We suppose that*

(1) $\sup_{r\in\mathcal{R}} \|r\|_\infty \le M_1$, *(2)* $\mathbb{E}\left(\sup_{r\in\mathcal{R}} |G(r)|\right) \le H$, *(3)* $\sup_{r\in\mathcal{R}} \frac{1}{N}\sum_{i=1}^N \mathrm{Var}(r(U_i)) \le v$.

Then, there exists $K > 0$, $K_1 > 0$, $K_2 > 0$ such that

$$\mathbb{E}[\sup_{r\in\mathcal{R}} |G(r)|^2 - 10H^2]_+ \le K\left(\frac{v}{N}e^{-K_1\frac{NH^2}{v}} + \frac{M_1^2}{N^2}e^{-K_2\frac{NH}{M_1}}\right)$$

Here $N = p_n$ and for $l \in \{0, \dots, p_n - 1\}$, $U_{l+1} = (X^*_{4lq_n+1}, .., X^*_{2(2l+1)q_n})$,

$$r_T(x_1, \dots, x_{2q_n}) = \frac{1}{q_n}\sum_{i=1, i \text{ odd}}^{2q_n-1} T(x_i, x_{i+1}) - \int T(x_i, y)\Pi(x_i, y)dy$$

and $\mathcal{R} = \{r_T, T \in B_f(m, m')\}$. Then

$$G(r_T) = \frac{1}{p_n}\sum_{l=0}^{p_n-1}\frac{1}{q_n}\sum_{i=4lq_n+1, i \text{ odd}}^{2(2l+1)q_n-1} \Gamma_i^*(T) = 4Z^*_{n,1}(T).$$

We now compute M_1, H and v.

(1)We recall that $S_m + S_{m'}$ is included in the model $S_{m''}$ with dimension $\max(D_{m_1}, D_{m_1'})$ $\max(D_{m_2}, D_{m_2'})$.

$$\begin{aligned}\sup_{T\in B}\|r_T\|_\infty &\le \sup_{T\in B}\|T\|_\infty \frac{1}{q_n}\sum_{i=1, i \text{ odd}}^{2q_n-1}\left(1 + \int \Pi(x_i, y)dy\right)\\ &\le 2\phi_0\sqrt{\max(D_{m_1}, D_{m_1'})\max(D_{m_2}, D_{m_2'})}\sup_{T\in B}\|T\| \le \frac{2\phi_0}{f_0}n^{1/3}.\end{aligned}$$

Then we set $M_1 = \dfrac{2\phi_0}{f_0}n^{1/3}$.

(2) Since A_0 and A_0^* have the same distribution, $r_T(U_1) = \frac{1}{q_n}\sum_{i=1, i \text{ odd}}^{2q_n-1} \Gamma_i^*(T)$ has the same distribution than $\frac{1}{q_n}\sum_{i=1, i \text{ odd}}^{2q_n-1} \Gamma_i(T)$. We observe that $\mathbb{E}(\Gamma_i(T)|X_i) = 0$ and then for all set I

$$\begin{aligned}&\mathbb{E}\left(\left[\sum_{i\in I}\Gamma_i(T)\right]^2\right) = \mathbb{E}\left(\sum_{i,j\in I}\Gamma_i(T)\Gamma_j(T)\right)\\ &= 2\mathbb{E}\left(\sum_{j<i}\mathbb{E}[\Gamma_i(T)\Gamma_j(T)|X_1, \dots, X_i]\right) + \sum_{i\in I}\mathbb{E}\left[\Gamma_i^2(T)\right]\\ &= 2\mathbb{E}\left(\sum_{j<i}\Gamma_j(T)\mathbb{E}[\Gamma_i(T)|X_i]\right) + \sum_{i\in I}\mathbb{E}\left[\Gamma_i^2(T)\right] = \sum_{i\in I}\mathbb{E}\left[\Gamma_i^2(T)\right].\end{aligned}$$

In particular

$$\begin{aligned}\mathrm{Var}[r_T(U_1)] &= \mathbb{E}\left(\left[\frac{1}{q_n}\sum_{i=1,i\text{ odd}}^{2q_n-1}\Gamma_i(T)\right]^2\right) = \frac{1}{q_n^2}\sum_{i=1,i\text{ odd}}^{2q_n-1}\mathbb{E}\left[\Gamma_i^2(T)\right]\\ &\le \frac{1}{q_n^2}\sum_{i=1,i\text{ odd}}^{2q_n-1}\mathbb{E}\left[T^2(X_i,X_{i+1})\right] \le \frac{1}{q_n}\|\Pi\|_{\infty,A}\|T\|_f^2.\end{aligned}$$

Then $v = \dfrac{\|\Pi\|_{\infty,A}}{q_n}$.

(3) Let $(\bar\varphi_j\otimes\psi_k)_{(j,k)\in\Lambda(m,m')}$ an orthonormal basis of $(S_m+S_{m'},\|.\|_f)$.

$$\begin{aligned}\mathbb{E}(\sup_{T\in B}|G^2(r_T)|) &\le \sum_{j,k}\mathbb{E}(G^2(r_{\bar\varphi_j\otimes\psi_k}))\\ &\le \sum_{j,k}\frac{1}{p_n^2q_n^2}\mathbb{E}\left(\left[\sum_{l=0}^{p_n-1}\sum_{i=4lq_n+1,i\text{ odd}}^{2(2l+1)q_n-1}\Gamma_i^*(\bar\varphi_j\otimes\psi_k)\right]^2\right)\\ &\le \sum_{j,k}\frac{16}{n^2}\sum_{l=0}^{p_n-1}\mathbb{E}\left(\left[\sum_{i=4lq_n+1,i\text{ odd}}^{2(2l+1)q_n-1}\Gamma_i^*(\bar\varphi_j\otimes\psi_k)\right]^2\right)\end{aligned}$$

where we used the independence of the A_l^*. Now we can replace Γ_i^* by Γ_i in the sum because A_l and A_l^* have the same distribution and we use as previously the martingale property of the Γ_i.

$$\begin{aligned}\mathbb{E}(\sup_{T\in B}|G^2(r_T)|) &\le \sum_{j,k}\frac{16}{n^2}\sum_{l=0}^{p_n-1}\mathbb{E}\left(\left[\sum_{i=4lq_n+1,i\text{ odd}}^{2(2l+1)q_n-1}\Gamma_i(\bar\varphi_j\otimes\psi_k)\right]^2\right)\\ &\le \sum_{j,k}\frac{16}{n^2}\sum_{l=0}^{p_n-1}\sum_{i=4lq_n+1,i\text{ odd}}^{2(2l+1)q_n-1}\mathbb{E}\left(\Gamma_i^2(\bar\varphi_j\otimes\psi_k)\right)\\ &\le \sum_{j,k}\frac{4}{n}\|\Pi\|_{\infty,A}\|\bar\varphi_j\otimes\psi_k\|_f^2 \le 4\|\Pi\|_{\infty,A}\frac{D(m,m')}{n}.\end{aligned}$$

Then $\mathbb{E}^2(\sup_{T\in B}|G(r_T)|) \le 4\|\Pi\|_{\infty,A}\dfrac{D(m,m')}{n}$ and $H^2 = 4\|\Pi\|_{\infty,A}\dfrac{D(m,m')}{n}$.

Remark 2.2 *We have also*

$$\mathbb{E}(\sup_{T\in B}|G^2(r_T)|) \le \frac{4}{n}\|\sum_j\bar\varphi_j^2\|_\infty\|\sum_k\psi_k^2\|_\infty \le 4\phi_0^2f_0^{-1}\frac{\dim(S_{m''})}{n}$$

using property M2. So we could choose $H^2 = 4\phi_0^2 f_0^{-1}\dim(S_{m''})/n$ *where* $S_{m''}$ *has dimension* $\max(D_{m_1}, D_{m_1'})\max(D_{m_2}, D_{m_2'})$. *It is a bad choice for an anisotropic estimation but in the case of an isotropic estimation, the spaces* S_m *are nested so that* $\dim(S_m + S_{m'}) = \dim(S_{m''})$ *and it allows to avoid the term* $\|\Pi\|_{\infty,A}$ *in* H^2 *and then in the penalty.*

According to Lemma 2.1, there exists $K' > 0$, $K_1 > 0$, $K_2' > 0$ such that

$$\mathbb{E}\left[\sup_{T\in B_f(m,m')}(4Z_{n,1}^*)^2(T) - 10H^2\right]_+ \leq K'\left[\frac{1}{n}e^{-K_1 D(m,m')} + n^{-4/3}q_n^2 e^{-K_2' n^{1/6}\sqrt{D(m,m')}/q_n}\right].$$

But $q_n \leq n^c$ with $c < \frac{1}{6}$. So

$$\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left[\sup_{T\in B_f(m,m')} Z_{n,1}^{*2}(T) - \frac{p(m,m')}{4}\right]_+$$
$$\leq \frac{K'}{n}\left[\sum_{m'\in\mathcal{M}_n} e^{-K_1 D(m,m')} + n^{2c-1/3}|\mathcal{M}_n|e^{-K_2' n^{1/6-c}}\right] \leq \frac{A_1}{n}. \tag{2.16}$$

In the same way, $\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left[\sup_{T\in B_f(m,m')} Z_{n,r}^{*2}(T) - p(m,m')/4\right]_+ \leq A_r/n$ for $r = 2, 3, 4$. And then

$$\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left(\left[\sup_{T\in B_f(m,m')} Z_n^2(T) - p(m,m')\right]_+ \mathbb{1}_{\Omega^*}\right)$$
$$= \sum_{m'\in\mathcal{M}_n}\mathbb{E}\left(\left[\sup_{T\in B_f(m,m')} Z_n^{*2}(T) - p(m,m')\right]_+ \mathbb{1}_{\Omega^*}\right) \leq \frac{C_1}{n}.$$

□

2.7.4 Proof of Proposition 2.3

First we observe that

$$P(\Omega_\rho^c \cap \Omega^*) \leq P\left(\sup_{T\in\mathcal{B}}|\nu_n(T^2)| > 1 - 1/\rho\right)$$

where $\nu_n(T) = \frac{1}{n}\sum_{i=1}^n \int [T(X_i^*, y) - \mathbb{E}(T(X_i^*, y))]dy$ and $\mathcal{B} = \{T \in \mathcal{S} \quad \|T\|_f = 1\}$.
But, if $T(x,y) = \sum_{j,k} a_{j,k}\varphi_j(x)\psi_k(y)$, then

$$\nu_n(T^2) = \sum_{j,j'}\sum_k a_{j,k}a_{j',k}\bar{\nu}_n(\varphi_j\varphi_{j'})$$

where

$$\bar{\nu}_n(u) = \frac{1}{n}\sum_{i=1}^{n}[u(X_i^*) - \mathbb{E}(u(X_i^*))]. \tag{2.17}$$

Let $b_j = (\sum_k a_{j,k}^2)^{1/2}$, then $|\nu_n(T^2)| \leq \sum_{j,j'} b_j b_{j'} |\bar{\nu}_n(\varphi_j\varphi_{j'})|$ and, if $T \in \mathcal{B}$,

$$\sum_j b_j^2 = \sum_j \sum_k a_{j,k}^2 = \|T\|^2 \leq f_0^{-1}.$$

Thus

$$\sup_{T\in\mathcal{B}} |\nu_n(T^2)| \leq f_0^{-1} \sup_{\sum b_j^2=1} \sum_{j,l} b_j b_l |\bar{\nu}_n(\varphi_j\varphi_l)|.$$

Lemma 2.2 *Let $B_{j,l} = \|\varphi_j\varphi_l\|_\infty$ and $V_{j,l} = \|\varphi_j\varphi_l\|$. Let, for any symmetric matrix $(A_{j,l})$*

$$\bar{\rho}(A) = \sup_{\sum a_j^2=1} \sum_{j,l} |a_j a_l| A_{j,l}$$

and $L(\varphi) = \max\{\bar{\rho}^2(V), \bar{\rho}(B)\}$. Then, if M2 is satisfied, $L(\varphi) \leq \phi_1 \mathcal{D}_n^2$.

This lemma is proved in Baraud *et al.* (2001).

Let $x = \dfrac{f_0^2(1-1/\rho)^2}{40\|f\|_{\infty,A_1} L(\varphi)}$ and $\Delta = \left\{\forall j \forall l \quad |\bar{\nu}_n(\varphi_j\varphi_l)| \leq 4\left[B_{j,l}x + V_{j,l}\sqrt{2\|f\|_{\infty,A_1}x}\right]\right\}$.
On the set Δ:

$$\begin{aligned}
\sup_{T\in\mathcal{B}} |\nu_n(T^2)| &\leq 4f_0^{-1} \sup_{\sum b_j^2=1} \sum_{j,l} b_j b_l \left[B_{j,l}x + V_{j,l}\sqrt{2\|f\|_{\infty,A_1}x}\right] \\
&\leq 4f_0^{-1}\left[\bar{\rho}(B)x + \bar{\rho}(V)\sqrt{2\|f\|_{\infty,A_1}x}\right] \\
&\leq (1-1/\rho)\left[\frac{f_0(1-1/\rho)}{10\|f\|_{\infty,A_1}}\frac{\bar{\rho}(B)}{L(\varphi)} + \frac{2}{\sqrt{5}}\left(\frac{\bar{\rho}^2(V)}{L(\varphi)}\right)^{1/2}\right] \\
&\leq (1-1/\rho)\left[\frac{1}{10} + \frac{2}{\sqrt{5}}\right] \leq (1-1/\rho).
\end{aligned}$$

Then $P\left(\sup_{T\in\mathcal{B}} |\nu_n(T^2)| > 1 - \dfrac{1}{\rho}\right) \leq P(\Delta^c)$. But $\bar{\nu}_n(u) = 2\bar{\nu}_{n,1}(u) + 2\bar{\nu}_{n,2}(u)$ with

$$\bar{\nu}_{n,r}(u) = \frac{1}{p_n}\sum_{l=0}^{p_n-1} Y_{l,r}(u) \qquad r = 1, 2$$

with $\begin{cases} Y_{l,1}(u) &= \dfrac{1}{2q_n}\sum_{i=4lq_n+1}^{2(2l+1)q_n}[u(X_i^*) - \mathbb{E}(u(X_i^*))], \\ Y_{l,2}(u) &= \dfrac{1}{2q_n}\sum_{i=2(2l+1)q_n+1}^{2(2l+2)q_n}[u(X_i^*) - \mathbb{E}(u(X_i^*))]. \end{cases}$

To bound $P(\bar{\nu}_{n,1}(\varphi_j\varphi_l) \geq B_{j,l}x + V_{j,l}\sqrt{2\|f\|_{\infty,A_1}x})$, we will use the Bernstein inequality given in Birgé and Massart (1998). That is why we bound $\mathbb{E}|Y_{l,1}(u)|^m$:

$$\begin{aligned}
\mathbb{E}|Y_{l,1}(u)|^m &\leq \frac{1}{4q_n^2}(2\|u\|_\infty)^{m-2}\mathbb{E}\left|\sum_{i=4lq_n+1}^{2(2l+1)q_n}[u(X_i^*) - \mathbb{E}(u(X_i^*))]\right|^2 \\
&\leq (2\|u\|_\infty)^{m-2}\frac{1}{4q_n^2}\mathbb{E}\left|\sum_{i=4lq_n+1}^{2(2l+1)q_n}[u(X_i) - \mathbb{E}(u(X_i))]\right|^2 \\
&\leq (2\|u\|_\infty)^{m-2}\frac{1}{2q_n}\sum_{i=4lq_n+1}^{2(2l+1)q_n}\mathbb{E}\left[u(X_1) - \mathbb{E}(u(X_1))\right]|^2
\end{aligned}$$

since $X_i^* = X_i$ on Ω^* and the X_i have the same distribution than X_1. Thus

$$\begin{aligned}
\mathbb{E}|Y_{l,1}(u)|^m &\leq (2\|u\|_\infty)^{m-2}\mathbb{E}|u(X_1) - \mathbb{E}(u(X_1))|^2 \leq (2\|u\|_\infty)^{m-2}\int u^2(x)f(x)dx \\
&\leq 2^{m-2}(\|u\|_\infty)^{m-2}(\sqrt{\|f\|_{\infty,A_1}}\|u\|)^2. \qquad (2.18)
\end{aligned}$$

With $u = \varphi_j\varphi_{j'}$, $\mathbb{E}|Y_{l,1}(\varphi_j\varphi_{j'})|^m \leq 2^{m-2}(B_{j,j'})^{m-2}(\sqrt{\|f\|_{\infty,A_1}}V_{j,j'})^2$. And then

$$P(|\bar{\nu}_{n,r}(\varphi_j\varphi_l)| \geq B_{j,l}x + V_{j,l}\sqrt{2\|f\|_{\infty,A_1}x}) \leq 2e^{-p_n x}.$$

Given that $P(\Omega_\rho^c \cap \Omega^*) \leq P(\Delta^c) = \sum_{j,l} P\left(|\bar{\nu}_n(\varphi_j\varphi_l)| > 4(B_{j,l}x + V_{j,l}\sqrt{2\|f\|_{\infty,A_1}x})\right)$,

$$\begin{aligned}
P(\Omega_\rho^c \cap \Omega^*) &\leq 4\mathcal{D}_n^2 \exp\left\{-\frac{p_n f_0^2(1-1/\rho)^2}{40\|f\|_{\infty,A_1}L(\varphi)}\right\} \\
&\leq 4n^{2/3}\exp\left\{-\frac{f_0^2(1-1/\rho)^2}{160\|f\|_{\infty,A_1}}\frac{n}{q_nL(\varphi)}\right\}.
\end{aligned}$$

But $L(\varphi) \leq \phi_1\mathcal{D}_n^2 \leq \phi_1 n^{2/3}$ and $q_n \leq n^{1/6}$ so

$$P(\Omega_\rho^c \cap \Omega^*) \leq 4n^{2/3}\exp\left\{-\frac{f_0^2(1-1/\rho)^2}{160\|f\|_{\infty,A_1}\phi_1}n^{1/6}\right\} \leq \frac{C}{n^{7/3}}. \qquad (2.19)$$

□

2.7.5 Proof of Theorem 2.2

We recall that $\|\Pi\|_{\infty,A}$ denotes $\|\Pi \mathbb{1}_A\|_\infty$ and we introduce the following set:

$$\Lambda = \left\{ \left| \frac{\|\hat{\Pi}\|_\infty}{\|\Pi \mathbb{1}_A\|_\infty} - 1 \right| < \frac{1}{2} \right\}.$$

As previously, we decompose the space:

$$\mathbb{E}\|\tilde{\Pi} - \Pi \mathbb{1}_A\|_n^2 = \mathbb{E}\left(\|\tilde{\Pi} - \Pi \mathbb{1}_A\|_n^2 \mathbb{1}_{\Omega_\rho^* \cap \Lambda}\right) + \mathbb{E}\left(\|\tilde{\Pi} - \Pi \mathbb{1}_A\|_n^2 \mathbb{1}_{\Omega_\rho^* \cap \Lambda^c}\right) + \mathbb{E}\left(\|\tilde{\Pi} - \Pi \mathbb{1}_A\|_n^2 \mathbb{1}_{\Omega_\rho^{*c}}\right)$$

We have already dealt with the third term. For the first term, we can proceed as in the proof of Theorem 2.1 as soon as

$$\kappa p(m, m') \leq \overline{\text{pen}}(m) + \overline{\text{pen}}(m')$$

with $\kappa = 3\rho = 9/2$ and $p(m, m') = 10\|\Pi\|_{\infty,A} D(m, m')/n$. But on Λ, $\|\Pi\|_{\infty,A} < 2\|\hat{\Pi}\|_\infty$ and so

$$\begin{aligned} \kappa p(m, m') &= 10\kappa \|\Pi\|_{\infty,A} \frac{D(m, m')}{n} \leq 20\kappa \|\hat{\Pi}\|_\infty \frac{D(m, m')}{n} \\ &\leq 20\kappa \|\hat{\Pi}\|_\infty \frac{D_{m_1} D_{m_2}}{n} + 20\kappa \|\hat{\Pi}\|_\infty \frac{D_{m_1'} D_{m_2'}}{n} \end{aligned}$$

It is sufficient to set $\overline{K_0} = 20\kappa$.

Now, inequality (2.15) gives

$$\mathbb{E}\left(\|\Pi \mathbb{1}_A - \hat{\Pi}_{\hat{m}}\|_n^2 \mathbb{1}_{\Omega_\rho^* \cap \Lambda^c}\right) \leq (\|\Pi\|_{\infty,A} + 4\phi_2 n^{1/3}) P(\Omega_\rho^* \cap \Lambda^c).$$

It remains to prove that $P(\Omega_\rho^* \cap \Lambda^c) \leq C n^{-4/3}$ for some constant C.

$$\begin{aligned} P(\Omega_\rho^* \cap \Lambda^c) &= P(|\|\hat{\Pi}\|_\infty - \|\Pi \mathbb{1}_A\|_\infty| \mathbb{1}_{\Omega_\rho^*} \geq \|\Pi\|_{\infty,A}/2) \leq P(\|\hat{\Pi} - \Pi \mathbb{1}_A\|_\infty \mathbb{1}_{\Omega_\rho^*} \geq \|\Pi\|_{\infty,A}/2) \\ &\leq P(\|\hat{\Pi} - \Pi_{m*}\|_\infty \mathbb{1}_{\Omega_\rho^*} \geq \|\Pi\|_{\infty,A}/4) + P(\|\Pi_{m*} - \Pi \mathbb{1}_A\|_\infty \geq \|\Pi\|_{\infty,A}/4) \\ &\leq P\left(\|\hat{\Pi} - \Pi_{m*}\| \mathbb{1}_{\Omega_\rho^*} \geq \frac{\|\Pi\|_{\infty,A}}{4\phi_0 \sqrt{D_{m_1*} D_{m_2*}}}\right) + P(\|\Pi_{m*} - \Pi \mathbb{1}_A\|_\infty \geq \|\Pi\|_{\infty,A}/4) \end{aligned}$$

since $\|\hat{\Pi} - \Pi_{m*}\|_\infty \leq \phi_0 \sqrt{D_{m*_1} D_{m*_2}} \|\hat{\Pi} - \Pi_{m^*}\|$.

Furthermore the inequality $\gamma_n(\hat{\Pi}) \leq \gamma_n(\Pi_{m*})$ leads to

$$\|\hat{\Pi} - \Pi \mathbb{1}_A\|_n^2 \leq \|\Pi_{m*} - \Pi \mathbb{1}_A\|_n^2 + \frac{1}{\kappa'} \|\hat{\Pi} - \Pi_{m*}\|_f^2 + \kappa' \sup_{T \in B_f(m*)} Z_n^2(T)$$

where $B_f(m) = \{t \in S_m \quad \|t\|_f = 1\}$ and then, on Ω_ρ,

$$\begin{aligned}
\|\hat{\Pi} - \Pi_{m*}\|_f^2 \left(1 - \frac{2\rho}{\kappa'}\right) &\leq 4\rho\|\Pi_{m*} - \Pi\mathbb{1}_A\|_n^2 + 2\rho\kappa' \sup_{T\in B_f(m*)} Z_n^2(T) \\
\text{so} \qquad \|\hat{\Pi} - \Pi_{m*}\|^2 &\leq \frac{4\rho\kappa' f_0^{-1}}{\kappa' - 2\rho}\|\Pi_{m*} - \Pi\mathbb{1}_A\|_n^2 + \frac{2\rho\kappa'^2 f_0^{-1}}{\kappa' - 2\rho} \sup_{T\in B_f(m*)} Z_n^2(T) \\
&\leq 12\rho f_0^{-1}|A_2|\|\Pi_{m*} - \Pi\mathbb{1}_A\|_\infty^2 + 18\rho^2 f_0^{-1} \sup_{T\in B_f(m*)} Z_n^2(T)
\end{aligned}$$

with $\kappa' = 3\rho$ and by remarking that for T with support A, $\|T\|_n^2 \leq |A_2|\|T\|_\infty^2$. Thus

$$\begin{aligned}
P(\Omega_\rho^* \cap \Lambda^c) \leq\ & P(\sup_{T\in B_f(m*)} Z_n^2(T)\mathbb{1}_{\Omega_\rho^*} \geq \frac{\|\Pi\|_{\infty,A}^2}{32\phi_0^2 n^{1/3}} \frac{1}{18\rho^2 f_0^{-1}}) \\
&+ P(\|\Pi_{m*} - \Pi\mathbb{1}_A\|_\infty^2 \geq \frac{\|\Pi\|_{\infty,A}^2}{32\phi_0^2 D_{m_1*} D_{m_2*}} \frac{1}{12\rho f_0^{-1}|A_2|}) \\
&+ P(\|\Pi_{m*} - \Pi\mathbb{1}_A\|_\infty \geq \|\Pi\|_{\infty,A}/4) \\
\leq\ & P(\sup_{T\in B_f(m*)} Z_n^2(T)\mathbb{1}_{\Omega^*} \geq \frac{a}{n^{1/3}}) + P(D_{m_1*}D_{m_2*}\|\Pi_{m*} - \Pi\mathbb{1}_A\|_\infty^2 \geq b) \\
&+ P(\|\Pi_{m*} - \Pi\mathbb{1}_A\|_\infty \geq \frac{\|\Pi\|_{\infty,A}}{4})
\end{aligned} \tag{2.20}$$

with $a = \dfrac{\|\Pi\|_{\infty,A}^2}{32\phi_0^2} \dfrac{1}{18\rho^2 f_0^{-1}}$ and $b = \dfrac{\|\Pi\|_{\infty,A}^2}{32\phi_0^2} \dfrac{1}{12\rho f_0^{-1}|A_2|}$.

We will first study the two last terms in (2.20). Since the restriction Π_A of Π belongs to $B_{2,\infty}^{(\alpha_1,\alpha_2)}(A)$, the imbedding theorem proved in Nikol′skiĭ (1975) p.236 implies that Π_A belongs to $B_{\infty,\infty}^{(\beta_1,\beta_2)}(A)$ with $\beta_1 = \alpha_1(1 - 1/\bar{\alpha})$ and $\beta_2 = \alpha_2(1 - 1/\bar{\alpha})$. Then the approximation lemma 2.3 (which is still valid for the trigonometric polynomial spaces with the infinite norm instead of the L^2 norm) yields to

$$\|\Pi_{m*} - \Pi\mathbb{1}_A\|_\infty \leq C(D_{m_1*}^{-\beta_1} + D_{m_2*}^{-\beta_2}).$$

And then, since $D_{m_1*} = D_{m_2*}$,

$$\begin{aligned}
D_{m_1*}D_{m_2*}\|\Pi_{m*} - \Pi\mathbb{1}_A\|_\infty^2 &\leq C'(D_{m_1*}^{2-2\beta_1} + D_{m_1*}^{2-2\beta_2}) \\
&\leq C'((\log n)^{2-2\beta_1} + (\log n)^{2-2\beta_2}) \to 0.
\end{aligned}$$

Indeed $\begin{cases} 2 - 2\beta_1 < 0 \Leftrightarrow 2\alpha_1\alpha_2 - 3\alpha_2 - \alpha_1 > 0 \\ 2 - 2\beta_2 < 0 \Leftrightarrow 2\alpha_1\alpha_2 - 3\alpha_1 - \alpha_2 > 0 \end{cases}$ and this double condition is ensured when $\alpha_1 > 3/2$ and $\alpha_2 > \max(\frac{\alpha_1}{2\alpha_1 - 3}, \frac{3\alpha_1}{2\alpha_1 - 1})$. Consequently, for n large enough,

$$P(D_{m_1*}D_{m_2*}\|\Pi_{m*} - \Pi\mathbb{1}_A\|_\infty^2 \geq b) + P(\|\Pi_{m*} - \Pi\|_\infty \geq \frac{\|\Pi\|_{\infty,A}}{4}) = 0.$$

We will now prove that

$$P\left(\sup_{T\in B_f(m*)} Z_n^2(T)\mathbb{1}_{\Omega^*} \geq \frac{a}{n^{1/3}}\right) \leq \frac{C}{n^{4/3}}$$

and then using (2.20), we will have $P(\Omega_\rho^* \cap \Lambda^c) \leq Cn^{-4/3}$. We remark that, if $(\varphi_j \otimes \psi_k)_{j,k}$ is a base of $(S_{m*}, \|.\|_f)$,

$$\sup_{T\in B_f(m*)} Z_n^2(T) \leq \sum_{j,k} Z_n^2(\varphi_j \otimes \psi_k)$$

and we recall that, on Ω^*, $Z_n(T) = \sum_{r=1}^4 Z_{n,r}^*(T)$ (see the proof of Proposition 2.2). So we are interested in

$$P\left(Z_{n,1}^{*2}(\varphi_j \otimes \psi_k)\mathbb{1}_{\Omega^*} \geq \frac{a}{4D_{m_1*}D_{m_2*}n^{1/3}}\right).$$

Let $x = Bn^{-2/3}$ with B such that $2f_0^{-2}B^2 + 4\|\Pi\|_{\infty,A}B \leq a/4$ (for example $B = \inf(1, a/8(f_0^{-2} + 2\|\Pi\|_{\infty,A}))$. Then

$$(\sqrt{2\|\Pi\|_{\infty,A}x} + \sqrt{D_{m_1*}D_{m_2*}}f_0^{-1}x)^2 \leq \frac{a}{4D_{m_1*}D_{m_2*}n^{1/3}}.$$

So we will now bound $P(Z_{n,1}^*(\varphi_j \otimes \psi_k)\mathbb{1}_{\Omega^*} \geq \sqrt{2\|\Pi\|_{\infty,A}x} + \sqrt{D_{m_1*}D_{m_2*}}f_0^{-1}x)$ by using the Bernstein inequality given in Birgé and Massart (1998). That is why we bound $\mathbb{E}|\frac{1}{4q_n}\sum_{i=1,i \text{ odd}}^{2q_n-1}\Gamma_i^*(T)|^m$ for all integer $m \geq 2$,

$$\begin{aligned}
&\mathbb{E}|\frac{1}{4q_n}\sum_{i=1,i \text{ odd}}^{2q_n-1}\Gamma_i^*(T)|^m \\
&\leq \frac{(2\|T\|_\infty q_n)^{m-2}}{(4q_n)^m}\mathbb{E}\left|\sum_{i=1,i \text{ odd}}^{2q_n-1}[T(X_i^*, X_{i+1}^*) - \int T(X_i^*, y)\Pi(X_i^*, y)dy]\right|^2 \\
&\leq \left(\frac{\|T\|_\infty}{2}\right)^{m-2}\frac{1}{16q_n^2}\mathbb{E}\left|\sum_{i=1,i \text{ odd}}^{2q_n-1}[T(X_i, X_{i+1}) - \int T(X_i, y)\Pi(X_i, y)dy]\right|^2 \\
&\leq \left(\frac{\|T\|_\infty}{2}\right)^{m-2}\frac{1}{16}\int T^2(x,y)f(x)\Pi(x,y)dxdy \\
&\leq \frac{1}{2^{m+2}}(\|T\|_\infty)^{m-2}\|\Pi\|_{\infty,A}\|T\|_f^2.
\end{aligned}$$

Then

$$\mathbb{E}|\frac{1}{4q_n}\sum_{i=1,i\text{ odd}}^{2q_n-1}\Gamma_{i*}(\varphi_j\otimes\psi_k)|^m \;\leq\; \frac{1}{2^{m+2}}(\sqrt{D_{m_1*}D_{m_2*}}f_0^{-1})^{m-2}\|\Pi\|_{\infty,A}.$$

Thus the Bernstein inequality gives

$$P(|Z^*_{n,1}(\varphi_j\otimes\psi_k)|\geq\sqrt{D_{m_1*}D_{m_2*}}f_0^{-1}x+\sqrt{2\|\Pi\|_{\infty,A}x})\leq 2e^{-p_nx}.$$

Hence

$$\begin{aligned}P(\sup_{T\in B_f(m*)}Z^{*2}_{n,1}(T)\mathbb{1}_{\Omega^*}\geq\frac{a}{4n^{1/3}}) &\leq 2D_{m_1*}D_{m_2*}\exp\{-p_nBn^{-2/3}\}\\ &\leq 2n^{2/3}\exp\{-\frac{B}{4}\frac{n^{1/3}}{q_n}\}.\end{aligned}$$

But $2n^{2/3}\exp\{-\frac{B}{4}\frac{n^{1/3}}{q_n}\}\leq Cn^{-4/3}$ since $q_n\leq n^{1/6}$ and so

$$P\left(\sup_{T\in B_f(m*)}Z_n^2(T)\mathbb{1}_{\Omega^*}\geq\frac{a}{n^{1/3}}\right)\leq\frac{4C}{n^{4/3}}.$$

□

2.7.6 Proof of Corollary 2.1

To control the bias term, we use the following lemma

Lemma 2.3 *Let Π_A belong to $B^{\boldsymbol{\alpha}}_{2,\infty}(A)$. We consider that S'_m is one of the following spaces on A:*

- *a space of piecewise polynomials of degrees bounded by $s_i>\alpha_i-1$ $(i=1,2)$ based on a partition with rectangles of vertices $1/D_{m_1}$ and $1/D_{m_2}$,*
- *a linear span of $\{\phi_\lambda\psi_\mu,\lambda\in\cup_0^{m_1}\Lambda(j),\mu\in\cup_0^{m_2}M(k)\}$ where $\{\phi_\lambda\}$ and $\{\psi_\mu\}$ are orthonormal wavelet bases of respective regularities $s_1>\alpha_1-1$ and $s_2>\alpha_2-1$ (here $D_{m_i}=2^{m_i},i=1,2$),*
- *the space of trigonometric polynomials with degree smaller than D_{m_1} in the first direction and smaller than D_{m_2} in the second direction.*

Let Π'_m be the orthogonal projection of Π_A on S'_m. Then, there exists a positive constant C_0 such that

$$\left(\int_A|\Pi_A-\Pi'_m|^2\right)^{1/2}\leq C_0[D_{m_1}^{-\alpha_1}+D_{m_2}^{-\alpha_2}].$$

Proof: It is proved in Hochmuth (2002) for S'_m a space of wavelets or polynomials and in Nikol′skiĭ (1975) (p. 191 and 200) for a space of trigonometric polynomials that

$$\left(\int_A |\Pi_A - \Pi'_m|^2\right)^{1/2} \leq C[\omega_{s_1+1,1}(\Pi, D_{m_1}^{-1}) + \omega_{s_2+1,2}(\Pi, D_{m_2}^{-1})].$$

The definition of $B^{\boldsymbol{\alpha}}_{2,\infty}(A)$ implies $\left(\int_A |\Pi_A - \Pi'_m|^2\right)^{1/2} \leq C_0[D_{m_1}^{-\alpha_1} + D_{m_2}^{-\alpha_2}]$. □

If we choose for S'_m the set of the restrictions to A of the functions of S_m and Π_A the restriction of Π to A, we can apply Lemma 2.3. But Π'_m is also the restriction to A of Π_m so that

$$\|\Pi \mathbb{1}_A - \Pi_m\| \leq C_0[D_{m_1}^{-\alpha_1} + D_{m_2}^{-\alpha_2}].$$

According to Theorem 2.1

$$\mathbb{E}\|\tilde{\Pi} - \Pi \mathbb{1}_A\|_n^2 \leq C'' \inf_{m \in \mathcal{M}_n} \left\{ D_{m_1}^{-2\alpha_1} + D_{m_2}^{-2\alpha_2} + \frac{D_{m_1} D_{m_2}}{n} \right\}.$$

In particular, if m^* is such that $D_{m_1^*} = \lfloor n^{\frac{\alpha_2}{\alpha_1+\alpha_2+2\alpha_1\alpha_2}} \rfloor$ and $D_{m_2^*} = \lfloor (D_{m_1^*})^{\frac{\alpha_1}{\alpha_2}} \rfloor$ then

$$\mathbb{E}\|\tilde{\Pi} - \Pi \mathbb{1}_A\|_n^2 \leq C''' \left\{ D_{m_1^*}^{-2\alpha_1} + \frac{D_{m_1^*}^{1+\alpha_1/\alpha_2}}{n} \right\} = O\left(n^{-\frac{2\alpha_1\alpha_2}{\alpha_1+\alpha_2+2\alpha_1\alpha_2}}\right).$$

But the harmonic mean of α_1 and α_2 is $\bar{\alpha} = 2\alpha_1\alpha_2/(\alpha_1+\alpha_2)$. Then $\mathbb{E}\|\tilde{\Pi} - \Pi \mathbb{1}_A\|_n^2 = O(n^{-\frac{2\bar{\alpha}}{2\bar{\alpha}+2}})$.

The condition $D_{m_1} \leq n^{1/3}$ allows this choice of m only if $\frac{\alpha_2}{\alpha_1+\alpha_2+2\alpha_1\alpha_2} < \frac{1}{3}$ i.e. if $\alpha_1 - 2\alpha_2 + 2\alpha_1\alpha_2 > 0$. In the same manner, the condition $\alpha_2 - 2\alpha_1 + 2\alpha_1\alpha_2 > 0$ must be verified. □

2.7.7 Proof of Theorem 2.3

We use the same notations as for the proof of Theorem 2.1. Let us write

$$\mathbb{E}\|\tilde{\Pi}^* - \Pi \mathbb{1}_A\|^2 = B_1 + B_2 + B_3$$

$$\text{with } \begin{cases} B_1 = \mathbb{E}\left(\|\tilde{\Pi}^* - \Pi \mathbb{1}_A\|^2 \mathbb{1}_{\Omega_\rho^*} \mathbb{1}_{\|\tilde{\Pi}\| \leq k_n}\right) \\ B_2 = \mathbb{E}\left(\|\tilde{\Pi}^* - \Pi \mathbb{1}_A\|^2 \mathbb{1}_{\Omega_\rho^*} \mathbb{1}_{\|\tilde{\Pi}\| > k_n}\right) \\ B_3 = \mathbb{E}\left(\|\tilde{\Pi}^* - \Pi \mathbb{1}_A\|^2 \mathbb{1}_{\Omega_\rho^{*c}}\right) \end{cases}$$

To bound the first term, we observe that for all $m \in \mathcal{M}_n$, on Ω_ρ^*, $\|\tilde{\Pi} - \Pi_m\|^2 \leq f_0^{-1}\rho\|\tilde{\Pi} - \Pi_m\|_n^2$. Then

$$\begin{aligned} \|\tilde{\Pi} - \Pi \mathbb{1}_A\|^2 \mathbb{1}_{\Omega_\rho^*} &\leq 2\|\tilde{\Pi} - \Pi_m\|^2 \mathbb{1}_{\Omega_\rho^*} + 2\|\Pi_m - \Pi \mathbb{1}_A\|^2 \\ &\leq 2f_0^{-1}\rho\|\tilde{\Pi} - \Pi_m\|_n^2 \mathbb{1}_{\Omega_\rho^*} + 2\|\Pi_m - \Pi \mathbb{1}_A\|^2 \\ &\leq 2f_0^{-1}\rho\{2\|\tilde{\Pi} - \Pi \mathbb{1}_A\|_n^2 \mathbb{1}_{\Omega_\rho^*} + 2\|\Pi_m - \Pi \mathbb{1}_A\|_n^2\} + 2\|\Pi_m - \Pi \mathbb{1}_A\|^2 \end{aligned}$$

Thus

$$B_1 \leq \mathbb{E}\left(\|\tilde{\Pi} - \Pi \mathbb{1}_A\|^2 \mathbb{1}_{\Omega_\rho^*}\right) \leq 4f_0^{-1}\rho \mathbb{E}\left(\|\tilde{\Pi} - \Pi \mathbb{1}_A\|_n^2 \mathbb{1}_{\Omega_\rho^*}\right) + (4f_0^{-1}\rho\|f\|_{\infty,A_1} + 2)\|\Pi_m - \Pi \mathbb{1}_A\|^2.$$

But, using (2.14), we obtain

$$B_1 \leq (24f_0^{-1}\rho\|f\|_{\infty,A_1} + 2)\|\Pi_m - \Pi \mathbb{1}_A\|^2 + 24f_0^{-1}\rho \mathrm{pen}(m) + 36f_0^{-1}\rho^2 \frac{C_1}{n}.$$

Since $\rho = 3/2$, by setting $C = \max(36f_0^{-1}\|f\|_{\infty,A_1} + 1, 36f_0^{-1})$,

$$B_1 \leq C\{\|\Pi_m - \Pi \mathbb{1}_A\|^2 + \mathrm{pen}(m)\} + \frac{81 f_0^{-1} C_1}{n}$$

for all $m \in \mathcal{M}_n$.

Next, the definition of $\tilde{\Pi}^*$ and the Markov inequality provide

$$B_2 \leq \mathbb{E}\left(\|\Pi \mathbb{1}_A\|^2 \mathbb{1}_{\Omega_\rho^*} \mathbb{1}_{\|\tilde{\Pi}\| > k_n}\right) \leq \|\Pi\|_A^2 \frac{\mathbb{E}(\|\tilde{\Pi}\|^2 \mathbb{1}_{\Omega_\rho^*})}{k_n^2}. \tag{2.21}$$

But $\|\tilde{\Pi}\|^2 \mathbb{1}_{\Omega_\rho^*} \leq \rho f_0^{-1}\|\tilde{\Pi}\|_n^2 \leq 2\rho f_0^{-1}(\|\tilde{\Pi} - \Pi \mathbb{1}_A\|_n^2 + \|\Pi \mathbb{1}_A\|_n^2)$. Now we use (2.15) to state

$$\begin{aligned}
\|\tilde{\Pi}\|^2 \mathbb{1}_{\Omega_\rho^*} &\leq 2\rho f_0^{-1}(\|\Pi\|_{\infty,A} + 4\phi_2 n^{1/3} + \|\Pi \mathbb{1}_A\|_n^2) \\
&\leq 2\rho f_0^{-1}(\|\Pi\|_{\infty,A} + 4\phi_2 n^{1/3} + \frac{1}{n}\sum_{i=1}^n \|\Pi\|_{\infty,A} \int \Pi(X_i, y) dy] \\
&\leq 2\rho f_0^{-1}(2\|\Pi\|_{\infty,A} + 4\phi_2 n^{1/3}).
\end{aligned}$$

Then, since $k_n = n^{2/3}$, (2.21) becomes

$$B_2 \leq \|\Pi\|_A^2 \frac{2\rho f_0^{-1}(2\|\Pi\|_{\infty,A} + 4\phi_2 n^{1/3})}{k_n^2} \leq 4\rho f_0^{-1}\|\Pi\|_A^2 \left(\frac{\|\Pi\|_{\infty,A}}{n^{4/3}} + \frac{2\phi_2}{n}\right).$$

Lastly

$$B_3 \leq \mathbb{E}\left(2(\|\tilde{\Pi}^*\|^2 + \|\Pi \mathbb{1}_A\|^2) \mathbb{1}_{\Omega_\rho^{*c}}\right) \leq 2(k_n^2 + \|\Pi\|_A^2) P(\Omega_\rho^{*c}).$$

We now remark that $P(\Omega_\rho^{*c}) = P(\Omega^{*c}) + P(\Omega_\rho^c \cap \Omega^*)$. In the geometric case $\beta_{2q_n} \leq e^{-\theta c \log(n)} \leq n^{-\theta c}$ and in the other case $\beta_{2q_n} \leq (2q_n)^{-\theta} \leq n^{-\theta c}$. Then

$$P(\Omega^{*c}) \leq 4p_n \beta_{2q_n} \leq n^{1-c\theta}.$$

But, if $\theta > 20$ in the arithmetic case, we can choose c such that $c\theta > \dfrac{10}{3}$ and so $P(\Omega^{*c}) \leq n^{-7/3}$. Then, using Proposition 2.3,

$$B_3 \leq 2(n^{4/3} + \|\Pi\|_A^2) \frac{1 + C_2}{n^{7/3}} \leq \frac{2(C_2 + 1)(1 + \|\Pi\|_A^2)}{n}.$$

□

2.8 Annex: Lower bound for the estimation of the transition density

For all function $g : \mathbb{R}^2 \mapsto \mathbb{R}$, we denote by

$$\|g\|_{p,A} = \left(\int_A |g|^p \right)^{1/p}$$

the norm in $L^p(A)$. We set

$$\mathcal{B} = \{\Pi \text{ transition density on } \mathbb{R} \text{ of a positive recurrent Markov chain such that } \|\Pi\|_{B^{\boldsymbol{\alpha}}_{r,q}(A)} \leq L\}$$

where $B^{\boldsymbol{\alpha}}_{r,q}$ is the anisotropic Besov space on A with regularity $r, q, \boldsymbol{\alpha}$. For all set C (included in $\mathbb{R}$ or $\mathbb{R}^2$), we denote by $|C|$ the Lebesgue measure of C. We recall that $\bar{\alpha}$ means the harmonic mean of α_1 and α_2:

$$\bar{\alpha} = \frac{2\alpha_1\alpha_2}{\alpha_1 + \alpha_2}.$$

Last $\mathbb{E}_\Pi$ means the expectation corresponding to the distribution of $X_1, \ldots, X_n$ if the true transition density of the Markov chain is Π and the initial distribution is the stationary distribution.

The following theorem is a generalization of the one of Clémençon (2000).

Theorem 2.5 *Let $1 \leq q \leq \infty$, $1 \leq p, r < \infty$. Then, there exists a positive constant C such that, if n is large enough,*

$$\inf_{\hat{\Pi}_n} \sup_{\Pi \in \mathcal{B}} (\mathbb{E}_\Pi \|\hat{\Pi}_n - \Pi\|^p_{p,A})^{1/p} \geq C n^{-\frac{\bar{\alpha}}{2\bar{\alpha}+2}}$$

where the infimum is taken over all estimators $\hat{\Pi}_n$ of Π based on the observations $X_1, \ldots, X_{n+1}$. If moreover $\bar{\alpha}(1 - 1/\alpha_1) > 2/r$ and $\mathcal{B}' = \mathcal{B} \cap \{\sum_{k\geq 0} \beta_k < B_1\}$

$$\inf_{\hat{\Pi}_n} \sup_{\Pi \in \mathcal{B}'} (\mathbb{E}_\Pi \|\hat{\Pi}_n - \Pi\|^p_{p,A})^{1/p} \geq C \left(\frac{\log n}{n} \right)^{\frac{\frac{\bar{\alpha}}{2} + \frac{1}{p} - \frac{1}{r}}{\bar{\alpha} + 1 - 2/r}}$$

We can reword the theorem :

Theorem 2.6 *Let $1 \leq q \leq \infty$, $1 \leq p < \infty$, $2/[\bar{\alpha}(1 - 1/\alpha_1)] < r < \infty$ and*

$$\mathcal{B}' = \{\Pi \text{ transition density on } \mathbb{R} \text{ of a positive recurrent Markov chain such that } \|\Pi\|_{B^{\boldsymbol{\alpha}}_{r,q}(A)} \leq L\} \cap \{\sum_{k\geq 0} \beta_k < B_1\}.$$

Then, for n large enough,

$$\inf_{\hat{\Pi}_n} \sup_{\Pi \in \mathcal{B}'} (\mathbb{E}_\Pi \|\hat{\Pi}_n - \Pi\|_{p,A}^p)^{1/p} \geq C r_n$$

where

$$r_n = \begin{cases} n^{-\frac{\bar{\alpha}}{2\bar{\alpha}+2}} & \text{if } r > \dfrac{p}{\bar{\alpha}+1}, \\ \left(\dfrac{\log n}{n}\right)^{\frac{\frac{\bar{\alpha}}{2}+\frac{1}{p}-\frac{1}{r}}{\bar{\alpha}+1-2/r}} & \text{if } r \leq \dfrac{p}{\bar{\alpha}+1}. \end{cases}$$

We have then a lower bound on the minimax risk over the anisotropic Besov classes. The obtained result is the same as the one of Clémençon (2000) by replacing the single regularity by the harmonic mean of the two regularities. As in the case of the estimation of a probability density, we observe a noteworthy phenomenon : an "elbow" for the rate of convergence. It divides the space of values (r, p) into two zones :

1. regular zone : $r > p/(\bar{\alpha}+1)$
2. sparse zone : $r \leq p/(\bar{\alpha}+1)$

An explanation concerning the names "regular" or "sparse" can be found in Härdle *et al.* (1998).

Proof of Theorem 2.5:

2.8.1 Regular case

We use the method described in detail in Härdle *et al.* (1998) and the proof of Theorem 54 in Clémençon (1999). Let ψ be a very regular wavelet with compact support. For $J = (j_1, j_2) \in \mathbb{Z}^2$ to be chosen below and $K = (k_1, k_2) \in \mathbb{Z}^2$, we set

$$\psi_{JK}(x, y) = 2^{(j_1+j_2)/2} \psi(2^{j_1}x - k_1)\psi(2^{j_2}y - k_2).$$

Let $\Pi_0(x, y) = c_0 \mathbb{1}_B(y)$ with B a compact set such that

- $A \subset B \times B$
- $|B| \geq 2|A|^{1/r}/L$

and $c_0 = |B|^{-1}$. So Π_0 is a transition density with $\|\Pi_0\|_{B^{\alpha}_{r,q}(A)} \leq L/2$, since $\|1\|_{B^{\alpha}_{r,q}(A)} = \|1\|_{r,A} = |A|^{1/r}$. Actually Π_0 is the transition of i.i.d. random variables with uniform distribution on B. Now we set R_J the maximal subset of $\mathbb{Z}^2$ such that

$$\begin{aligned} \text{Supp}(\psi_{JK}) \subset A \quad & \forall K \in R_J \\ \text{Supp}(\psi_{JK}) \cap \text{Supp}(\psi_{JK'}) = \emptyset \quad & \text{if } K \neq K' \end{aligned}$$

The cardinal of R_J is $|R_J| = c2^{j_1+j_2}$, with c a positive constant which depends only on A and the support of ψ. Let, for all $\varepsilon = (\varepsilon_K) \in \{-1,1\}^{|R_J|}$

$$\Pi_\varepsilon = \Pi_0 + \frac{1}{\sqrt{n}} \sum_{K \in R_J} \varepsilon_K \psi_{JK}.$$

Let us denote by $\mathcal{G}$ the set of all such Π_ε. We remark that, since the ψ_{JK} have disjoint supports, the sum in the previous equality is actually composed of a single term for all x, y. Since $\int \psi = 0$ and Π_0 is a transition density,

$$\forall x \in \mathbb{R} \int \Pi_\varepsilon(x,y)dy = 1.$$

Additionally $\Pi_\varepsilon(x,y) = \Pi_0(x,y) \geq 0$ if $(x,y) \notin A$, and if $(x,y) \in A$:

$$\Pi_\varepsilon \geq c_0 - \frac{2^{(j_1+j_2)/2}}{\sqrt{n}} \|\psi\|_\infty^2$$

and then $\Pi_\varepsilon(x,y) \geq c_0/2 > 0$ as soon as

$$\left(\frac{2^{j_1+j_2}}{n}\right)^{1/2} \leq \frac{c_0}{2\|\psi\|_\infty^2}. \tag{2.22}$$

Thus, if (2.22) holds, $\Pi_\varepsilon(x,y) \geq (c_0/2)\mathbb{1}_B(y)$ for all x, y. It implies that the underlying Markov chain is Doeblin recurrent and then positive recurrent. We observe that $f = c_0\mathbb{1}_B$ is the stationary density since

$$\int f(x)\Pi_\varepsilon(x,y)dx = \int f(x)\Pi_0(x,y)dx = f(y).$$

To prove that $\Pi_\varepsilon \in \mathcal{B}$, it remains to compute $\|\Pi_\varepsilon\|_{B^{\boldsymbol{\alpha}}_{r,q}(A)}$,

$$\|\Pi_\varepsilon\|_{B^{\boldsymbol{\alpha}}_{r,q}(A)} \leq \|\Pi_0\|_{B^{\boldsymbol{\alpha}}_{r,q}(A)} + \frac{1}{\sqrt{n}} \| \sum_{K \in R_J} \varepsilon_K \psi_{JK} \|_{B^{\boldsymbol{\alpha}}_{r,q}(A)}.$$

But Hochmuth (2002) proves that

$$\| \sum_{K \in R_J} \varepsilon_K \psi_{JK} \|_{B^{\boldsymbol{\alpha}}_{r,q}(A)} \leq (2^{j_1\alpha_1} + 2^{j_2\alpha_2}) \| \sum_{K \in R_J} \varepsilon_K \psi_{JK} \|_{r,A}$$

for $r < \infty$ and ψ smooth enough. Since

$$\| \textstyle\sum_{K \in R_J} \varepsilon_K \psi_{JK} \|_{r,A}^r = \sum_{K \in R_J} |\varepsilon_K|^r \int |\psi_{JK}|^r$$
$$= \textstyle\sum_{K \in R_J} |\varepsilon_K|^r 2^{(j_1+j_2)(r/2-1)} \|\psi\|_r^{2r} = c2^{j_1+j_2} 2^{(j_1+j_2)(r/2-1)} \|\psi\|_r^{2r},$$

then

$$\|\Pi_\varepsilon\|_{B^\alpha_{r,q}(A)} \leq \frac{L}{2} + \frac{2^{j_1\alpha_1} + 2^{j_2\alpha_2}}{\sqrt{n}} c^{1/r} \|\psi\|_r^2 2^{(j_1+j_2)/2}.$$

From now on, we suppose that Condition C is verified where

$$\text{Condition C:} \qquad \frac{(2^{j_1\alpha_1} + 2^{j_2\alpha_2})2^{(j_1+j_2)/2}}{\sqrt{n}} \leq \frac{L}{2c^{1/r}\|\psi\|_r^2}.$$

It entails in particular that (2.22) holds if j_1 and j_2 are great enough.Then for all ε, $\Pi_\varepsilon \in \mathcal{B}$.

Now we denote by $\Lambda_n(\Pi_{\varepsilon'}, \Pi_\varepsilon)$ the likelihood ratio between $\Pi_{\varepsilon'}$ and Π_ε. We use the following Lemma.

Lemma 2.4 Härdle *et al.* (1998) p.160
Let $\delta = \inf_{\varepsilon \neq \varepsilon'} \|\Pi_\varepsilon - \Pi_{\varepsilon'}\|_{p,A}/2$. *For* $\varepsilon \in \{-1, +1\}^{|R_J|}$, *put* $\varepsilon_{*K} = (\varepsilon'_I)_{I \in R_J}$ *such that:*

$$\varepsilon'_I = \begin{cases} -\varepsilon_I & \text{if } I = K, \\ \varepsilon_I & \text{else .} \end{cases}$$

If there exists $\lambda > 0$ *and* $p_0 > 0$ *such that*

$$\forall \varepsilon \quad \forall n \qquad P_{\Pi_\varepsilon}(\Lambda_n(\Pi_{\varepsilon_{*K}}, \Pi_\varepsilon) > e^{-\lambda}) \geq p_0$$

then, for any estimator $\hat{\Pi}_n$ *and for* n *large enough,*

$$\max_{\Pi_\varepsilon \in \mathcal{G}} \mathbb{E}_{\Pi_\varepsilon} \|\hat{\Pi}_n - \Pi_\varepsilon\|_{p,A}^p \geq \frac{|R_J|}{2} \delta^p e^{-\lambda} p_0.$$

So we now prove that there exists $\lambda > 0$ and p_0 such that

$$\forall \varepsilon, \forall n, \forall K \in R_J \qquad P_{\Pi_\varepsilon}(\Lambda_n(\Pi_{\varepsilon_{*K}}, \Pi_\varepsilon) > e^{-\lambda}) \geq p_0. \tag{2.23}$$

The likelihood ratio can be written

$$\Lambda_n(\Pi_{\varepsilon_{*K}}, \Pi_\varepsilon) = \prod_{i=1}^n \frac{\Pi_{\varepsilon_{*K}}(X_i, X_{i+1})}{\Pi_\varepsilon(X_i, X_{i+1})}.$$

Note that $\Pi_\varepsilon(X_i, X_{i+1}) > 0$ P_{Π_ε}- and $P_{\Pi_{\varepsilon_{*K}}}$- almost surely (actually the chain "lives" on B). Then

$$\log(\Lambda_n(\Pi_{\varepsilon_{*K}}, \Pi_\varepsilon)) = \sum_{i=1}^n \log\left(1 - \frac{2}{\sqrt{n}} \frac{\varepsilon_K \psi_{JK}(X_i, X_{i+1})}{\Pi_\varepsilon(X_i, X_{i+1})}\right).$$

We set $U_{JK}(X_i, X_{i+1}) = \dfrac{-\varepsilon_K \psi_{JK}(X_i, X_{i+1})}{\Pi_\varepsilon(X_i, X_{i+1})}$ so that

$$\begin{aligned}
&\log(\Lambda_n(\Pi_{\varepsilon_{*K}}, \Pi_\varepsilon)) = \sum_{i=1}^n \log\left(1 + \frac{2}{\sqrt{n}} U_{JK}(X_i, X_{i+1})\right) \\
&= \sum_{i=1}^n \left\{ \theta\left(\frac{2}{\sqrt{n}} U_{JK}(X_i, X_{i+1})\right) + \frac{2}{\sqrt{n}} U_{JK}(X_i, X_{i+1}) - \frac{2}{n} U_{JK}^2(X_i, X_{i+1}) \right\} \\
&= u_n + v_n - w_n
\end{aligned}$$

with θ the function defined by $\theta(u) = \log(1+u) - u + \dfrac{u^2}{2}$. Now we will prove the three following assertions

1° $\mathbb{E}_{\Pi_\varepsilon}(|u_n|) = \mathbb{E}_{\Pi_\varepsilon}\left(\left|\sum_{i=1}^n \theta\left(\frac{2}{\sqrt{n}} U_{JK}(X_i, X_{i+1})\right)\right|\right) \underset{n\to\infty}{\to} 0$

2° $\mathbb{E}_{\Pi_\varepsilon}(w_n) = \mathbb{E}_{\Pi_\varepsilon}\left(\dfrac{2}{n}\sum_{i=1}^n U_{JK}^2(X_i, X_{i+1})\right) \leq 4$

3° $\mathbb{E}_{\Pi_\varepsilon}(v_n^2) = \mathbb{E}_{\Pi_\varepsilon}\left(\dfrac{4}{n}|\sum_{i=1}^n U_{JK}(X_i, X_{i+1})|^2\right) \leq 8$

1° : First we observe that $\|\dfrac{2}{\sqrt{n}} U_{JK}\|_\infty \leq \dfrac{2}{\sqrt{n}} \dfrac{2^{(j_1+j_2)/2}\|\psi\|_\infty^2}{c_0/2} = O\left(\dfrac{2^{(j_1+j_2)/2}}{\sqrt{n}}\right)$ and $\dfrac{2^{(j_1+j_2)}}{n} \to 0$ since Condition C holds. So there exists some integer n_0 such that $\forall n \geq n_0$

$$\forall x, y \quad \left|\theta\left(\frac{2}{\sqrt{n}} U_{JK}(x,y)\right)\right| \leq \left|\frac{2}{\sqrt{n}} U_{JK}(x,y)\right|^3.$$

But

$$\begin{aligned}
&\iint \left|\frac{2U_{JK}(x,y)}{\sqrt{n}}\right|^3 f(x)\Pi_\varepsilon(x,y)dxdy = \frac{8}{n\sqrt{n}} \iint \frac{|\psi_{JK}(x,y)|^3}{\Pi_\varepsilon(x,y)^2} f(x)dxdy \\
&\leq \frac{8}{n\sqrt{n}} \frac{2^{(j_1+j_2)/2}\|\psi\|_\infty^2 c_0}{(c_0/2)^2} \iint \psi_{JK}(x,y)^2 dxdy \leq \frac{32\|\psi\|_\infty^2}{c_0 n} \left(\frac{2^{(j_1+j_2)}}{n}\right)^{1/2}.
\end{aligned}$$

Then

$$\begin{aligned}
&\mathbb{E}_{\Pi_\varepsilon}|u_n| \leq \sum_{i=1}^n \iint \left|\theta\left(\frac{2}{\sqrt{n}} U_{JK}(x,y)\right) f(x)\Pi_\varepsilon(x,y)\right| dxdy \\
&\leq \sum_{i=1}^n \frac{32\|\psi\|_\infty^2}{c_0 n} \left(\frac{2^{(j_1+j_2)}}{n}\right)^{1/2} \underset{n\to\infty}{\longrightarrow} 0
\end{aligned}$$

2° : We bound the expectation of $U_{JK}(X_i, X_{i+1})^2$

$$\mathbb{E}_{\Pi_\varepsilon}(U_{JK}(X_i,X_{i+1})^2) = \mathbb{E}_{\Pi_\varepsilon}\left(\frac{\psi_{JK}^2(X_i,X_{i+1})}{\Pi_\varepsilon^2(X_i,X_{i+1})}\right) = \iint \frac{\psi_{JK}^2(x,y)}{\Pi_\varepsilon(x,y)} f(x)dxdy$$

where $f = c_0 \mathbb{1}_B$ is the stationary density. So

$$\mathbb{E}_{\Pi_\varepsilon}(U_{JK}(X_i,X_{i+1})^2) \leq c_0 \iint_A \frac{\psi_{JK}^2(x,y)}{c_0/2} dxdy \leq 2. \tag{2.24}$$

And then $\mathbb{E}_{\Pi_\varepsilon}(w_n) = \mathbb{E}_{\Pi_\varepsilon}\left(\frac{2}{n}\sum_{i=1}^n U_{JK}(X_i,X_{i+1})^2\right) \leq 4$.

3° : We observe that

$$\begin{aligned}\mathbb{E}_{\Pi_\varepsilon}(U_{JK}(X_i,X_{i+1})|X_1,\ldots,X_i) &= \int U_{JK}(X_i,y)\Pi_\varepsilon(X_i,y)dy \\ &= -\varepsilon_K \int \psi_{JK}(X_i,y)dy = 0\end{aligned}$$

and thus $\sum_{i=1}^n U_{JK}(X_i,X_{i+1})$ is a martingale. A classic property of square integrable martingales involves

$$E_{\Pi_\varepsilon}\left(\left[\sum_{i=1}^n U_{JK}(X_i,X_{i+1})\right]^2\right) = \sum_{i=1}^n \mathbb{E}_{\Pi_\varepsilon}\left[U_{JK}(X_i,X_{i+1})^2\right].$$

Thus, using (2.24)

$$\mathbb{E}_{\Pi_\varepsilon}(v_N^2) = \frac{4}{n}\sum_{i=1}^n \mathbb{E}_{\Pi_\varepsilon}\left[U_{JK}(X_i,X_{i+1})^2\right] \leq 8.$$

Now we deduce from the three previous assertions 1°, 2° and 3° that there exists $\lambda > 0$ and p_0 such that $P_{\Pi_\varepsilon}(\Lambda_n(\Pi_{\varepsilon_{*K}}, \Pi_\varepsilon) > e^{-\lambda}) \geq p_0$. Since $\mathbb{E}_{\Pi_\varepsilon}(|u_n|) \to 0$, there exists a positive constant M such that for all $n \geq 1$, $\mathbb{E}_{\Pi_\varepsilon}(|u_n|) \leq M$. Now

$$\begin{aligned}P_{\Pi_\varepsilon}(u_n + v_n - w_n > -\lambda) &\geq 1 - P_{\Pi_\varepsilon}(-u_n + w_n \geq \frac{\lambda}{2}) - P_{\Pi_\varepsilon}(-v_n \geq \frac{\lambda}{2}) \\ &\geq 1 - \frac{\mathbb{E}_{\Pi_\varepsilon}|-u_n+w_n|}{(\lambda/2)} - \frac{\mathbb{E}_{\Pi_\varepsilon}|v_n|^2}{(\lambda/2)^2} \geq 1 - 2\frac{M+4}{\lambda} - 4\frac{8}{\lambda^2} \geq \frac{1}{2}\end{aligned}$$

for λ large enough. Thus we have proved equality (2.23) and according to Lemma 2.4,

$$\max_{\Pi_\varepsilon \in \mathcal{G}} \mathbb{E}_{\Pi_\varepsilon} \|\hat{\Pi}_n - \Pi_\varepsilon\|_{p,A}^p \geq \frac{|R_J|}{2}\delta^p e^{-\lambda} p_0$$

where

$$\delta = \inf_{\varepsilon \neq \varepsilon'} \|\Pi_\varepsilon - \Pi_{\varepsilon'}\|_{p,A}/2 = \left\| \frac{1}{\sqrt{n}} \varepsilon_K \psi_{JK} \right\|_{p,A} = \frac{2^{(j_1+j_2)(\frac{1}{2}-\frac{1}{p})} \|\psi\|_p^2}{\sqrt{n}}.$$

Finally

$$\max_{\Pi_\varepsilon \in \mathcal{G}} \mathbb{E}_{\Pi_\varepsilon} \|\hat{\Pi}_n - \Pi_\varepsilon\|_{p,A}^p \geq \frac{ce^{-\lambda} p_0 \|\psi\|_p^{2p}}{2} \frac{2^{(j_1+j_2)(1+p(\frac{1}{2}-\frac{1}{p}))}}{n^{p/2}} = \frac{ce^{-\lambda} p_0 \|\psi\|_p^{2p}}{2} \left(\frac{2^{j_1+j_2}}{n} \right)^{\frac{p}{2}}.$$

Now for all n we choose $J = J(n) = (j_1(n), j_2(n))$ such that

$$\begin{cases} \dfrac{c_1}{2} n^{\frac{\alpha_2}{\alpha_1+\alpha_2+2\alpha_1\alpha_2}} \leq 2^{j_1} \leq c_1 n^{\frac{\alpha_2}{\alpha_1+\alpha_2+2\alpha_1\alpha_2}} \\ \dfrac{c_2}{2} n^{\frac{\alpha_1}{\alpha_1+\alpha_2+2\alpha_1\alpha_2}} \leq 2^{j_2} \leq c_2 n^{\frac{\alpha_1}{\alpha_1+\alpha_2+2\alpha_1\alpha_2}} \end{cases}$$

with c_1 and c_2 such that

$$\frac{(2^{j_1\alpha_1} + 2^{j_2\alpha_2}) 2^{(j_1+j_2)/2}}{\sqrt{n}} \leq (c_1^{\alpha_1} + c_2^{\alpha_2}) \sqrt{c_1 c_2} \leq \frac{L}{2c^{1/r} \|\psi\|_r^2}$$

so that Condition C is satisfied. Moreover, we have

$$\frac{2^{j_1+j_2}}{n} \geq \frac{c_1 c_2}{4} n^{\frac{\alpha_2+\alpha_1}{\alpha_1+\alpha_2+2\alpha_1\alpha_2}-1} \geq \frac{c_1 c_2}{4} n^{\frac{-2\alpha_1\alpha_2}{\alpha_1+\alpha_2+2\alpha_1\alpha_2}}.$$

Thus

$$\max_{\Pi_\varepsilon \in \mathcal{G}} \mathbb{E}_{\Pi_\varepsilon} \|\hat{\Pi}_n - \Pi_\varepsilon\|_{p,A}^p \geq \frac{ce^{-\lambda} p_0 \|\psi\|_p^{2p} (c_1 c_2)^{p/2}}{2^{p+1}} n^{\frac{-p\alpha_1\alpha_2}{\alpha_1+\alpha_2+2\alpha_1\alpha_2}} \geq C n^{\frac{-p\bar{\alpha}}{2\bar{\alpha}+2}}.$$

And then for all estimator

$$\sup_{\Pi \in \mathcal{B}} \mathbb{E}_{\Pi} \|\hat{\Pi}_n - \Pi\|_{p,A}^p \geq C n^{-\frac{p\bar{\alpha}}{2\bar{\alpha}+2}}$$

with $C = ce^{-\lambda} p_0 \|\psi\|_p^{2p} (c_1 c_2)^{p/2} / 2^{p+1}$.

2.8.2 Sparse case

In this case, let d a positive real small enough and $\mathcal{G}$ the set of functions

$$\Pi_K(x,y) = \Pi_0(x,y) + d\sqrt{\frac{\log n}{n}} \psi_{JK}(x,y)$$

for K in R_J. Let us prove that $\mathcal{G} \subset \mathcal{B}$.

First $\int \Pi_K(.,y)dy = 1$. Now if $(x,y) \in A$,

$$\Pi_K(x,y) \geq c_0 - d\sqrt{\frac{\log n}{n}} 2^{(j_1+j_2)/2} \|\psi\|_\infty^2.$$

And then $\Pi_K \geq (c_0/2)\mathbb{1}_A$ as soon as

$$\text{Condition D:} \qquad 2^{(j_1+j_2)/2}\sqrt{\frac{\log n}{n}} \leq \frac{c_0}{2\|\psi\|_\infty^2 d}.$$

If the Condition D is verified, the underlying Markov chain is positive recurrent with stationary density $f = c_0\mathbb{1}_B$.

Let us now prove that $\|\Pi_K\|_{B^{\boldsymbol{\alpha}}_{r,q}(A)} \leq L$.

$$\|\Pi_K\|_{B^{\boldsymbol{\alpha}}_{r,q}(A)} \leq \|\Pi_0\|_{B^{\boldsymbol{\alpha}}_{r,q}(A)} + d\sqrt{\frac{\log n}{n}}\|\psi_{JK}\|_{B^{\boldsymbol{\alpha}}_{r,q}(A)} \leq \frac{L}{2} + d\sqrt{\frac{\log n}{n}}(2^{j_1\alpha_1} + 2^{j_2\alpha_2})\|\psi_{JK}\|_r$$

using Hochmuth (2002). But $\|\psi_{JK}\|_r^r = 2^{(j_1+j_2)(r/2-1)}\|\psi\|_r^{2r}$. Then

$$\|\Pi_K\|_{B^{\boldsymbol{\alpha}}_{r,q}(A)} \leq \frac{L}{2} + d\sqrt{\frac{\log n}{n}}(2^{j_1\alpha_1} + 2^{j_2\alpha_2})2^{(j_1+j_2)(\frac{1}{2}-\frac{1}{r})}\|\psi\|_r^2.$$

From now on, we suppose that Condition E is verified where

$$\text{Condition E:} \qquad (2^{j_1\alpha_1} + 2^{j_2\alpha_2})2^{(j_1+j_2)(\frac{1}{2}-\frac{1}{r})}\sqrt{\frac{\log n}{n}} \leq \frac{L}{2d\|\psi\|_r^2}$$

Then, if conditions D and E hold, $\Pi_K \in \mathcal{B}$ for all $K \in R_J$. Now choose

$$\begin{cases} \dfrac{c_1}{2}\left(\sqrt{\dfrac{n}{\log n}}\right)^{\frac{\bar{\alpha}/\alpha_1}{\bar{\alpha}+1-2/r}} \leq 2^{j_1} \leq c_1\left(\sqrt{\dfrac{n}{\log n}}\right)^{\frac{\bar{\alpha}/\alpha_1}{\bar{\alpha}+1-2/r}} \\ \dfrac{c_2}{2}\left(\sqrt{\dfrac{n}{\log n}}\right)^{\frac{\bar{\alpha}/\alpha_2}{\bar{\alpha}+1-2/r}} \leq 2^{j_2} \leq c_2\left(\sqrt{\dfrac{n}{\log n}}\right)^{\frac{\bar{\alpha}/\alpha_2}{\bar{\alpha}+1-2/r}} \end{cases}$$

We compute

$$2^{(j_1+j_2)/2}\sqrt{\frac{\log n}{n}} \leq \sqrt{c_1c_2}\left(\frac{n}{\log n}\right)^{\frac{2/r-\bar{\alpha}}{2(\bar{\alpha}+1-2/r)}}$$

and

$$(2^{j_1\alpha_1} + 2^{j_2\alpha_2})2^{(j_1+j_2)(\frac{1}{2}-\frac{1}{r})}\sqrt{\frac{\log n}{n}} \leq \max(c_1^{\alpha_1}, c_2^{\alpha_2})(c_1c_2)^{(\frac{1}{2}-\frac{1}{r})}.$$

Since $2/r - \bar{\alpha} < 0$, the condition D holds for n sufficiently large and the conditions E is verified if d is small enough.

We use now the following Lemma.

Lemma 2.5 Korostelëv and Tsybakov (1993)
Let $\mathcal{B}$ contain the functions $(\Pi_K)_{K\in R_J}$ such that

1° : $\|\Pi_K - \Pi_{K'}\|_{p,A} \geq \delta > 0$, *for* $K \in R_J, K \neq K'$,

2° : $|R_J| \geq \exp(\lambda_n)$, *for some* $\lambda_n > 0$,

3° : $\Lambda_n(\Pi_0, \Pi_K) = \exp(z_n^K - v_n^K)$ *where* z_n^K *is a random variable such that there exists* p_0 *with* $P_{\Pi_K}(z_n^K > 0) \geq p_0$, *and* v_n^K *are constants,*

4° : $\sup_K v_n^K \leq \lambda_n$.

Then

$$\sup_{\Pi\in\mathcal{B}} P_\Pi\left(\|\Pi - \hat{\Pi}_n\|_{p,A} \geq \frac{\delta}{2}\right) \geq \sup_{K\in R_J} P_\Pi\left(\|\Pi_K - \hat{\Pi}_n\|_{p,A} \geq \frac{\delta}{2}\right) \geq \frac{p_0}{2},$$

for an arbitrary estimator $\hat{\Pi}_n$.

Let us verify the four points of this lemma

1° : $\|\Pi_K - \Pi_{K'}\|_p = d\sqrt{\frac{\log n}{n}}\|\psi_{JK} - \psi_{JK'}\|_p$. But ψ_{JK} and $\psi_{JK'}$ have disjoint supports. So

$$\|\psi_{JK} - \psi_{JK'}\|_p^p = \|\psi_{JK}\|_p^p + \|\psi_{JK'}\|_p^p = 2.2^{(j_1+j_2)(p/2-1)}\|\psi\|_p^{2p}.$$

And

$$\|\Pi_K - \Pi_{K'}\|_p = d\sqrt{\frac{\log n}{n}}2^{1/p}2^{(j_1+j_2)(\frac{1}{2}-\frac{1}{p})}\|\psi\|_p^2 =: \delta \tag{2.25}$$

2° : We have already observed that $|R_J| = c2^{j_1+j_2}$. So

$$|R_J| \geq c\frac{c_1c_2}{4}\left(\sqrt{\frac{n}{\log n}}\right)^{\frac{\bar{\alpha}/\alpha_1+\bar{\alpha}/\alpha_2}{\bar{\alpha}+1-2/r}} = c\frac{c_1c_2}{4}\left(\frac{n}{\log n}\right)^{\frac{1}{\bar{\alpha}+1-2/r}}$$

and

$$\log|R_J| \geq \log(c\frac{c_1c_2}{4}) + \frac{1}{\bar{\alpha}+1-2/r}[\log n - \log\log n] \geq \lambda_n$$

for n large enough and with

$$\lambda_n = \frac{\log n}{2(\bar{\alpha}+1-2/r)} \tag{2.26}$$

3° : The likelihood ratio can be written

$$\Lambda_n(\Pi_0, \Pi_K) = \prod_{i=1}^n \frac{\Pi_0(X_i, X_{i+1})}{\Pi_K(X_i, X_{i+1})}$$

and then

$$\begin{aligned}\log(\Lambda_n(\Pi_0,\Pi_K)) &= \sum_{i=1}^{n}\log\left(1-d\sqrt{\frac{\log n}{n}}\frac{\psi_{JK}(X_i,X_{i+1})}{\Pi_K(X_i,X_{i+1})}\right)\\ &= \sum_{i=1}^{n}\log\left(1+d\sqrt{\frac{\log n}{n}}V_{JK}(X_i,X_{i+1})\right)\end{aligned}$$

by setting $V_{JK}(X_i,X_{i+1})=-\dfrac{\psi_{JK}(X_i,X_{i+1})}{\Pi_K(X_i,X_{i+1})}$. Thus, if $\theta(u)=\log(1+u)-u+\dfrac{u^2}{2}$,

$$\begin{aligned}\log(\Lambda_n(\Pi_0,\Pi_K) &= \sum_{i=1}^{n}\left\{\theta\left(d\sqrt{\frac{\log n}{n}}V_{JK}(X_i,X_{i+1})\right)+d\sqrt{\frac{\log n}{n}}V_{JK}(X_i,X_{i+1})\right.\\ &\left.-\frac{d^2}{2}\frac{\log n}{n}V_{JK}^2(X_i,X_{i+1})\right\}=z_n^K-v_n^K\end{aligned}$$

with

$$\begin{aligned}z_n^K &= \sum_{i=1}^{n}\theta\left(d\sqrt{\frac{\log n}{n}}V_{JK}(X_i,X_{i+1})\right)+d\sqrt{\frac{\log n}{n}}\sum_{i=1}^{n}V_{JK}(X_i,X_{i+1})\\ &\quad-\frac{d^2}{2}\frac{\log n}{n}\sum_{i=1}^{n}[V_{JK}^2(X_i,X_{i+1})-\mathbb{E}_{\Pi_K}(V_{JK}^2(X_i,X_{i+1}))]\\ v_n^K &= \frac{d^2}{2}\frac{\log n}{n}\sum_{i=1}^{n}\mathbb{E}_{\Pi_K}(V_{JK}^2(X_i,X_{i+1})).\end{aligned}$$

We have to prove that there exists p_0 with $P_{\Pi_K}(z_n^K>0)\geq p_0$. We split z_n^K into four terms: $z_n^K=z_{n,1}+z_{n,2}+z_{n,3}+z_{n,4}$ with

$$\begin{cases}z_{n,1}=\displaystyle\sum_{i=1}^{n}\theta\left(d\sqrt{\frac{\log n}{n}}V_{JK}(X_i,X_{i+1})\right)\\ z_{n,2}=-\dfrac{d^2}{2}\dfrac{\log n}{n}\displaystyle\sum_{i=1}^{n}[V_{JK}^2(X_i,X_{i+1})-\mathbb{E}_{\Pi_K}(V_{JK}^2(X_i,X_{i+1})|X_i)]\\ z_{n,3}=-\dfrac{d^2}{2}\dfrac{\log n}{n}\displaystyle\sum_{i=1}^{n}[\mathbb{E}_{\Pi_K}(V_{JK}^2(X_i,X_{i+1})|X_i)-\mathbb{E}_{\Pi_K}(V_{JK}^2(X_i,X_{i+1}))]\\ z_{n,4}=d\sqrt{\dfrac{\log n}{n}}\displaystyle\sum_{i=1}^{n}V_{JK}(X_i,X_{i+1}).\end{cases}$$

We show that $\mathbb{E}_{\Pi_K}\left|\dfrac{z_{n,j}}{\sqrt{\log n}}\right|\to 0$ for $j=1,2,3$ and $\dfrac{z_{n,4}}{d\sqrt{\log n}}\overset{\mathcal{L}(P_{\Pi_K})}{\longrightarrow}\mathcal{N}(0,1)$.

• First

$$\left| d\sqrt{\frac{\log n}{n}} V_{JK}(X_i, X_{i+1}) \right| \leq d\sqrt{\frac{\log n}{n}} \frac{2^{(j_1+j_2)/2}\|\psi\|_\infty^2}{c_0/2} \underset{n\to\infty}{\longrightarrow} 0.$$

Then, for n large enough, $|\theta(d\sqrt{\frac{\log n}{n}} V_{JK}(X_i, X_{i+1}))| \leq |d\sqrt{\frac{\log n}{n}} V_{JK}(X_i, X_{i+1})|^3$. But

$$\begin{aligned}
&\iint \left| d\sqrt{\frac{\log n}{n}} V_{JK}(x,y) \right|^3 f(x)\Pi_\varepsilon(x,y)dxdy \\
&= d^3 \left(\frac{\log n}{n}\right)^{3/2} \iint \frac{|\psi_{JK}(x,y)|^3}{\Pi_\varepsilon(x,y)^2} f(x)dxdy \\
&\leq d^3 \left(\frac{\log n}{n}\right)^{3/2} \frac{c_0\|\psi_{JK}\|_\infty}{(c_0/2)^2} \iint \psi_{JK}(x,y)^2 dxdy \leq M 2^{(j_1+j_2)/2} \left(\frac{\log n}{n}\right)^{3/2}
\end{aligned}$$

with $M = d^3\|\psi\|_\infty/(4c_0)$. Thus $\mathbb{E}_{\Pi_K}|z_{n,1}| \leq M 2^{(j_1+j_2)/2} \frac{(\log n)^{3/2}}{\sqrt{n}}$ and

$$\frac{\mathbb{E}_{\Pi_K}|z_{n,1}|}{\sqrt{\log n}} \leq M\sqrt{\log n}\sqrt{c_1 c_2} \left(\frac{n}{\log n}\right)^{\frac{2/r-\bar{\alpha}}{2(\bar{\alpha}+1-2/r)}} \underset{n\to\infty}{\longrightarrow} 0.$$

• Next we remark that $\mathbb{E}_{\Pi_K}[V_{JK}^2(X_i, X_{i+1}) - \mathbb{E}_{\Pi_K}(V_{JK}^2(X_i, X_{i+1})|X_i)|X_i] = 0$. This implies (cf. remark p.97)

$$\mathbb{E}_{\Pi_K}|z_{n,2}|^2 = \frac{d^4}{4}\frac{(\log n)^2}{n^2} \sum_{i=1}^n \mathbb{E}_{\Pi_K}[(V_{JK}^2(X_i, X_{i+1}) - \mathbb{E}_{\Pi_K}(V_{JK}^2(X_i, X_{i+1})|X_i))^2].$$

Therefore

$$\begin{aligned}
\mathbb{E}_{\Pi_K}|z_{n,2}|^2 &\leq \frac{d^4}{4}\frac{(\log n)^2}{n^2} \sum_{i=1}^n \mathbb{E}_{\Pi_K}[V_{JK}^4(X_i, X_{i+1})] \\
&\leq \frac{d^4}{4}\frac{(\log n)^2}{n^2} \sum_{i=1}^n \iint \frac{\psi_{JK}^4(x,y)}{\Pi_K^3(x,y)} f(x)dxdy \leq \frac{2d^4}{c_0^2}\frac{(\log n)^2}{n}\|\psi_{JK}^2\|_\infty \iint \psi_{JK}^2(x,y)dxdy.
\end{aligned}$$

Consequently

$$\mathbb{E}_{\Pi_K}|z_{n,2}|^2 \leq \frac{2d^4}{c_0^2}\frac{(\log n)^2}{n} 2^{j_1+j_2}\|\psi\|_\infty^4$$

and

$$\mathbb{E}_{\Pi_K}|\frac{z_{n,2}}{\sqrt{\log n}}| \leq \frac{\sqrt{2}d^2\|\psi\|_\infty^2}{c_0}\sqrt{\frac{\log n}{n} 2^{j_1+j_2}} \to 0.$$

- If Φ denotes the function

$$\Phi(x) = \int \frac{\psi_{JK}^2(x,y)}{\Pi_K(x,y)} dy, \tag{2.27}$$

then $\mathbb{E}_{\Pi_K}(V_{JK}^2(X_i, X_{i+1})|X_i) = \Phi(X_i)$, so that

$$\mathbb{E}_{\Pi_K}|z_{n,3}|^2 \leq \frac{d^4}{4}\frac{(\log n)^2}{n^2}\mathrm{Var}_{\Pi_K}[\sum_{i=1}^n \Phi(X_i)].$$

So we use the following lemma.

Lemma 2.6 *(Volkonskiĭ and Rozanov (1959)) Let (T_i) a strictly stationnary process with β-mixing coefficients β_k. Then, for all function ψ (such that $\mathbb{E}[\psi^2(T_1)] < \infty$) and for all n*

$$\mathrm{Var}(\sum_{i=1}^n \psi(T_i)) \leq 4n \sum_k \beta_k \|\psi\|_\infty^2.$$

Thus

$$\mathbb{E}_{\Pi_K}|z_{n,3}|^2 \leq d^4\frac{(\log n)^2}{n}\sum_k \beta_k \|\Phi\|_\infty^2 \leq d^4\frac{(\log n)^2}{n}\sum_k \beta_k \left(\frac{2^{j_1}\|\psi\|_\infty^2}{c_0/2}\right)^2.$$

Hence

$$\mathbb{E}_{\Pi_K}|\frac{z_{n,3}}{\sqrt{\log n}}| \leq \frac{2d^2\|\psi\|_\infty^2}{c_0}\sqrt{\sum_k \beta_k}\sqrt{\frac{\log n}{n}}2^{j_1} \to 0$$

since $\bar{\alpha}(1 - 1/\alpha_1) > 2/r$.

- Now we use the following central limit theorem to prove that

$$\frac{z_{n,4}}{d\sqrt{\log n}} = \frac{1}{\sqrt{n}}\sum_{i=1}^n V_{JK}(X_i, X_{i+1}) \overset{\mathcal{L}(P_{\Pi_K})}{\longrightarrow} \mathcal{N}(0,1).$$

Lemma 2.7 *(Theorem 2.8.41 in Dacunha-Castelle and Duflo (1983))*
Let M be a square integrable martingale adapted to a filtration $(\mathcal{F}_k)$ with square bracket process $< M >$. We assume that, for a deterministic positive sequence (a_n) growing to infinity,

K1) $a_n^{-1} < M >_n$ tends to Γ in probability;

K2) The Lindeberg condition is verified, i.e. for all $\varepsilon > 0$,

$$a_n^{-1}\sum_{k=1}^n \mathbb{E}[|M_k - M_{k-1}|^2 \mathbb{1}_{|M_k - M_{k-1}| \geq \varepsilon a_n^{1/2}}|\mathcal{F}_{k-1}]$$

tends to 0 in probability.

Then $a_n^{-1} M_n \xrightarrow{a.s.} 0$ *and* $a_n^{-1/2} M_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Gamma)$.

Let $\mathcal{F}_k$ be the σ-algebra generated by the variables $X_1, \ldots, X_{k+1}$ and

$$M_n = \sum_{i=1}^{n} V_{JK}(X_i, X_{i+1}).$$

Then the process (M_n) is a martingale adapted to $(\mathcal{F}_n)$. By setting $a_n = n$, Assumption K2) in Lemma 2.7 can be written

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\Pi_K}[|V_{JK}(X_i, X_{i+1})|^2 \mathbb{1}_{|V_{JK}(X_i, X_{i+1})| \geq \varepsilon n^{1/2}} | X_1, \ldots, X_i] \xrightarrow{P_{\Pi_K}} 0.$$

But, for $\delta > 0$,

$$\begin{aligned}
&P_{\Pi_K}\left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\Pi_K}[|V_{JK}(X_i, X_{i+1})|^2 \mathbb{1}_{|V_{JK}(X_i, X_{i+1})| \geq \varepsilon n^{1/2}} | X_1, \ldots, X_i] > \delta \right) \\
&\leq \frac{1}{\delta n} \sum_{i=1}^{n} \mathbb{E}_{\Pi_K}\left[|V_{JK}(X_i, X_{i+1})|^2 \mathbb{1}_{|V_{JK}(X_i, X_{i+1})| \geq \varepsilon n^{1/2}} \right] \\
&\leq \frac{1}{\delta n} \sum_{i=1}^{n} \mathbb{E}_{\Pi_K}\left[|V_{JK}(X_i, X_{i+1})|^2 \left| \frac{V_{JK}(X_i, X_{i+1})}{\varepsilon n^{1/2}} \right|^2 \right] \\
&\leq \frac{1}{\delta \varepsilon^2} \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}_{\Pi_K}[|V_{JK}(X_i, X_{i+1})|^4]
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}_{\Pi_K} |V_{JK}(X_i, X_{i+1})|^4 = \frac{1}{n^2} \sum_{i=1}^{n} \iint |V_{JK}(x,y)|^4 \Pi_K(x,y) f(x) dx dy \\
&= \frac{1}{n} \iint \frac{|\psi_{JK}(x,y)|^4}{\Pi_K(x,y)^3} f(x) dx dy \leq \frac{1}{n} \frac{\|f\|_\infty \|\psi_{JK}\|_\infty^2}{(c_0/2)^3} \iint \psi_{JK}(x,y)^2 dx dy \\
&\leq \frac{8 \|\psi\|_\infty^4}{c_0^2} \left(\frac{2^{j_1 + j_2}}{n} \right) \underset{n \to \infty}{\longrightarrow} 0.
\end{aligned}$$

Therefore the Lindeberg condition is verified. Now we write

$$< M >_n = \sum_{i=1}^{n} \mathbb{E}_{\Pi_K}[V_{JK}(X_i, X_{i+1})^2 | X_1, \ldots, X_i].$$

So, to prove that Assumption K1) in Lemma 2.7 holds, we have to show that $(1/n)\sum_{i=1}^n \mathbb{E}_{\Pi_K}(V_{JK}^2(X_i,X_{i+1})|X_i) \stackrel{P_{\Pi_K}}{\longrightarrow} 1$, i.e. $(1/n)\sum_{i=1}^n \Phi(X_i) \to 1$ where Φ is defined in (2.27). We use that

$$\mathbb{E}_{\Pi_K}\left|\frac{1}{n}\sum_{i=1}^n \Phi(X_i) - 1\right|^2 = \mathrm{Var}_{\Pi_K}\left(\frac{1}{n}\sum_{i=1}^n \Phi(X_i)\right) + \left|\mathbb{E}_{\Pi_K}\left(\frac{1}{n}\sum_{i=1}^n \Phi(X_i)\right) - 1\right|^2.$$

But, using again Lemma 2.6,

$$\mathrm{Var}_{\Pi_K}\left[\frac{1}{n}\sum_{i=1}^n \Phi(X_i)\right] \le \frac{4}{n}\sum_k \beta_k \|\Phi^2\|_\infty \le \frac{16\|\psi\|_\infty^4}{c_0^2}\sum_k \beta_k \frac{2^{2j_1}}{n} = o\left(\frac{1}{\log n}\right).$$

Now we remark that $\iint \dfrac{\psi_{JK}^2(x,y)}{\Pi_0(x,y)} f(x)dxdy = 1$, and so

$$\begin{aligned}
\left|\mathbb{E}_{\Pi_K}\left(\frac{1}{n}\sum_{i=1}^n \Phi(X_i)\right) - 1\right| &= \left|\iint \frac{\psi_{JK}^2(x,y)}{\Pi_K(x,y)} f(x)dxdy - \iint \frac{\psi_{JK}^2(x,y)}{\Pi_0(x,y)} f(x)dxdy\right| \\
&\le \iint \psi_{JK}^2(x,y)\left|\frac{1}{\Pi_K(x,y)} - \frac{1}{\Pi_0(x,y)}\right| f(x)dxdy \\
&\le \iint \psi_{JK}^2(x,y)\frac{\|\Pi_K - \Pi_0\|_\infty}{c_0^2/2} f(x)dxdy \\
&\le \frac{2}{c_0^2}\iint \psi_{JK}^2(x,y) d\sqrt{\frac{\log n}{n}}\|\psi_{JK}\|_\infty f(x)dxdy \\
&\le \frac{2d\|\psi\|_\infty^2}{c_0}\left(\frac{\log n}{n}2^{j_1+j_2}\right)^{1/2} \to 0.
\end{aligned}$$

Consequently $|\mathbb{E}_{\Pi_K}(1/n\sum_{i=1}^n \Phi(X_i)) - 1|^2 \to 0$ and $(1/n)\sum_{i=1}^n \Phi(X_i) \to 1$. Thus K1) in Lemma 2.7 is satisfied and $n^{-1/2}\sum_{i=1}^n V_{JK}(X_i,X_{i+1}) \stackrel{\mathcal{L}(P_{\Pi_K})}{\longrightarrow} \mathcal{N}(0,1)$.

Since $\dfrac{z_{n,4}}{\sqrt{\log n}} \stackrel{\mathcal{L}(P_{\Pi_K})}{\longrightarrow} \mathcal{N}(0,1)$, there exists $\lambda > 0$ such that

$$P\left(\frac{z_{n,4}}{\sqrt{\log n}} \le \lambda\right) \to \frac{1}{3}.$$

Thus, for n large enough,

$$P(z_n^K > 0) = P\left(\frac{z_n^K}{\sqrt{\log n}} > 0\right) \geq 1 - \sum_{j=1}^{3} P\left(\frac{z_{n,j}}{\sqrt{\log n}} \leq -\frac{\lambda}{3}\right) - P\left(\frac{z_{n,4}}{\sqrt{\log n}} \leq \lambda\right)$$
$$\geq 1 - \sum_{j=1}^{3} P\left(|\frac{z_{n,j}}{\sqrt{\log n}}| \geq \frac{\lambda}{3}\right) - \frac{1}{3} \geq \frac{1}{3}.$$

Then $\Lambda_n(\Pi_0, \Pi_K) = \exp(z_n^K - v_n^K)$ with v_n^K constant and $P_{\Pi_K}(z_n^K > 0) \geq 1/3$ and the third point is verified.

4° : Regarding the fourth point,

$$\mathbb{E}_{\Pi_K}(V_{JK}(X_i, X_{i+1})^2) = \mathbb{E}_{\Pi_K}\left(\frac{\psi_{JK}^2(X_i, X_{i+1})}{\Pi_K^2(X_i, X_{i+1})}\right) = \iint \frac{\psi_{JK}^2(x,y)}{\Pi_K(x,y)} f(x)dxdy.$$

So

$$\mathbb{E}_{\Pi_K}(V_{JK}(X_i, X_{i+1})^2) \leq \|f\|_\infty \iint_A \frac{\psi_{JK}^2(x,y)}{c_0/2} dxdy \leq 2.$$

And then, using (2.26),

$$\sup_K v_n^K \leq d^2 \log n \leq \lambda_n$$

if $d^2 \leq 1/2(\bar{\alpha} + 1 - 2/r)$.

Finally Lemma 2.5 and Markov inequality yield

$$\sup_{\Pi \in \mathcal{B}} \mathbb{E}_\Pi \|\hat{\Pi}_n - \Pi\|_{p,A}^p \geq \sup_{\Pi \in \mathcal{B}} P_\Pi\left(\|\hat{\Pi}_n - \Pi\|_{p,A}^p \geq \left(\frac{\delta}{2}\right)^p\right)\left(\frac{\delta}{2}\right)^p \geq \frac{p_0}{2}\left(\frac{\delta}{2}\right)^p.$$

Then we use (2.25) to write

$$\sup_{\Pi \in \mathcal{B}} \mathbb{E}_\Pi \|\hat{\Pi}_n - \Pi\|_{p,A}^p \geq \frac{p_0}{2^p} d^p \left(\sqrt{\frac{\log n}{n}}\right)^p 2^{(j_1+j_2)(\frac{p}{2}-1)} \|\psi\|_p^{2p}.$$

Since

$$2^{(j_1+j_2)} \geq (c_1 c_2/4)\left(\frac{n}{\log n}\right)^{\frac{1}{\bar{\alpha}+1-2/r}},$$

we obtain

$$\sup_{\Pi \in \mathcal{B}} \mathbb{E}_\Pi \|\hat{\Pi}_n - \Pi\|_{p,A}^p \geq C\left(\frac{\log n}{n}\right)^{p\frac{\frac{\bar{\alpha}}{2}+\frac{1}{p}-\frac{1}{r}}{\bar{\alpha}+1-2/r}}$$

with $C = p_0 d^p \|\psi\|_p^{2p} (c_1 c_2)^{p/2-1}/2^{2p-2}$.

□

Deuxième partie

Estimation pour des chaînes de Markov cachées

Chapitre 3

Vitesses de convergence en déconvolution

Version modifiée de l'article *Rates of convergence for nonparametric deconvolution* paru aux Comptes Rendus de l'Académie des Sciences, vol. 342.

3.1 Introduction

We consider the following deconvolution problem:

$$Y_i = X_i + \varepsilon_i \qquad i = 1, \dots, n$$

where the X_i's are independent and identically distributed random variables with an unknown density g and the random variables ε_i are i.i.d with known density q. Moreover (X_i) and (ε_i) are independent. The aim is to estimate g from data $Y_1, \dots, Y_n$.

The hypothesis framework is the following. Denote, for all function u, u^* the Fourier transform of u: $u^*(x) = \int e^{-ixt} u(t)dt$. We suppose that noise is such that for all x in $\mathbb{R}$, $q^*(x) \neq 0$ and that it satisfies the following assumption:

H5 There exist $s \geq 0, b \geq 0, \gamma \in \mathbb{R}$ ($\gamma > 0$ if $s = 0$) and $k_0, k_1 > 0$ such that

$$k_0(x^2+1)^{-\gamma/2} \exp(-b|x|^s) \leq |q^*(x)| \leq k_1(x^2+1)^{-\gamma/2} \exp(-b|x|^s)$$

We assume that g belongs to the space

$$\mathcal{A}_{\delta,r,a}(L) = \{g : \mathbb{R} \mapsto \mathbb{R}, \int |g^*(x)|^2 (x^2+1)^{\delta} \exp(2a|x|^r) \leq L\}$$

with $r \geq 0, a \geq 0, \delta \in \mathbb{R}$ ($\delta > 1/2$ if $r = 0$), $L > 0$. When $r > 0$ the function is known as supersmooth, and as ordinary smooth else. The terminology is the same for noise.

This problem has been extensively studied for a function g belonging to a Sobolev or Hölder class (i.e. $r = 0$): see among others Carroll and Hall (1988), Devroye (1989), Fan (1991, 1993), Liu and Taylor (1989), Stefanski (1990). The bad rates of convergence (power of $\log n$) for supersmooth noise (and then in particular for Gaussian distributions) lead to consider supersmooth functions. First Pensky and Vidakovic (1999) and more recently Butucea (2004), Butucea and Tsybakov (2007) and Comte *et al.* (2006b) studied estimators in this context.

The contribution of this chapter is to provide exact and explicit rates of convergence, even in the case $r > 0$ and $s > 0$ where up to now the rates were not explicitly available except in very particular cases.

3.2 Estimators and preliminar results

3.2.1 Estimators

The classical kernel estimator is the following:

$$\hat{g}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - Y_i}{h}\right)$$

where K is the function defined as the inverse Fourier transform of

$$K^*(x) = \mathbb{1}_{\{|x|\leq 1\}}/q^*(-x/h).$$

The pointwise mean squared error (denoted by MSE) and mean integrated squared error (denoted by MISE) are established in Butucea and Tsybakov (2007):

Proposition 3.1 *If g belongs to $\mathcal{A}_{\delta,r,a}(L)$, then under Assumption H5,*

$$MISE = \mathbb{E}\|g - \hat{g}_n\|^2 = O\left(h^{2\delta}\exp(-2a/h^r) + \frac{h^{s-1-2\gamma}}{n}\exp(2b/h^s)\right) \quad \text{and}$$

$$MSE = \mathbb{E}|g(x) - \hat{g}_n(x)|^2 = O\left(h^{2\delta+r-1}\exp(-2a/h^r) + \min(1, h^{s-1})\frac{h^{s-1-2\gamma}}{n}\exp(2b/h^s)\right)$$

This estimator has the advantage to be optimal in the sharp asymptotic minimax sense (see Butucea and Tsybakov (2007)) but provides an adaptive estimator only in particular cases. That is why we present the projection estimator introduced in Comte *et al.* (2006b).

Let $\varphi(x) = \sin(\pi x)/(\pi x)$ and $\varphi_{m,j}(x) = \sqrt{L_m}\varphi(L_m x - j)$. Consider v_t the inverse Fourier transform of

$$v_t^*(x) = t^*(x)/q^*(-x).$$

Then the projection estimator is defined by

$$\hat{g}_m(x) = \sum_{|j|\leq K_n} \hat{a}_{m,j}\varphi_{m,j} \qquad \text{where } \hat{a}_{m,j} = \frac{1}{n}\sum_{i=1}^{n} v_{\varphi_{m,j}}(Y_i).$$

For this estimator, the following result is proved in Comte *et al.* (2006b):

Proposition 3.2 *Assume that $q \in L^2$ (i.e. $\gamma > 1/2$ when $s = 0$) and that g is a L^2 function which verifies $\int x^2 g^2(x)dx \leq M$. If g belongs to $\mathcal{A}_{\delta,r,a}(L)$, then under Assumption H5,*

$$MISE = \mathbb{E}\|g - \hat{g}_m\|^2 = O\left(L_m^{-2\delta}\exp(-2a\pi^r L_m^r) + \frac{L_m^{2\gamma+1-s}}{n}\exp(2b\pi^s L_m^s)\right)$$

Then, an adaptive estimator can be defined using a model selection method (see Comte *et al.* (2006b) for details).

3.2.2 Rates of convergence

We can observe that both estimators have the same L^2 rate of convergence (take $h^{-1} = \pi L_m$). To compute this rate, we have to minimize the risk orders in h (or L_m). By setting to zero the derivative of this quantity we obtain the equation

$$\exp\left(\frac{2b}{h^s} + \frac{2a}{h^r}\right)h^\alpha = O(n) \tag{3.1}$$

where $\alpha = r-2\delta-2\gamma-1$ if we consider the integrated error and $\alpha = -2\delta-2\gamma+(s-1)_+$ if we consider the pointwise error. In most cases, the solution of this equation is well known and leads to Tables 3.1 and 3.2 where different regularities for g and q are examined.

	$s=0$	$s>0$
$r=0$	$n^{-\frac{2\delta}{2\delta+2\gamma+1}}$	$(\log n)^{-\frac{2\delta}{s}}$
$r>0$	$\frac{(\log n)^{\frac{2\gamma+1}{r}}}{n}$	Theorem 3.1

Table 3.1: Rates of convergence for the MISE.

	$s=0$	$s>0$
$r=0$	$n^{\frac{1-2\delta}{2\delta+2\gamma}}$	$(\log n)^{\frac{1-2\delta}{s}}$
$r>0$	$\frac{(\log n)^{\frac{2\gamma+1}{r}}}{n}$	Theorem 3.1

Table 3.2: Rates of convergence for the MSE.

Except for the bottom right cells (to be completed in the next section), these rates are known to be optimal minimax rates: see Fan (1991) and Butucea (2004) for the lower bounds.

3.3 Results

The rates of convergence in the case $(r>0, s>0)$ depend on the integer k such that r/s or s/r belongs to the interval $(k/(k+1),(k+1)/(k+2)]$:

Theorem 3.1 *We assume $r>0$ and $s>0$. Let $k\in\mathbb{N}$ and $\lambda=\mu^{-1}=r/s$. Then*

- *if $r=s$, if $\xi=[2\delta b+(s-2\gamma-1)a]/[(a+b)s]$*

$$MISE = O\left(n^{-a/(a+b)}(\log n)^{-\xi}\right);$$

$$MSE = O\left(n^{-a/(a+b)}(\log n)^{-\xi+\frac{(1-s)_+ b}{(a+b)s}}\right)$$

- *if $r<s$ and $\dfrac{k}{k+1}<\lambda\leq\dfrac{k+1}{k+2}$, there exist reals b_i such that*

$$MISE = O\left((\log n)^{-2\delta/s}\exp[\sum_{i=0}^{k} b_i(\log n)^{(i+1)\lambda-i}]\right);$$

$$MSE = O\left((\log n)^{(-2\delta-r+1)/s}\exp[\sum_{i=0}^{k} b_i(\log n)^{(i+1)\lambda-i}]\right)$$

- *if* $r > s$ *and* $\dfrac{k}{k+1} < \mu \leq \dfrac{k+1}{k+2}$, *there exist reals* d_i *such that,*

$$MISE = O\left(\frac{(\log n)^{(1+2\gamma-s)/r}}{n}\exp[-\sum_{i=0}^{k} d_i(\log n)^{(i+1)\mu-i}]\right);$$

$$MSE = O\left(\frac{(\log n)^{(1+2\gamma-s-(s-1)_+)/r}}{n}\exp[-\sum_{i=0}^{k} d_i(\log n)^{(i+1)\mu-i}]\right)$$

The coefficients b_i and d_i are computable, see Section 3.4 for the exact form of reals b_i. Notice that these original rates have the property to decrease faster than any logarithmic function. Moreover, they are optimal in the cases where the corresponding lower bounds are known, i.e $r = s = 1$ (see Tsybakov (2000)) and $r < s$ (see Butucea and Tsybakov (2007)). We can also remark that, given the complexity of these rates, it is worth finding adaptive estimators, i.e. estimators whose risk automatically achieves the minimax rates.

3.4 Proof

Here we prove Theorem 3.1. We denote by $W(h)$ the quantity $h^\alpha \exp\left(\dfrac{2b}{h^s} + \dfrac{2a}{h^r}\right)$. We look for the optimal h^* such that $W(h^*) = O(n)$. Now let

$$G(h) = h^{2\delta}\exp(-2a/h^r) + \frac{h^{s-1-2\gamma}}{n}\exp(2b/h^s)$$

the order of the integrated risk and

$$L(h) = h^{2\delta+r-1}\exp(-2a/h^r) + \min(1, h^{s-1})\frac{h^{s-1-2\gamma}}{n}\exp(2b/h^s)$$

the order of the pointwise risk.

- **case** $r = s$**.** Let

$$h = \left(\frac{\log n + \alpha/s\log\log n}{2a+2b}\right)^{-1/s}.$$

Then

$$\begin{aligned} W(h) &= h^\alpha e^{(2a+2b)/h^s} = \left(\frac{\log n + \alpha/s\log\log n}{2a+2b}\right)^{-\alpha/s} e^{\log n+\alpha/s\log\log n} \\ &= Kn(\log n + \alpha/s\log\log n)^{-\alpha/s}(\log n)^{\alpha/s} \\ &= Kn(1+\frac{\alpha/s\log\log n}{\log n})^{-\alpha/s} = O(n). \end{aligned}$$

This $h = h^*$ is thus appropriate whatever α. We can now compute the corresponding risk, by denoting $w_n = (\alpha \log\log n)/(s \log n)$,

$$\begin{aligned} G(h^*) &= \left(\frac{\log n + \frac{\alpha}{s}\log\log n}{2a+2b}\right)^{\frac{-2\delta}{s}} \exp\left(-\frac{a}{a+b}(\log n + \frac{\alpha}{s}\log\log n)\right) \\ &\quad + \left(\frac{\log n + \frac{\alpha}{s}\log\log n}{2a+2b}\right)^{\frac{1-s+2\gamma}{s}} \frac{1}{n}\exp\left(\frac{b}{a+b}(\log n + \frac{\alpha}{s}\log\log n)\right) \\ &= n^{-\frac{a}{a+b}}[K_1(1+w_n)^{-\frac{2\delta}{s}}(\log n)^{-\frac{2\delta}{s}-\frac{a}{a+b}\frac{\alpha}{s}} \\ &\quad + K_2(1+w_n)^{\frac{1-s+2\gamma}{s}}(\log n)^{\frac{1-s+2\gamma}{s}+\frac{b}{a+b}\frac{\alpha}{s}}] \\ &= O(n^{-\frac{a}{a+b}}[(\log n)^{\frac{-2\delta(a+b)-a\alpha}{(a+b)s}} + (\log n)^{\frac{(1-s+2\gamma)(a+b)+b\alpha}{(a+b)s}}]) \qquad (3.2) \\ &= O(n^{-\frac{a}{a+b}}(\log n)^{\frac{-2\delta b - a(s-2\gamma-1)}{(a+b)s}}) \end{aligned}$$

since $\alpha = s - 2\delta - 2\gamma - 1$. Then

$$MISE = O\left(n^{-\frac{a}{a+b}}(\log n)^{-\frac{2\delta b+(s-2\gamma-1)a}{(a+b)s}}\right).$$

For the pointwise risk, it is sufficient to write $\alpha = -2\delta - 2\gamma + (s-1)_+$ in (3.2):

$$\begin{aligned} L(h^*) &= \begin{cases} O\left(n^{-\frac{a}{a+b}}(\log n)^{\frac{-2\delta b - a(s-2\gamma-1)}{(a+b)s}}\right) & \text{if } s \geq 1 \\ O\left(n^{-\frac{a}{a+b}}(\log n)^{\frac{-(2\delta+s-1)b - a(s-2\gamma-1)}{(a+b)s}}[(\log n)^{\frac{(a+b)(s-1)}{(a+b)s}} + 1]\right) & \text{if } s < 1 \end{cases} \\ &= \begin{cases} O\left(n^{-\frac{a}{a+b}}(\log n)^{\frac{-2\delta b - a(s-2\gamma-1)}{(a+b)s}}\right) & \text{if } s \geq 1 \\ O\left(n^{-\frac{a}{a+b}}(\log n)^{\frac{-(2\delta+s-1)b - a(s-2\gamma-1)}{(a+b)s}}\right) & \text{if } s < 1 \end{cases} \end{aligned}$$

And then

$$MSE = O\left(n^{-\frac{a}{a+b}}(\log n)^{-\frac{(2\delta-(1-s)_+)b+(s-2\gamma-1)a}{(a+b)s}}\right).$$

- **case** $r < s \Leftrightarrow \lambda = r/s < 1$. Let $c = \dfrac{2a}{(2b)^\lambda}$. The optimal h^* depends on the value of λ. Let $k \in \mathbb{N}$ and

$$h = (2b)^{1/s}\left[\log n + \frac{\alpha}{s}\log\log n + \sum_{i=0}^{k} b_i (\log n)^{(i+1)\lambda - i}\right]^{-1/s}$$

where the b_i's are reals to be specified later. Let us calculate $W(h)$:

$$\begin{aligned} W(h) &= \exp[\frac{2b}{h^s}+\frac{2a}{h^r}+\alpha\log h] = \exp\Big[\log n+\frac{\alpha}{s}\log\log n+\sum_{i=0}^{k} b_i(\log n)^{(i+1)\lambda-i} \\ &\quad +\frac{2a}{(2b)^{r/s}}(\log n)^{r/s}\left(1+\frac{\alpha}{s}\frac{\log\log n}{\log n}+\sum_{i=0}^{k} b_i(\log n)^{(i+1)\lambda-(i+1)}\right)^{r/s} \\ &\quad +\frac{\alpha}{s}\log(2b)-\frac{\alpha}{s}\log\log n \\ &\quad -\frac{\alpha}{s}\log\left(1+\frac{\alpha}{s}\frac{\log\log n}{\log n}+\sum_{i=0}^{k} b_i(\log n)^{(i+1)\lambda-(i+1)}\right)\Big] \\ &= Kn\exp[\sum_{i=0}^{k} b_i(\log n)^{(i+1)\lambda-i}+c(\log n)^{\lambda}(1+u_n)^{\lambda}-\frac{\alpha}{s}\log(1+u_n)] \end{aligned}$$

where $u_n := \frac{\alpha}{s}\frac{\log\log n}{\log n}+\sum_{i=0}^{k} b_i(\log n)^{(i+1)\lambda-(i+1)} = o(1)$. Then

$$\begin{aligned} W(h) &= Kn\exp[\sum_{i=0}^{k} b_i(\log n)^{(i+1)\lambda-i}-\frac{\alpha}{s}\log(1+o(1)) \\ &\quad +c(\log n)^{\lambda}(1+\lambda u_n+..+\frac{\lambda(\lambda-1)..(\lambda-k)}{(k+1)!}u_n^{k+1}+o(u_n^{k+1}))] \\ &= Kn\exp[o(1)+\sum_{i=0}^{k} b_i(\log n)^{(i+1)\lambda-i}+c(\log n)^{\lambda} \\ &\quad +A_1+..+A_{k+1}+o((\log n)^{\lambda}u_n^{k+1})] \end{aligned}$$

where $A_j = c\frac{\lambda\ldots(\lambda-j+1)}{j!}(\log n)^{\lambda}u_n^j$ for $j=1,..,k+1$. But

$$\begin{aligned} u_n^j &= \left[\frac{\alpha}{s}\frac{\log\log n}{\log n}+\sum_{i=0}^{k} b_i(\log n)^{(\lambda-1)(i+1)}\right]^j \\ &= O\left(\frac{\log\log n}{\log n}\right)+\left[\sum_{p=1}^{k+1} b_{p-1}(\log n)^{(\lambda-1)p}\right]^j \\ &= O\left(\frac{\log\log n}{\log n}\right)+\sum_{i=j}^{k+1}\sum_{p_1+\ldots+p_j=i} b_{p_1-1}\ldots b_{p_j-1}(\log n)^{(\lambda-1)i}+o((\log n)^{(k+1)(\lambda-1)}). \end{aligned}$$

And then

$$\begin{aligned} A_j &= o(1)+c\frac{\lambda\ldots(\lambda-j+1)}{j!}\sum_{i=j}^{k+1}\sum_{p_1+\ldots+p_j=i} b_{p_1-1}\ldots b_{p_j-1}(\log n)^{(i+1)\lambda-i} \\ &\quad +o((\log n)^{(k+2)\lambda-(k+1)}). \end{aligned}$$

By writing together the terms before $(\log n)^{(i+1)\lambda-i}$, we obtain

$$W(h) = Kn \exp[o(1) + \sum_{i=0}^{k+1} M_i (\log n)^{(i+1)\lambda-i} + o((\log n)^{(k+2)\lambda-(k+1)})]$$

with $\begin{cases} M_0 = b_0 + c, \\ M_i = b_i + c\sum_{j=1}^{i} \frac{\lambda\ldots(\lambda-j+1)}{j!} \sum_{p_1+\ldots+p_j=i} b_{p_1-1}\ldots b_{p_j-1} & \text{for } 1 \le i \le k, \\ M_{k+1} = c\sum_{j=1}^{k+1} \frac{\lambda\ldots(\lambda-j+1)}{j!} \sum_{p_1+\ldots+p_j=i} b_{p_1-1}\ldots b_{p_j-1}. \end{cases}$

We observe that for all i, M_i depends only on $b_0, b_1, \ldots, b_i$. We can thus define the coefficients b_i by setting the conditions $M_0 = 0, M_1 = 0, .., M_k = 0$. We have then

$$W(h) = O(n) \exp[M_{k+1}(\log n)^{(k+2)\lambda-(k+1)} + o((\log n)^{(k+2)\lambda-(k+1)})].$$

If $\lambda \le (k+1)/(k+2)$ then $W(h) = O(n)$. So we choose this $h = h^*$ if $k/(k+1) < \lambda \le (k+1)/(k+2)$. If $\lambda \le k/(k+1)$, we use a shorter expansion

$$h = (2b)^{1/s} \left[\log n + \frac{\alpha}{s} \log\log n + \sum_{i=0}^{k-1} b_i (\log n)^{(i+1)\lambda-i} \right]^{-1/s}.$$

Now we have to compute the corresponding risk:

$$\begin{aligned} G(h^*) &= (2b)^{2\delta/s} (\log n(1+u_n))^{\frac{-2\delta}{s}} \exp\left[-\frac{2a}{(2b)^{r/s}} (\log n)^{r/s} (1+u_n)^{r/s} \right] \\ &\quad + \frac{(2b)^{\frac{s-1-2\gamma}{s}}}{n} (\log n(1+u_n))^{\frac{2\gamma+1-s}{s}} \exp\left[\log n + \frac{\alpha}{s}\log\log n + \sum_{i=0}^{k} b_i (\log n)^{(i+1)\lambda-i} \right] \\ &= K_1(1+o(1))(\log n)^{\frac{-2\delta}{s}} \exp[-c(\log n)^{\lambda}(1+u_n)^{\lambda}] \\ &\quad + K_2(1+o(1))(\log n)^{\frac{2\gamma+1-s+\alpha}{s}} \exp[\sum_{i=0}^{k} b_i (\log n)^{(i+1)\lambda-i}]. \end{aligned}$$

But, using the previous calculations

$$\begin{aligned} c(\log n)^{\lambda}(1+u_n)^{\lambda} &= c(\log n)^{\lambda} + A_1 + A_2 + .. + A_{k+1} + o((\log n)^{(k+2)\lambda-(k+1)}) \\ &= \sum_{i=0}^{k} (M_i - b_i)(\log n)^{(i+1)\lambda-i} + O(1) \\ &= -\sum_{i=0}^{k} b_i (\log n)^{(i+1)\lambda-i} + O(1). \end{aligned}$$

So the risk can be written

$$G(h^*) = O\Bigg((\log n)^{\frac{-2\delta}{s}} \exp[\sum_{i=0}^{k} b_i (\log n)^{(i+1)\lambda-i} + O(1)] + (\log n)^{\frac{2\gamma+1-s+\alpha}{s}} \exp[\sum_{i=0}^{k} b_i (\log n)^{(i+1)\lambda-i}] \Bigg).$$

And finally, since $2\gamma+1-s+\alpha=-2\delta$,

$$MISE = O\left((\log n)^{\frac{-2\delta}{s}}\exp[\sum_{i=0}^{k} b_i(\log n)^{(i+1)\lambda-i}]\right).$$

Regarding the pointwise risk:

$$\begin{aligned} L(h^*) &= K_1(\log n(1+u_n))^{\frac{-2\delta-r+1}{s}}\exp\left[-\frac{2a}{(2b)^{r/s}}(\log n)^{r/s}(1+u_n)^{r/s}\right] \\ &+\frac{K_2}{n}(\log n(1+u_n))^{\frac{2\gamma+1-s-(s-1)_+}{s}}\exp\left[\log n+\frac{\alpha}{s}\log\log n+\sum_{i=0}^{k} b_i(\log n)^{(i+1)\lambda-i}\right] \\ &= O\left([(\log n)^{\frac{-2\delta-r+1}{s}}+(\log n)^{\frac{2\gamma+1-s-(s-1)_++\alpha}{s}}]\exp[\sum_{i=0}^{k} b_i(\log n)^{(i+1)\lambda-i}]\right). \end{aligned}$$

Now we use that in this case $\alpha=-2\delta-2\gamma+(s-1)_+$ and so

$$L(h^*) = O\left([1+(\log n)^{\frac{r-s}{s}}](\log n)^{\frac{-2\delta-r+1}{s}}\exp[\sum_{i=0}^{k} b_i(\log n)^{(i+1)\lambda-i}]\right).$$

Given that $r-s<0$, $MSE = O\left((\log n)^{\frac{-2\delta-r+1}{s}}\exp[\sum_{i=0}^{k} b_i(\log n)^{(i+1)\lambda-i}]\right)$.

- **case** $r>s \Leftrightarrow \mu=s/r<1$. Let $q=\dfrac{2b}{(2a)^{\mu}}$. We can proceed as in the case $r<s$. For $k/(k+1)<\mu\leq(k+1)(k+2)$, let

$$h^* = (2a)^{1/r}[\log n+\frac{\alpha}{r}\log\log n+\sum_{i=0}^{k} d_i(\log n)^{(i+1)\mu-i}]^{-1/r}$$

with $\begin{cases} d_0=-q, \\ d_i=-q\sum_{j=1}^{i}\frac{\mu\ldots(\mu-j+1)}{j!}\sum_{p_1+\ldots+p_j=i} d_{p_1-1}\ldots d_{p_j-1} & \text{for } 1\leq i\leq k. \end{cases}$

This h^* verifies $W(h^*)=O(n)$. Let $v_n=\alpha\log\log n/r\log n+\sum_{i=0}^{k} d_i(\log n)^{(i+1)\mu-(i+1)}$. We compute the L^2 risk for $k/(k+1)<\mu\leq(k+1)/(k+2)$:

$$\begin{aligned} G(h^*) &= K_1(\log n(1+v_n))^{\frac{-2\delta}{r}}\exp\left[-\log n-\frac{\alpha}{r}\log\log n-\sum_{i=0}^{k} d_i(\log n)^{(i+1)\mu-i}\right] \\ &+\frac{K_2}{n}(\log n(1+v_n))^{\frac{2\gamma+1-s}{r}}\exp\left[\frac{2b}{(2a)^{s/r}}(\log n)^{s/r}(1+v_n)^{s/r}\right] \\ &= O\Big(\frac{(\log n)^{\frac{-2\delta-\alpha}{r}}}{n}\exp\left[-\sum_{i=0}^{k} d_i(\log n)^{(i+1)\mu-i}\right] \\ &+\frac{(\log n)^{\frac{2\gamma+1-s}{r}}}{n}\exp[q(\log n)^{\mu}(1+v_n)^{\mu}]\Big). \end{aligned}$$

But, as previously, $q(\log n)^{\mu}(1+v_n)^{\mu} = -\sum_{i=0}^{k} d_i(\log n)^{(i+1)\mu-i} + O(1)$ and then

$$MISE = O\left(\frac{(\log n)^{\frac{2\gamma+1-s}{r}}}{n}\exp[-\sum_{i=0}^{k} d_i(\log n)^{(i+1)\mu-i}]\right).$$

It remains to compute $L(h^*)$.

$$\begin{aligned}
L(h^*) &= K_1(\log n(1+v_n))^{\frac{-2\delta-r+1}{r}}\exp\left[-\log n - \frac{\alpha}{r}\log\log n - \sum_{i=0}^{k} d_i(\log n)^{(i+1)\mu-i}\right] \\
&\quad + \frac{K_2}{n}(\log n(1+v_n))^{\frac{2\gamma+1-s-(s-1)_+}{r}}\exp\left[\frac{2b}{(2a)^{s/r}}(\log n)^{s/r}(1+v_n)^{s/r}\right] \\
&= O\Bigg(\frac{(\log n)^{\frac{-2\delta-r+1-\alpha}{r}}}{n}\exp\left[-\sum_{i=0}^{k} d_i(\log n)^{(i+1)\mu-i}\right] \\
&\quad + \frac{(\log n)^{\frac{2\gamma+1-s-(s-1)_+}{r}}}{n}\exp[q(\log n)^{\mu}(1+v_n)^{\mu}]\Bigg) \\
&= O\left(\frac{(\log n)^{\frac{2\gamma+1-s-(s-1)_+}{r}}}{n}[(\log n)^{\frac{s-r}{r}}+1]\exp[-\sum_{i=0}^{k} d_i(\log n)^{(i+1)\mu-i}]\right),
\end{aligned}$$

and finally

$$MSE = O\left(\frac{(\log n)^{\frac{2\gamma+1-s-(s-1)_+}{r}}}{n}\exp[-\sum_{i=0}^{k} d_i(\log n)^{(i+1)\mu-i}]\right).$$

Chapitre 4

Estimation de la densité de transition par quotient

Version modifiée de l'article *Adaptive estimation of the transition density of a particular hidden Markov chain* accepté pour publication à Journal of Multivariate Analysis.

4.1 Introduction

Let us consider the following model:

$$Y_i = X_i + \varepsilon_i \qquad i = 1, \ldots, n+1 \tag{4.1}$$

where $(X_i)_{i\geq 1}$ is an irreducible and positive recurrent Markov chain and $(\varepsilon_i)_{i\geq 1}$ is a noise independent of $(X_i)_{i\geq 1}$. We assume that $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed random variables with known distribution. This model belongs to the class of hidden Markov models.

The Hidden Markov Models constitute a very famous class of discrete-time stochastic processes, with many applications in various areas such as biology, speech recognition or finance. For a general reference on these models, we refer to Cappé *et al.* (2005). Here, we study a simple model of HMM where the noise is additive (which allows to deal also with multiplicative noise by use of logarithm). In standard HMM, it is assumed that the joint density of (X_i, Y_i) has a parametric form and the aim is then to infer the parameter from the observations $Y_1, ..., Y_n$, generally by maximizing the likelihood. For this type of study, we can cite among others Baum and Petrie (1966), Leroux (1992), Bakry *et al.* (1997), Bickel *et al.* (1998), Jensen and Petersen (1999), Douc *et al.* (2004).

Here, we are interested in a nonparametric approach of the estimation of the hidden chain transition. A nonparametric model is particularly useful in the financial field (for instance in stochastic volatility model) where the form of the chain, which is usually derived from a diffusion, can be entirely unknown. So we assume that the Markov chain law is entirely unknown. Matias (2002) and Butucea and Matias (2005) considered the semiparametric problem where X_i follows an unknown distribution and the emission distribution has an unknown variance. The identifiability requires then for the signal density to be less regular than the density of the noise. Here, we assume that all regularities (in the sense defined below) for both distributions are possible but the noise distribution is completely known.

Our model is then a convolution model but with dependent variables X_i. The estimation of the density of X_i from the observations $Y_1, \ldots, Y_n$ when the X_i's are i.i.d. (the so-called convolution model) has been extensively studied, see e.g. Carroll and Hall (1988), Fan (1991), Stefanski (1990), Pensky and Vidakovic (1999), Comte *et al.* (2006b). However, very few authors study the case where (X_i) is a Markov chain. We can cite Dorea and Zhao (2002) who estimate the density of Y_i in a very general context of HMM, Masry (1993) who is interested in the estimation of the multivariate density in a mixing framework and Clémençon (2003) who estimates the stationary density and the transition density of the hidden chain in the model (4.1). More precisely he introduces an estimator of the transition density based on the thresholding of a wavelet-vaguelette decomposition and he studies its performance in the case of an ordinary smooth noise, that is with a polynomial decay of its Fourier transform.

Here, we are also interested in the estimation of the transition density of (X_i) but we

consider a larger class of noise distributions. In Clémençon (2003) there is no study of supersmooth noise (i.e. with exponentially decreasing Fourier transform), as the Gaussian distribution. However, the study of such noise is essential for the applications and gives interesting rates of convergence, in particular when the chain density is also supersmooth. In the present chapter, the four cases (ordinary smooth or supersmooth noise with ordinary smooth or supersmooth chain) are considered.

The aim of this chapter is to estimate the transition density Π of the Markov chain (X_i) from the observations $Y_1, \dots, Y_n$. To do this, we assume that the regime is stationary and we note that $\Pi = F/f$ where F is the density of (X_i, X_{i+1}) and f the stationary density. The estimation of f comes down to a problem of deconvolution, as does the estimation of F. We use contrast minimization and a model selection method inspired by Barron *et al.* (1999) to find adaptive estimators of f and F. Our estimator of Π is then the quotient of the two previous estimators. Note that it is worth finding an adaptive estimator, i.e. an estimator whose risk automatically achieves the minimax rate, because the regularity of the densities f and F is generally very hard to compute, even if the chain can be fully described (as it is the case for a diffusion or an autoregressive process).

We study the performance of our estimator by computing the rate of convergence of the integrated risk. We improve the result of Clémençon (2003) (case of an ordinary smooth noise) since we obtain the minimax rate without logarithmic loss. Moreover, we observe noticeable rates of convergence when both the noise and the chain are supersmooth.

The chapter is organized as follows. Section 4.2 is devoted to notations and assumptions while the estimation procedure is developed in Section 4.3. After describing the projection spaces to which the estimators belong, we define separately the estimator of the stationary density f, the one of the joint density F and in the end the estimator $\tilde{\Pi}$ of the transition density. Section 4.4 states the results obtained for our estimators. To illustrate the theorems, some examples are provided in Section 4.5 as the AR(1) model, the Cox-Ingersoll-Ross process or the stochastic volatility model. Some simulations are presented in Section 4.6 and the proofs are to be found in Section 4.7.

4.2 Notations and Assumptions

For the sake of clarity, we use lowercase letters for dimension 1 and capital letters for dimension 2. For a function $t : \mathbb{R} \mapsto \mathbb{R}$, we denote by $\|t\|$ the L^2 norm that is $\|t\|^2 = \int_{\mathbb{R}} t^2(x)dx$. The Fourier transform t^* of t is defined by

$$t^*(u) = \int e^{-ixu} t(x)dx$$

Note that the function t is the inverse Fourier transform of t^* and can be written $t(x) = 1/(2\pi) \int e^{ixu} t^*(u)du$. The convolution product is defined by $(t * s)(x) = \int t(x-y)s(y)dy$.

In the same way, for a function $T : \mathbb{R}^2 \mapsto \mathbb{R}$, $\|T\|^2 = \iint_{\mathbb{R}^2} T^2(x,y)dxdy$ and

$$T^*(u,v) = \iint e^{-ixu-iyv} T(x,y)dxdy, \quad (T * S)(x,y) = \iint T(x-z, y-w)S(z,w)dzdw.$$

We denote by $t \otimes s$ the function: $(x,y) \mapsto (t \otimes s)(x,y) = t(x)s(y)$.

The density of ε_i is named q and is known. We denote by p the unknown density of Y_i. We have $p = f * q$ and then $p^* = f^* q^*$.

Now the assumptions on the model are the following:

H1 The chain is irreducible, positive recurrent and stationary with (unknown) density f.

H2b The chain is geometrically β-mixing ($\beta_q \leq Me^{-\theta q}$), or arithmetically β-mixing ($\beta_q \leq Mq^{-\theta}$) with $\theta > 8$.

H3: **H3c** $f \in (L^2 \cap L^\infty)(\mathbb{R})$
H3d $F \in (L^2 \cap L^\infty)(\mathbb{R}^2)$

H5 Function q^* never vanishes and there exist $s \geq 0, b > 0$, $\gamma \in \mathbb{R}$ ($\gamma > 0$ if $s = 0$) and $k_0, k_1 > 0$ such that

$$k_0(x^2+1)^{-\gamma/2} \exp(-b|x|^s) \leq |q^*(x)| \leq k_1(x^2+1)^{-\gamma/2} \exp(-b|x|^s).$$

The condition H2b is verified as soon as the chain is uniformly ergodic. A definition of the β-mixing coefficients (in general and in the case of a Markov chain) can be found in Doukhan (1994). A lot of Markov chains satisfy these assumptions A4, see examples in Section 1.2.2 in Chapter 1.

In the sequel we consider the following smoothness spaces:

$$\mathcal{A}_{\delta,r,a}(l) = \{f : \mathbb{R} \mapsto \mathbb{R}, \int |f^*(x)|^2 (x^2+1)^\delta \exp(2a|x|^r)dx \leq l\}$$

with $r \geq 0, a > 0$, $\delta \in \mathbb{R}$ ($\delta > 1/2$ if $r = 0$), $l > 0$ and

$$\mathbb{A}_{\Delta,R,A}(L) = \{F : \mathbb{R}^2 \mapsto \mathbb{R}, \iint |F^*(x,y)|^2 (x^2+1)^\Delta (y^2+1)^\Delta \exp(2A(|x|^R + |y|^R))dxdy \leq L\}$$

with $R \geq 0, A > 0$, $\Delta \in \mathbb{R}$ ($\Delta > 1/2$ if $R = 0$), $L > 0$.

When $r > 0$ (respectively $R > 0$) the function f (resp. F) is known as supersmooth, and as ordinary smooth otherwise. In the same way, the noise distribution is called ordinary smooth if $s = 0$ and supersmooth otherwise. The spaces of ordinary smooth functions correspond to classic Sobolev classes, while supersmooth functions are infinitely differentiable. It includes for example normal ($r = 2$) and Cauchy ($r = 1$) densities.

It is worth noting that as F is the density of (X_i, X_{i+1}), the two directions play a similar role. Thus, there is no use considering more general functional spaces for F, like anisotropic ones (see Lepski and Levit (1999)).

4.3 Estimation procedure

Since $\Pi = F/f$ we proceed in 3 steps to estimate the transition density Π. First we find an estimator $\tilde{f}$ of f (see Section 4.3.2). Then we estimate F by $\tilde{F}$ (see Section 4.3.3). And finally we estimate Π with the quotient $\tilde{F}/\tilde{f}$ (Section 4.3.4).

All estimators defined here are projection estimators. We therefore start with describing the projection spaces.

4.3.1 Projection spaces

Let us consider the function

$$\varphi(x) = \sin(\pi x)/(\pi x)$$

and, for m in $\mathbb{N}^*$, j in $\mathbb{Z}$, $\varphi_{m,j}(x) = \sqrt{m}\varphi(mx-j)$. Note that $\{\varphi_{m,j}\}_{j\in\mathbb{Z}}$ is an orthonormal basis of the space of integrable functions having a Fourier transform with compact support included into $[-\pi m, \pi m]$. In the sequel, we use the following notations:

$$S_m = \mathrm{Span}\{\varphi_{m,j}\}_{j\in\mathbb{Z}}; \quad \mathbb{S}_m = \mathrm{Span}\{\varphi_{m,j} \otimes \varphi_{m,k}\}_{j,k\in\mathbb{Z}}$$

These spaces have particular properties, which are a consequence of the first point of Lemma 4.3 (see Section 4.7.8):

$$\forall t \in S_m \qquad \|t\|_\infty \leq \sqrt{m}\|t\|; \quad \forall T \in \mathbb{S}_m \qquad \|T\|_\infty \leq m\|T\| \tag{4.2}$$

where $\|t\|_\infty = \sup_{x\in\mathbb{R}} |t(x)|$ and $\|T\|_\infty = \sup_{(x,y)\in\mathbb{R}^2} |T(x,y)|$.

4.3.2 Estimation of f

Here, we estimate f, which is the density of the X_i's. It is the classic deconvolution problem. We choose to estimate f by minimizing a contrast. The standard contrast in density estimation is $(1/n)\sum_{i=1}^n [\|t\|^2 - 2t(X_i)]$. It is not possible to use this contrast here since we do not observe $X_1, \ldots, X_n$. Only the noisy data $Y_1, \ldots, Y_n$ are available. That is why we use the following lemma.

Lemma 4.1 *For all function t, let v_t be the inverse Fourier transform of $t^*/q^*(-.)$, i.e.*

$$v_t(x) = \frac{1}{2\pi}\int e^{ixu}\frac{t^*(u)}{q^*(-u)}du.$$

Then, for all $1 \leq k \leq n$,

1. $\mathbb{E}[v_t(Y_k)|X_1, ..., X_n] = t(X_k)$
2. $\mathbb{E}[v_t(Y_k)] = \mathbb{E}[t(X_k)]$

The second assertion in Lemma 4.1 is an obvious consequence of the first one and leads us to consider the following contrast:

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n}[\|t\|^2 - 2v_t(Y_i)] \quad \text{with} \quad v_t^*(u) = \frac{t^*(u)}{q^*(-u)}. \tag{4.3}$$

Indeed, since $t(X_i)$ and $v_t(Y_i)$ have the same expectation, it is natural to replace the unknown quantity $t(X_i)$ in the contrast by $v_t(Y_i)$.

We can observe that $\mathbb{E}\gamma_n(t) = (1/n)\sum_{i=1}^n[\|t\|^2 - 2\mathbb{E}[v_t(Y_i)]] = (1/n)\sum_{i=1}^n[\|t\|^2 - 2\mathbb{E}[t(X_i)]] = \|t\|^2 - 2\int tf = \|t-f\|^2 - \|f\|^2$ and then minimizing $\gamma_n(t)$ comes down to minimizing the distance between t and f. So we define

$$\hat{f}_m = \arg\min_{t\in S_m} \gamma_n(t) \tag{4.4}$$

or, equivalently,

$$\hat{f}_m = \sum_{j\in\mathbb{Z}} \hat{a}_j\varphi_{m,j} \quad \text{with} \quad \hat{a}_j = \frac{1}{n}\sum_{i=1}^{n} v_{\varphi_{m,j}}(Y_i).$$

It is sufficient to differentiate the contrast to obtain this expression of the estimator. Actually, we should define $\hat{f}_m = \sum_{|j|\le K_n} \hat{a}_j\varphi_{m,j}$ because we can estimate only a finite number of coefficients. If K_n is suitably chosen, it does not change the rate of convergence since the additional terms can be made negligible. For the sake of simplicity, we let the sum over $\mathbb{Z}$. For an example of detailed truncation see Comte *et al.* (2006b).

Let f_m be the orthogonal projection of f on S_m, then

$$f_m = \sum_{j\in\mathbb{Z}}(\int f\varphi_{m,j})\varphi_{m,j} = \sum_{j\in\mathbb{Z}} \mathbb{E}(\hat{a}_j)\varphi_{m,j}.$$

Conditionally to (X_i), the variance or stochastic error is

$$\begin{aligned}\mathbb{E}[\|\hat{f}_m - f_m\|^2|X_1,\dots,X_n] &= \mathbb{E}[\sum_j(\hat{a}_j - \mathbb{E}(\hat{a}_j))^2|X_1,\dots,X_n] \\ &= \sum_j \mathrm{Var}[\frac{1}{n}\sum_{i=1}^{n} v_{\varphi_{m,j}}(Y_i)|X_1,\dots,X_n] \le \frac{\|\sum_j v_{\varphi_{m,j}}^2\|_\infty}{n}\end{aligned} \tag{4.5}$$

since $Y_1,\dots,Y_n$ are independent conditionally to (X_i). Then, it follows from Lemma 4.3 (see Section 4.7.8) that $\|\sum_j v_{\varphi_{m,j}}^2\|_\infty = \Delta(m)$ where

$$\Delta(m) = \frac{1}{2\pi}\int_{-\pi m}^{\pi m} |q^*(u)|^{-2}du. \tag{4.6}$$

with q^* the characteristic function of the noise (ε_i). This implies that the order of the variance is $\Delta(m)/n$. That is why we introduce

$$\mathcal{M}_n = \left\{ m \geq 1, \quad \frac{\Delta(m)}{n} \leq 1 \right\}.$$

To complete the estimation, we choose the best estimator among the collection $(\hat{f}_m)_{m\in\mathcal{M}_n}$. To do this, we select the model which minimizes the following penalized criterion. Let

$$\hat{m} = \arg\min_{m\in\mathcal{M}_n} \{\gamma_n(\hat{f}_m) + \text{pen}(m)\}$$

where pen is a penalty term to be specified later (see Theorem 4.1). Finally, we define $\tilde{f} = \hat{f}_{\hat{m}}$ our estimator of the stationary density.

4.3.3 Estimation of the density F of (X_i, X_{i+1})

We proceed similarly to the estimation of f. To define the contrast to minimize, we use the following lemma:

Lemma 4.2 *For all function T, let V_T be the inverse Fourier transform of $T^*/(q^* \otimes q^*)(-.)$, i.e.*

$$V_T(x,y) = \frac{1}{4\pi^2} \iint e^{ixu+iyv} \frac{T^*(u,v)}{q^*(-u)q^*(-v)} dudv.$$

Then, for all $1 \leq k \leq n$,

1. $\mathbb{E}[V_T(Y_k, Y_{k+1})|X_1, ..., X_{n+1}] = T(X_k, X_{k+1})$
2. $\mathbb{E}[V_T(Y_k, Y_{k+1})] = \mathbb{E}[T(X_k, X_{k+1})]$

We can now adapt contrast (4.3) to the bivariate case. For any function T in $L^2(\mathbb{R}^2)$, we define the contrast

$$\Gamma_n(T) = \frac{1}{n} \sum_{i=1}^{n} [\|T\|^2 - 2V_T(X_i, X_{i+1})]$$

whose expectation is equal to $\|T\|^2 - 2/n \sum_{k=1}^{n} \mathbb{E}[T(X_k, X_{k+1})] = \|T - F\|^2 - \|F\|^2$. As previously, we can define an estimator by minimizing the contrast function.

$$\hat{F}_m = \arg\min_{T\in\mathbb{S}_m} \Gamma_n(T) \tag{4.7}$$

By differentiating Γ_n, we obtain

$$\hat{F}_m(x,y) = \sum_{j,k} \hat{A}_{j,k} \varphi_{m,j}(x) \varphi_{m,k}(y) \text{ with } \hat{A}_{j,k} = \frac{1}{n} \sum_{i=1}^{n} V_{\varphi_{m,j}\otimes\varphi_{m,k}}(Y_i, Y_{i+1}).$$

We choose again not to truncate the estimator for the sake of simplicity.

We have defined a collection of estimators $\{\hat{F}_m\}_{m\in\mathbb{M}_n}$ where we set

$$\mathbb{M}_n = \left\{ m \geq 1, \quad \frac{\Delta^2(m)}{n} \leq 1 \right\},$$

with $\Delta(m)$ defined by (4.6). Indeed, as $V_{t\otimes s}(x,y) = v_t(x)v_s(y)$, the variance of the estimator $\hat{F}_m$ is now of order $\Delta^2(m)/n$ (see (4.5)). To define an adaptive estimator we have to select the best model m. So let

$$\hat{M} = \arg\min_{m\in\mathbb{M}_n} \{\Gamma_n(\hat{F}_m) + \mathrm{Pen}(m)\}$$

where Pen is a penalty function which is specified in Theorem 4.2. Finally, we consider the estimator $\tilde{F} = \hat{F}_{\hat{M}}$.

4.3.4 Estimation of Π

Whereas the estimation of f and F is valid on the whole real line $\mathbb{R}$ or $\mathbb{R}^2$, we estimate Π on a compact set $A = A_1 \times A_2$ only, because we need a lower bound on the stationary density. More precisely, we need to set some additional assumptions:

H4 There exists a positive real f_0 such that $\forall x \in A_1, \quad f(x) \geq f_0$

H3b $\forall (x,y) \in A, \qquad \Pi(x,y) \leq \|\Pi\|_{A,\infty} < \infty$

Now, since $\Pi(x,y) = F(x,y)/f(x)$ we set

$$\tilde{\Pi}(x,y) = \begin{cases} \dfrac{\tilde{F}(x,y)}{\tilde{f}(x)} & \text{if } |\tilde{F}(x,y)| \leq n|\tilde{f}(x)|, \\ 0 & \text{otherwise.} \end{cases} \tag{4.8}$$

Here, the truncation allows to avoid the too small values of $\tilde{f}$ in the quotient. Now we evaluate upper bounds for the risk of our estimators.

4.4 Results

Our first theorem regards the problem of deconvolution. This result may be put together with results of Comte *et al.* (2006b) in the i.i.d. case and of Comte *et al.* (2006a) in various mixing frameworks.

Theorem 4.1 *Under Assumptions H1-H2b-H3c-H5, consider the estimator* $\tilde{f} = \hat{f}_{\hat{m}}$ *where for each* m, $\hat{f}_m$ *is defined by* (4.4) *and* $\hat{m} = \arg\min_{m \in \mathcal{M}_n} \{\gamma_n(\hat{f}_m) + \mathrm{pen}(m)\}$ *with*

$$\mathrm{pen}(m) = k \frac{(\pi m)^{[s-(1-s)_+/2]_+} \Delta(m)}{n}$$

where k *is a constant depending only on* k_0, k_1, b, γ, s. *Then there exists* $C > 0$ *such that*

$$\mathbb{E}\|\tilde{f} - f\|^2 \leq 4 \inf_{m \in \mathcal{M}_n} \{\|f_m - f\|^2 + \mathrm{pen}(m)\} + \frac{C}{n}$$

where f_m *is the orthogonal projection of* f *on* S_m.

The penalty is close to the variance order. It implies that the obtained rates of convergence are minimax in most cases. More precisely, the rates are given in the following corollary where $\lceil x \rceil$ denotes the ceiling function, i.e. the smallest integer larger than or equal to x.

Corollary 4.1 *Under Assumptions of Theorem 4.1, if* f *belongs to* $\mathcal{A}_{\delta,r,a}(l)$, *then*

- *If* $r = 0$ *and* $s = 0$ $\quad \mathbb{E}\|\tilde{f} - f\|^2 \leq C n^{-\frac{2\delta}{2\delta+2\gamma+1}}$
- *If* $r = 0$ *and* $s > 0$ $\quad \mathbb{E}\|\tilde{f} - f\|^2 \leq C (\log n)^{-2\delta/s}$
- *If* $r > 0$ *and* $s = 0$ $\quad \mathbb{E}\|\tilde{f} - f\|^2 \leq C \frac{(\log n)^{(2\gamma+1)/r}}{n}$
- *If* $r > 0$ *and* $s > 0$
  - *if* $r < s$ *and* $k = \lceil (s/r - 1)^{-1} \rceil - 1$, *there exist reals* b_i *such that*

$$\mathbb{E}\|\tilde{f} - f\|^2 \leq C (\log n)^{-2\delta/s} \exp[\sum_{i=0}^{k} b_i (\log n)^{(i+1)r/s - i}]$$

  - *if* $r = s$, *if* $\xi = [2\delta b + (s - 2\gamma - 1 - [s - (1-s)_+/2]_+)a]/[(a+b)s]$

$$\mathbb{E}\|\tilde{f} - f\|^2 \leq C n^{-a/(a+b)} (\log n)^{-\xi}$$

  - *if* $r > s$ *and* $k = \lceil (r/s - 1)^{-1} \rceil - 1$, *there exist reals* d_i *such that*

$$\mathbb{E}\|\tilde{f} - f\|^2 \leq C \frac{(\log n)^{(1+2\gamma-s+[s-(1-s)_+/2]_+)/r}}{n} \exp[-\sum_{i=0}^{k} d_i (\log n)^{(i+1)s/r - i}]$$

These rates are the same as those obtained in the case of i.i.d. variables X_i; they are studied in detail in Comte *et al.* (2006b). In this case, the rates $n^{-\frac{2\delta}{2\delta+2\gamma+1}}$ ($r=s=0$), $(\log n)^{-2\delta/s}$ ($r=0$, $s>0$) and $(\log n)^{(2\gamma+1)/r}/n$ ($s=0$, $r>0$) are proved to be optimal by Fan (1991) (first two cases) and Butucea (2004) (third case) for i.i.d. variables. If $r>0$ and $s>0$, we find the original rates obtained in the third chapter, proved as being optimal for $0<r<s$ in Butucea and Tsybakov (2007). In the other cases, we can compare the results of Theorem 4.1 to the one obtained with a nonadaptive estimator. There is a loss only in the case $r \geq s > 1/3$ where a logarithmic term is added. But in this case, the rates are faster than any power of logarithm.

Now let us study the risk for our estimator of the joint density F.

Theorem 4.2 *Under Assumptions H1-H2b-H3d-H5, consider the estimator* $\tilde{F} = \hat{F}_{\hat{M}}$ *where for each* m, $\hat{F}_m$ *is defined by* (4.7) *and* $\hat{M} = \arg\min_{m\in\mathbb{M}_n}\{\Gamma_n(\hat{F}_m) + \mathrm{Pen}(m)\}$ *with*

$$\mathrm{Pen}(m) = K\frac{(\pi m)^{[s-(1-s)_+]_+}\Delta^2(m)}{n}$$

where K *is a constant depending only on* k_0, k_1, b, γ, s. *Then there exists* $C>0$ *such that*

$$\mathbb{E}\|\tilde{F}-F\|^2 \leq 4\inf_{m\in\mathbb{M}_n}\{\|F_m-F\|^2 + \mathrm{Pen}(m)\} + \frac{C}{n}$$

where F_m *is the orthogonal projection of* F *on* $\mathbb{S}_m$.

The bases derived from the sine cardinal function are adapted to the estimation on the whole real line. The proof of Theorem 4.2 actually contains the proof of another result (see Proposition 4.2 in Section 4.7): the estimation of a bivariate density in a mixing framework on $\mathbb{R}^2$ and not only on a compact set. In this case of the absence of noise ($\varepsilon=0$), we obtain the same result with the penalty $\mathrm{Pen}(m) = K_0(\sum_k \beta_{2k})m^2/n$. This limit case gives the mixing coefficients back in the penalty, as it always appears in this kind of estimation (see e.g. Tribouley and Viennet (1998)).

It is then significant that in the presence of noise the penalty contains neither any mixing term nor any unknown quantity. It is entirely computable since it only depends on the characteristic function q^* of the noise which is known.

Theorem 4.2 enables us to give rates of convergence for the estimation of F.

Corollary 4.2 *Under Assumptions of Theorem 4.2, if* F *belongs to* $\mathbb{A}_{\Delta,R,A}(L)$, *then*

- *If* $R=0$ *and* $s=0$ $\qquad \mathbb{E}\|\tilde{F}-F\|^2 \leq Cn^{-\frac{2\Delta}{2\Delta+4\gamma+2}}$
- *If* $R=0$ *and* $s>0$ $\qquad \mathbb{E}\|\tilde{F}-F\|^2 \leq C(\log n)^{-2\Delta/s}$
- *If* $R>0$ *and* $s=0$ $\qquad \mathbb{E}\|\tilde{F}-F\|^2 \leq C\frac{(\log n)^{(4\gamma+2)/R}}{n}$

- *If* $R > 0$ *and* $s > 0$
  - *if* $R < s$ *and* $k = \lceil (s/R - 1)^{-1} \rceil - 1$, *there exist reals* b_i *such that*

$$\mathbb{E}\|\tilde{F} - F\|^2 \leq C(\log n)^{-2\Delta/s} \exp[\sum_{i=0}^{k} b_i (\log n)^{(i+1)R/s - i}]$$

  - *if* $R = s$ *if* $\xi = [4\Delta b + (2s - 4\gamma - 2 - [s - (1-s)_+]_+)A]/[(A+2b)s]$

$$\mathbb{E}\|\tilde{F} - F\|^2 \leq C n^{-A/(A+2b)} (\log n)^{-\xi}$$

  - *if* $R > s$ *and* $k = \lceil (R/s - 1)^{-1} \rceil - 1$, *there exist reals* d_i *such that*

$$\mathbb{E}\|\tilde{F} - F\|^2 \leq C \frac{(\log n)^{(2+4\gamma-2s+[s-(1-s)_+]_+)/R}}{n} \exp[-\sum_{i=0}^{k} d_i (\log n)^{(i+1)s/R - i}]$$

The rates of convergence look like the one of Corollary 4.1 with modifications due to the bivariate nature of F. We can compare this result to the one of Clémençon (2003) who studies only the case where $R = 0$ and $s = 0$. He shows that the minimax lower bound in that case is $n^{-\frac{2\Delta}{2\Delta+4\gamma+2}}$, so our procedure is optimal, whereas his estimator has a logarithmic loss for the upper bound. We remark that if $s > 0$ (supersmooth noise), the rate is logarithmic for F belonging to a classic ordinary smooth space. But if F is also supersmooth, better rates are recovered.

Except in the case where $R = 0$ and $s = 0$, there is, to our knowledge, no lower bound available for this estimation. We can, however, evaluate the performance of this estimator by comparing it with a nonadaptive estimator. If the smoothness of F is known, a value of m depending on R and Δ which minimizes the risk $\|F - F_m\|^2 + \Delta(m)^2/n$ can be exhibited and then some rates of convergence for this nonadaptive estimator are obtained. As soon as $s \leq 1/2$ (i.e. $[s-(1-s)_+]_+ = 0$), the penalty is $\Delta(m)^2/n$ and then the adaptive estimator recovers the same rates of convergence as those of a nonadaptive estimator if the regularity of F were known. It automatically minimizes the risk without prior knowledge on the regularity of F and there is no loss in the rates. If $s > 1/2$ a loss can appear but is not systematic. If $R < s$, the rate of convergence is unchanged since the bias dominates. It is only when $R \geq s > 1/2$ that an additional logarithmic term appears. But in this case the risk decreases faster than any logarithmic function so that the loss is negligible.

We can now state the main result regarding the estimation of the transition density Π.

Theorem 4.3 *Under Assumptions H1-H2b-H3b,c,d-H4-H5, consider the estimator* $\tilde{\Pi}$ *defined in* (4.8). *We assume that* f *belongs to* $\mathcal{A}_{\delta,r,a}(l)$ *and that we browse only the models* $m \in \mathcal{M}_n$ *such that*

$$m \geq \log\log n \quad \text{and} \quad m\Delta(m) \leq \frac{n}{(\log n)^2} \tag{4.9}$$

to define $\tilde{f}$. *Then* $\tilde{\Pi}$ *verifies, for* n *large enough,*

$$\mathbb{E}\|\tilde{\Pi} - \Pi\|_A^2 \leq C_1 \mathbb{E}\|\tilde{F} - F\|^2 + C_2 \mathbb{E}\|\tilde{f} - f\|^2 + \frac{C}{n}$$

where $\|T\|_A^2 = \iint_A T^2(x,y)dxdy$.

Note that, unlike in Theorems 4.1 and 4.2, this result is asymptotic. It states that the rate of convergence for Π is no larger than the maximum of the rates of f and F. The restrictions (4.9) do not modify the conclusion of Theorem 4.1 and the resulting rates of convergence. Thus if f and F have the same regularity, the rates of convergence for Π are those of F, given in Corollary 4.2.

If $s = 0$ i.e. if ε_i is ordinary smooth, then the rates of convergence are polynomial; moreover they are near the parametric rate $1/n$ if R and r are positive. In the other hand the smoother the error distribution, the harder the estimation. In the case of a supersmooth noise, the rates are logarithmic if f or F is ordinary smooth but faster than any power of logarithm if the hidden chain has supersmooth densities. The exact rates depend on all regularities γ, s, δ, r, Δ, R and are very tedious to write. That is why we prefer to give some detailed examples.

4.5 Examples

In this section, we give some examples to illustrate the previous results. In nonparametric examples, the quantities that allow to compute the rates of convergence, i.e. the regularities of the densities, remain unknown. It is besides an advantage of the procedure, not to need such information to reach good rates.

So the following models are parametric, but it is well-known that in the case where the state spaces of the hidden chains are not finite, nor bounded, classical parametric estimation is not proved to perform well.

4.5.1 Autoregressive process of order 1

Let us study the case where the Markov chain is defined by

$$X_{n+1} = \alpha X_n + \beta + \eta_{n+1}$$

where the η_n's are i.i.d. centered Gaussian with variance σ^2. This chain is irreducible, Harris recurrent and geometrically β-mixing. The stationary distribution is Gaussian with mean $\beta/(1-\alpha)$ and variance $\sigma^2/(1-\alpha^2)$. So

$$f^*(u) = \exp\left[-iu\left(\frac{\beta}{1-\alpha}\right) - \frac{\sigma^2}{2(1-\alpha^2)}u^2\right]$$

and then bias computing gives $\delta = 1/2$, $r = 2$. The function F is the density of a Gaussian vector with mean $(\beta/(1-\alpha), \beta/(1-\alpha))$ and variance matrix $\sigma^2/(1-\alpha^2)\begin{pmatrix}1 & \alpha\\ \alpha & 1\end{pmatrix}$. So

$$F^*(u,v) = \exp\left[-i(u+v)\left(\frac{\beta}{1-\alpha}\right) - \frac{\sigma^2}{2(1-\alpha^2)}(u^2+v^2+2\alpha uv)\right]$$

and $\Delta = 1/2$, $R = 2$.

We can compute the rates of convergence for different kinds of noise ε. If ε has a Laplace distribution, $q^*(u) = 1/(1+u^2)$ so $s = 0$, $\gamma = 2$. In this case, Corollary 4.1 gives $\mathbb{E}\|\tilde{f} - f\|^2 \leq C(\log n)^{5/2}/n$ and $\mathbb{E}\|\tilde{F} - F\|^2 \leq C(\log n)^5/n$. Consequently,

$$\mathbb{E}\|\tilde{\Pi} - \Pi\|_A^2 \leq C\frac{(\log n)^5}{n}$$

with $A_1 = A_2$ an interval $[-d, d]$. This rate is close to the parametric rate $1/n$; it is due to the great smoothness of the chain compared with that of error.

If now ε has a normal distribution with variance τ^2, then we compute

$$\mathbb{E}\|\tilde{\Pi} - \Pi\|_A^2 \leq Cn^{-\frac{\sigma^2}{\sigma^2+2\tau^2}}(\log n)^{-\frac{\tau^2}{\sigma^2+2\tau^2}}.$$

4.5.2 Cox-Ingersoll-Ross process

Another example is given by $X_n = R_{n\tau}$ with τ a fixed sampling interval and R_t the so-called Cox-Ingersoll-Ross process defined by

$$dR_t = (2\theta R_t + \kappa\sigma_0^2)dt + 2\sigma_0\sqrt{R_t}dW_t \qquad \theta < 0, \kappa \in \{2, 3, ...\}.$$

Calculation of f and its regularity

Since $\theta > 0$ and $\kappa \geq 2$, R_t is positive recurrent on $]0, \infty[$. Following Chaleyat-Maurel and Genon-Catalot (2006) Appendix A.3, the stationary distribution is a gamma distribution with parameters $\kappa/2$ and $1/(2\rho^2)$ with

$$\rho = \frac{\sigma_0}{\sqrt{2|\theta|}}.$$

Then the stationary density can be written

$$f(x) = \frac{1}{(2\rho^2)^{\kappa/2}\Gamma(\kappa/2)}x^{\kappa/2-1}e^{-x/(2\rho^2)}\mathbb{1}_{x>0}$$

and the characteristic function

$$f^*(u) = \mathbb{E}[e^{-iuX_1}] = (1+2\rho^2 iu)^{-\kappa/2} = \left(1 + iu\frac{\sigma_0^2}{|\theta|}\right)^{-\kappa/2}.$$

Now we compute the bias to determine the regularity of f. Given that $|f^*(u)|^2 = (1 + 4\rho^4 u^2)^{-\kappa/2}$, we obtain

$$\int_{|u|>\pi m} |f^*(u)|^2 = 2\int_{\pi m}^{\infty} (1+4\rho^4u^2)^{-\kappa/2} du \leq 2\int_{\pi m}^{\infty} (4\rho^4u^2)^{-\kappa/2} du$$
$$\leq \frac{2}{(2\rho^2)^\kappa}\left[\frac{u^{-\kappa+1}}{-\kappa+1}\right]_{\pi m}^{\infty} \leq \frac{2}{(2\rho^2)^\kappa}\frac{1}{1-\kappa}(\pi m)^{1-\kappa}.$$

Therefore, $r = 0$ and $1 - \kappa = -2\delta$, so

$$r = 0, \qquad \delta = \frac{\kappa - 1}{2} \tag{4.10}$$

Calculation of F^* and its regularity

It is shown in Chaleyat-Maurel and Genon-Catalot (2006) that R_t is the square of the Euclidean norm of a κ-dimensional vector $(\xi_t^1, \ldots, \xi_t^\kappa)$ whose components are i.i.d. Ornstein-Uhlenbeck processes verifying

$$d\xi_t^j = \theta\xi_t^j dt + \sigma_0 dW_t^j, \qquad j = 1, \ldots, \kappa.$$

Then we can write

$$X_n = (\xi_{n\tau}^1)^2 + \cdots + (\xi_{n\tau}^\kappa)^2.$$

For all $j \in \{1, \ldots, \kappa\}$, $(\xi_{n\tau}^j)$ is a discretised Ornstein-Uhlenbeck process. It verifies then[(1)]

$$\xi_{n\tau}^j = \alpha\xi_{(n-1)\tau}^j + \beta\eta_n^j$$

with (η_n^j) i.i.d. sequences of $\mathcal{N}(0,1)$ and

$$\alpha = e^{\theta\tau}, \quad \beta^2 = \sigma_0^2\frac{e^{2\theta\tau}-1}{2\theta}.$$

Knowing $\xi_{(n-1)\tau}^j = x_j$, each variable $\xi_{n\tau}^j$ follows a Gaussian distribution with mean αx_j and variance β^2. Now

$$\beta^{-2}X_n = \left(\frac{\xi_{n\tau}^1}{\beta}\right)^2 + \cdots + \left(\frac{\xi_{n\tau}^\kappa}{\beta}\right)^2.$$

The distribution of $\beta^{-2}X_n$ knowing $\xi_{(n-1)\tau}^j = x_j$ is then a noncentral chi-square with mean $\beta^{-2}\sum_{j=1}^\kappa(\alpha x_j)^2$. Thus , conditionally to $X_{n-1} = \sum_{j=1}^\kappa(\xi_{(n-1)\tau}^j)^2 = x$,

$$\beta^{-2}X_n \sim \chi'^2\left(\frac{\alpha^2 x}{\beta^2}, \kappa\right).$$

[(1)]It involves that $\xi_{n\tau}^j$ admits a centered Gaussian with variance $\beta^2/(1-\alpha^2) = \rho^2$ for stationary distribution. Then $\rho^{-2}X_n$ is a chi-square $\chi^2(\kappa)$ and we find again $X_n \sim \Gamma(\kappa/2, 1/(2\rho^2))$.

We use Appendix A.2 in Chaleyat-Maurel and Genon-Catalot (2006) (equation (71)) to write

$$\mathbb{E}[e^{-ivX_n}|X_{n-1}=x] = (1+iv\beta^2)^{-\kappa/2}\exp\left(-\frac{iv\alpha^2 x}{1+2iv\beta^2}\right).$$

This formula allows to compute F^*. Indeed $F^*(u,v) = \mathbb{E}[e^{-iuX_{n-1}-ivX_n}]$ and then

$$F^*(u,v) = \int \mathbb{E}[e^{-ivX_n}|X_{n-1}=x]e^{-iux}f(x)dx.$$

We explicitly know all the terms of the right member, so we can do the computation.

$$\begin{aligned} F^*(u,v) &= \int_0^\infty (1+iv\beta^2)^{-\kappa/2}\exp\left(-\frac{iv\alpha^2 x}{1+2iv\beta^2}\right)e^{-iux}\frac{x^{\kappa/2-1}e^{-x/(2\rho^2)}}{(2\rho^2)^{\kappa/2}\Gamma(\kappa/2)}dx \\ &= \frac{(1+iv\beta^2)^{-\kappa/2}}{(2\rho^2)^{\kappa/2}\Gamma(\kappa/2)}\int_0^\infty x^{\kappa/2-1}\exp[-A(u,v)x]dx \end{aligned}$$

with

$$A(u,v) = \frac{iv\alpha^2}{1+2iv\beta^2} + iu + \frac{1}{2\rho^2}.$$

Then we obtain

$$F^*(u,v) = \frac{(1+iv\beta^2)^{-\kappa/2}}{(2\rho^2)^{\kappa/2}\Gamma(\kappa/2)}\Gamma(\kappa/2)A(u,v)^{-\kappa/2} = [2\rho^2(1+iv\beta^2)A(u,v)]^{-\kappa/2}.$$

Now we rewrite $A(u,v)$:

$$\begin{aligned} A(u,v) &= \frac{2\rho^2 iv\alpha^2 + 2\rho^2(1+2iv\beta^2)iu + 1 + 2iv\beta^2}{2\rho^2(1+2iv\beta^2)} \\ &= \frac{1-4\rho^2\beta^2uv + 2\rho^2 iu + 2(\alpha^2\rho^2+\beta^2)iv}{2\rho^2(1+2iv\beta^2)}. \end{aligned}$$

But $\alpha^2\rho^2+\beta^2 = \alpha^2\rho^2 - \rho^2(\alpha^2-1) = \rho^2$ and then

$$A(u,v)2\rho^2(1+2iv\beta^2) = 1-4\rho^2\beta^2uv + 2\rho^2 i(u+v).$$

Then we obtain

$$F^*(u,v) = [1-4\rho^2\beta^2uv + 2\rho^2 i(u+v)]^{-\kappa/2}.$$

Now we compute the bias to find R and Δ. First

$$|F^*(u,v)|^2 = [(1-4\rho^2\beta^2uv)^2 + 4\rho^4(u+v)^2]^{-\kappa/2}.$$

This expression being symmetric in u and v, it is sufficient to evaluate $\int_{v\in\mathbb{R}}\int_{|u|>\pi m}|F^*(u,v)|^2dudv$. To do this, we write

$$|F^*(u,v)|^2 = [a(v)u^2 + b(v)u + c(v)]^{-\kappa/2}$$

$$\text{with} \left| \begin{array}{rcl} a(v) &=& 16\rho^4\beta^4v^2+4\rho^4, \\ b(v) &=& -8\rho^2\beta^2v+8\rho^4v=8\rho^4\alpha^2v, \\ c(v) &=& 1+4\rho^4v^2. \end{array} \right.$$

Then, for m large enough,

$$\begin{aligned} \int_{|u|>\pi m}|F^*(u,v)|^2du &\leq \int_{\pi m}^{\infty}(a(v)u^2+b(v)u)^{-\kappa/2}du+\int_{-\infty}^{-\pi m}(a(v)u^2+b(v)u)^{-\kappa/2}du \\ &\leq \int_{\pi m}^{\infty}a(v)^{-\kappa/2}u^{-\kappa}du+\int_{-\infty}^{-\pi m}\left(\frac{a(v)u^2}{2}\right)^{-\kappa/2}du \\ &\leq C(\kappa)a(v)^{-\kappa/2}(\pi m)^{1-\kappa}. \end{aligned}$$

Replacing $a(v)$ by its value, we find

$$\begin{aligned} \int_{v\in\mathbb{R}}\int_{|u|>\pi m}|F^*(u,v)|^2dudv &\leq C(\kappa)(\pi m)^{1-\kappa}\int(16\rho^4\beta^4v^2+4\rho^4)^{-\kappa/2}dv \\ &\leq C(\pi m)^{1-\kappa}. \end{aligned}$$

Consequently

$$R=0 \text{ and } \Delta=\frac{\kappa-1}{2}. \tag{4.11}$$

Calculation of the rates of convergence

If (ε_i) is Gaussian, then $s=2$ and $\gamma=0$. We use Corollary 4.1: it is the case $(r=0,s>0)$ where the rate is $n^{-2\delta/s}$. So, using (4.10), $\mathbb{E}\|f-\tilde{f}\|^2\leq Cn^{(1-\kappa)/2}$. In the same way, using Corollary 4.2 and (4.11), $\mathbb{E}\|F-\tilde{F}\|^2\leq Cn^{(1-\kappa)/2}$. Finally

$$\mathbb{E}\|\Pi-\tilde{\Pi}\|_A^2\leq Cn^{(1-\kappa)/2}.$$

In the case of a noise with distribution $\Gamma(\alpha,\lambda)$, the smoothness coefficients of the noise are $\gamma=\alpha$ and $s=0$. It is the case $(r=0,s=0)$ in Corollary 4.1:

$$\mathbb{E}\|f-\tilde{f}\|^2\leq Cn^{-\frac{2\delta}{2\delta+2\gamma+1}}\leq Cn^{-\frac{\kappa-1}{\kappa+2\alpha}}.$$

Corollary 4.2 gives

$$\mathbb{E}\|F-\tilde{F}\|^2\leq Cn^{-\frac{2\Delta}{2\Delta+4\gamma+2}}\leq Cn^{-\frac{\kappa-1}{\kappa+4\alpha+1}}$$

and then

$$\mathbb{E}\|\Pi-\tilde{\Pi}\|_A^2\leq Cn^{-\frac{\kappa-1}{\kappa+4\alpha+1}}.$$

4.5.3 Stochastic volatility model

Our work allows to study some multiplicative models as the so-called stochastic volatility model in finance (see Genon-Catalot *et al.* (2000) for the links between the standard continuous-time SV models and the hidden Markov models). Let us consider

$$Z_n = U_n^{1/2}\eta_n$$

where (U_n) is a nonnegative Markov chain, (η_n) a sequence of i.i.d. standard Gaussian variables, the two sequences being independent. Setting $X_n = \log(U_n)$ and $\varepsilon_n = \log(\eta_n^2)$ leads us back to our initial problem.

The noise distribution is the logarithm of a chi-square distribution and then verifies $q^*(x) = 2^{-ix}\Gamma(1/2 - ix)/\sqrt{\pi}$. Van Es *et al.* (2005) show that $|q^*(x)| \sim_{+\infty} \sqrt{2}e^{-\pi|x|/2}$ and then $s = 1, \gamma = 0$.

In the general case, the logarithm of the hidden chain X_n derives from a regular sampling of a diffusion process with unknown drift and diffusion coefficients. Then the rate of convergence for the estimation of the transition depends on the smoothness of f and F. If $R = r = 0$, then $\mathbb{E}\|\tilde{\Pi} - \Pi\|_A^2 \leq C(\log n)^{-2\delta}$. But if r and R are positive, better rates are recovered.

For example, we assume that the logarithm of the hidden chain X_n derives from a regular sampling of an Ornstein-Uhlenbeck process, i.e. $X_n = V_{n\tau}$ where V_t is defined by the equation

$$dV_t = \theta V_t dt + \sigma dB_t$$

with B_t a standard Brownian motion. Then all the assumptions are satisfied. Similarly to Section 4.5.1, the stationary distribution is Gaussian with mean 0 and variance $\sigma^2/2|\theta|$ and then $\delta = 1/2$, $r = 2$. In the same way F is the density of a centered Gaussian vector with variance matrix $\sigma^2/(2|\theta|)\begin{pmatrix} 1 & e^{\theta\tau} \\ e^{\theta\tau} & 1 \end{pmatrix}$ and then $\Delta = 1/2$, $R = 2$. We obtain the following rate of convergence on some interval $A_1 = A_2 = [-d, d]$

$$\mathbb{E}\|\tilde{\Pi} - \Pi\|_A^2 \leq C\sqrt{\log n}\frac{\exp[(\pi/\beta)\sqrt{\log n}]}{n}$$

with $\beta^2 = \sigma^2(e^{2\theta\tau} - 1)/(2\theta)$.

4.6 Simulations

For each m, the estimator $\hat{f}_m$ of f can be written

$$\hat{f}_m = \sum_{|j|\leq K_n} \hat{a}_j^m \varphi_{m,j} \quad \text{with} \quad \hat{a}_j^m = \frac{1}{n}\sum_{i=1}^n v_{\varphi_{m,j}}(Y_i).$$

To compute the coefficients $\hat{a}_j^m$, we will use the Inverse Fast Fourier Transform. Indeed, using Lemma 4.3

$$\begin{aligned}\hat{a}_j^m &= \frac{1}{n}\sum_{k=1}^{n}\frac{\sqrt{m}}{2\pi}\int_{-\pi}^{\pi}e^{-ijv}e^{iY_kvm}[q^*(-vm)]^{-1}dv \\ &= \frac{\sqrt{m}}{2}\int_{-1}^{1}e^{ijx\pi}(\frac{1}{n}\sum_{k=1}^{n}e^{-iY_kx\pi m})[q^*(x\pi m)]^{-1}dx \\ &= \frac{\sqrt{m}}{2}(-1)^j\int_0^2 e^{ijx\pi}h_m(x)dx\end{aligned}$$

where

$$h_m(x) = \frac{\frac{1}{n}\sum_{k=1}^{n}e^{-iY_k\pi m(x-1)}}{q^*(\pi m(x-1))}.$$

So $\hat{a}_j^m$ can be approximated by

$$\sqrt{m}(-1)^j\frac{1}{N}\sum_{k=0}^{N-1}e^{ij\pi\frac{2k}{N}}h_m(\frac{2k}{N})$$

Then, if H is the vector $(h_m(0), h_m(2/N), \ldots, h_m(2(N-1)/N)$, we can write

$$\hat{a}_j^m \approx \sqrt{m}(-1)^j(IFFT(H))_j$$

for $j = 0, ..., N-1$. For $j < 0$, it is sufficient to replace $h_m(x)$ by $h_m(-x) = \overline{h_m(x)}$, i.e. H by $\overline{H}$. Following Comte *et al.* (2006b), we choose $K_n = N - 1 = 2^8 - 1$.

To compute $\tilde{f}$, we use then the following steps:

- For each m and for each j, compute $\hat{a}_j^m$.
- For each m compute $\gamma_n(\hat{f}_m) + \text{pen}(m) = -\sum_j |\hat{a}_j^m|^2 + \text{pen}(m)$.
- Select the $\hat{m}$ which minimize $\gamma_n(\hat{f}_m) + \text{pen}(m)$.
- Compute $\tilde{f} = \sum_{|j|\leq K_n}\hat{a}_j^{\hat{m}}\varphi_{\hat{m},j}$.

To compute the estimator of F, the procedure is identical:

- For each m and for each j, k, compute $\hat{A}_{j,k}^m$ (using a multivariate IFFT).
- For each m compute $\Gamma_n(\hat{F}_m) + \text{Pen}(m) = -\sum_{j,k}|\hat{A}_{j,k}^m|^2 + \text{Pen}(m)$.
- Select the $\hat{M}$ which minimize $\Gamma_n(\hat{F}_m) + \text{Pen}(m)$.
- Compute $\tilde{F} = \sum_{|j|\leq K_n,|k|\leq K_n}\hat{A}_{j,k}^{\hat{M}}\varphi_{\hat{M},j}\otimes\varphi_{\hat{M},k}$.

Finally, we compute

$$\tilde{\Pi}(x,y) = \begin{cases} \dfrac{\tilde{F}(x,y)}{\tilde{f}(x)} & \text{if } |\tilde{F}(x,y)| \leq 0.01n|\tilde{f}(x)|, \\ 0 & \text{otherwise.} \end{cases}$$

The processes are those described in the section Simulations of Chapter 1. We consider two different noises:

Laplace noise In this case, the density of ε_i is given by

$$q(x) = \frac{\lambda}{2}e^{-\lambda|x|}; \quad q^*(x) = \frac{\lambda^2}{\lambda^2 + x^2}; \quad \lambda = 5.$$

The smoothness parameters are $\gamma = 2$ and $b = s = 0$. We compute

$$\Delta(m) = \frac{m}{2}\int_{-1}^{1} |q^*(v\pi m)|^{-2} dv = \frac{1}{\pi}\left[\pi m + \frac{2}{3\lambda^2}(\pi m)^3 + \frac{1}{5\lambda^4}(\pi m)^5\right].$$

Since $\text{pen}(m)$ must be larger than $k\Delta(m)/n$, we choose

$$\text{pen}(m) = \frac{1}{2n}\left[\pi m + \frac{2}{3\lambda^2}(\pi m)^3 + \frac{1}{5\lambda^4}(\pi m)^5\right].$$

For the estimation of F, we use

$$\text{Pen}(m) = \frac{1}{2n}\left[\pi m + \frac{2}{3\lambda^2}(\pi m)^3 + \frac{1}{5\lambda^4}(\pi m)^5\right]^2.$$

Gaussian noise In this case, the density of ε_i is given by

$$q(x) = \frac{1}{\lambda\sqrt{2\pi}}e^{-\frac{x^2}{2\lambda^2}}; \quad q^*(x) = e^{-\frac{\lambda^2 x^2}{2}}; \quad \lambda = 0.3.$$

So $\gamma = 0, b = \lambda^2/2$ and $s = 2$. The penalties have to verify $\text{pen}(m) \geq k(\pi m)^2\Delta(m)/n$ and $\text{Pen}(m) \geq K(\pi m)^2\Delta(m)^2/n$ with

$$\Delta(m) = m\int_0^1 e^{(\lambda\pi m x)^2} dx.$$

So we choose

$$\begin{aligned} \text{pen}(m) &= \frac{1}{2}(\pi m)^3 \int_0^1 e^{(\lambda\pi m x)^2} dx, \\ \text{Pen}(m) &= \frac{1}{2}(\pi m)^4 (\int_0^1 e^{(\lambda\pi m x)^2} dx)^2. \end{aligned}$$

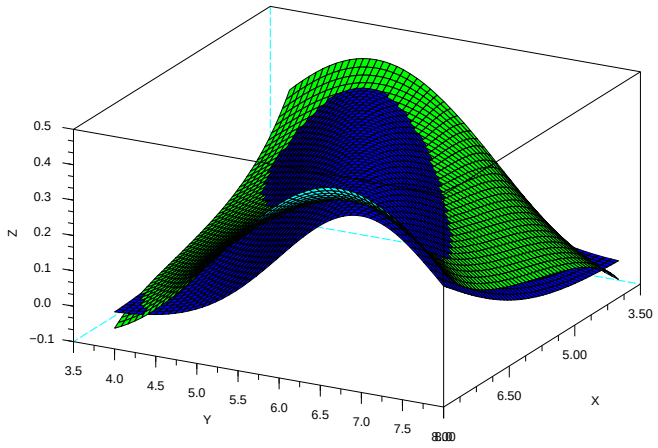


Figure 4.1: Estimator (light surface) and true transition (dark surface) for the process AR(i) observed with a Gaussian noise, $n = 500$.

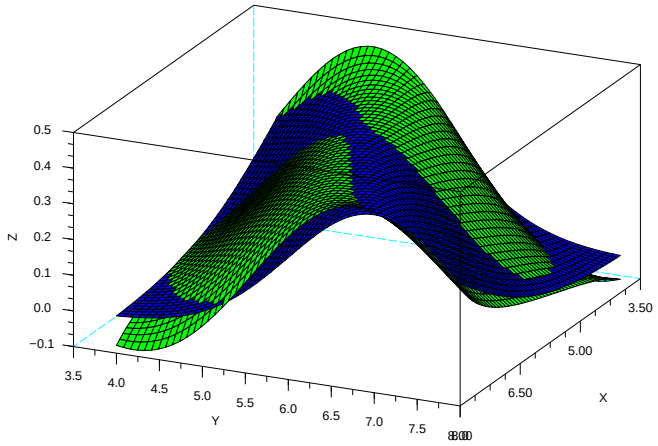


Figure 4.2: Estimator (light surface) and true transition (dark surface) for the process AR(i) observed with a Laplace noise, $n = 500$.

As in Comte *et al.* (2006b), we consider that m can be fractional. More precisely, we

take $\pi m = k/2$ with $k \in \mathbb{N}^*$. Figures 4.1, 4.2 illustrate the result for the process AR(i).

We can also observe on Figure 4.3 the estimator of the transition density of the process ARCH when the noise is Gaussian. Figure 4.4 shows sections of the previous surfaces. We can see the curves $z = \Pi(-3.2, y)$ versus $z = \tilde{\Pi}(-3.2, y)$ and the curves $z = \Pi(x, -4.2)$ versus $z = \tilde{\Pi}(x, -4.2)$. More precise results are presented in Table 4.1.

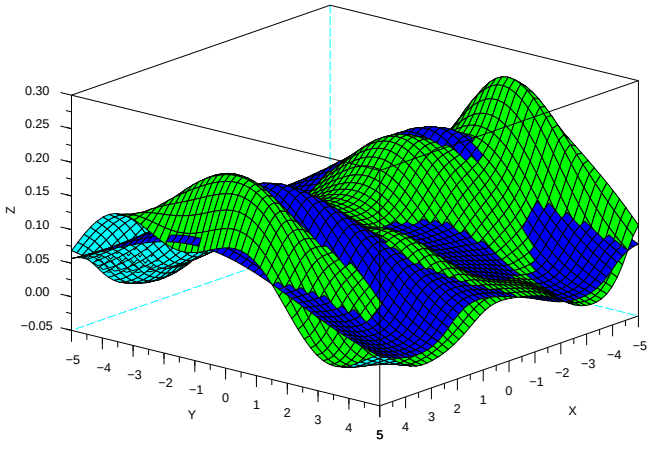
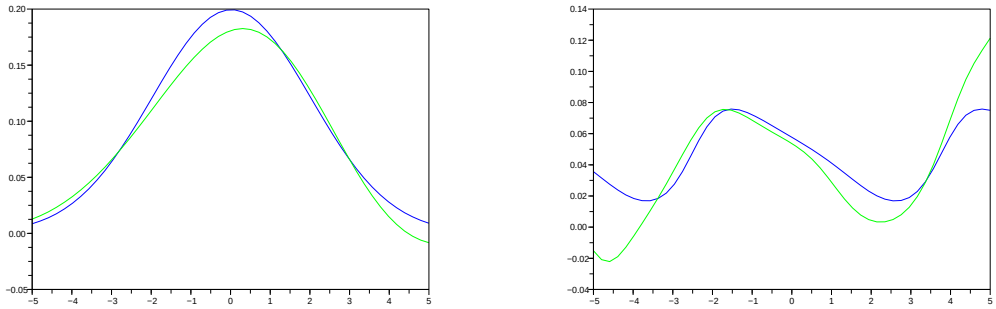


Figure 4.3: Estimator (light surface) and true transition (dark surface) for the process ARCH observed with a Gaussian noise, $n = 500$.



$x = -3.2$ $y = -4.2$

Figure 4.4: Sections for ARCH process, $n = 500$, dark line: true function, light line: estimator.

n	100	250	500	1000	noise
AR(i)	0.305	0.139	0.047	0.031	Lapl
	0.066	0.047	0.021	0.009	Gauss
AR(ii)	0.598	0.174	0.049	0.031	Lapl
	0.065	0.064	0.044	0.017	Gauss
$\sqrt{\text{CIR}}$	0.821	0.443	0.120	0.057	Lapl
	0.060	0.021	0.011	0.007	Gauss
CIR(iii)	0.545	0.453	0.185	0.104	Lapl
	0.853	0.793	0.776	0.771	Gauss
CIR(iv)	0.379	0.143	0.104	0.079	Lapl
	0.410	0.225	0.122	0.115	Gauss
ARCH	0.283	0.178	0.098	0.065	Lapl
	0.334	0.239	0.100	0.047	Gauss

Table 4.1: MISE $\mathbb{E}\|\Pi - \tilde{\Pi}\|^2$ averaged over $N = 100$ samples.

We remark in Table 4.1 that the number of observations must be great enough to have good results, which is usual in deconvolution problems. But, when n is large, the results are very satisfactory, sometimes better than in the case without noise. We can explain this fact by the performance of the basis derived from the sine cardinal. As expected after theoretical study, the decrease of the risk is faster when the noise follows a Laplace distribution. This phenomenon is underlined on the following graph.

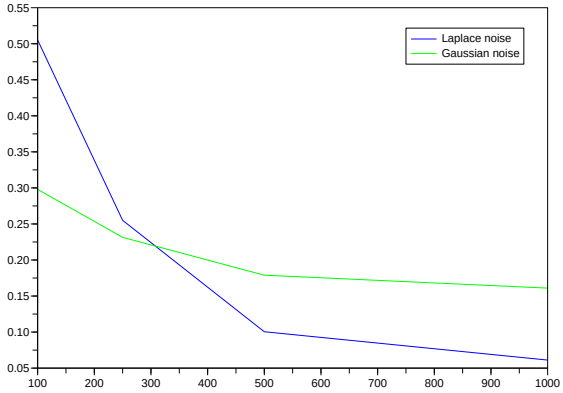


Figure 4.5: Mean of the MISE for the 6 processes when n increases.

4.7 Proofs

Here, we do not prove the results concerning the estimation of f. Indeed, they are similar to the ones concerning F (but actually simpler) and the ones of Comte *et al.* (2006b). It is then sufficient to use corresponding proofs for F *mutatis mutandis*.

For the sake of simplicity, all constants in the following are denoted by C, even if they have different values.

4.7.1 Proof of Lemma 4.2

It is sufficient to prove the first assertion. First we write that

$$V_T(Y_k, Y_{k+1}) = 1/4\pi^2 \int e^{iY_k u + iY_{k+1} v} T^*(u,v)/q^*(-u)q^*(-v)dudv$$

so that

$$\mathbb{E}[V_T(Y_k, Y_{k+1})|X_1, ..., X_{n+1}] = \frac{1}{4\pi^2} \int \mathbb{E}[e^{iY_k u + iY_{k+1} v}|X_1, ..., X_{n+1}] \frac{T^*(u,v)}{q^*(-u)q^*(-v)} dudv.$$

By using the independence between (X_i) and (ε_i), we compute

$$\begin{aligned}
&\mathbb{E}[e^{iY_k u + iY_{k+1} v}|X_1, .., X_{n+1}] = \mathbb{E}[e^{iX_k u + iX_{k+1} v} e^{i\varepsilon_k u + i\varepsilon_{k+1} v}|X_1, .., X_{n+1}] \\
&= e^{iX_k u + iX_{k+1} v} \mathbb{E}[e^{i\varepsilon_k u}]\mathbb{E}[e^{i\varepsilon_{k+1} v}] = e^{iX_k u + iX_{k+1} v} \int e^{ixu} q(x)dx \int e^{iyv} q(y)dy \\
&= e^{iX_k u + iX_{k+1} v} q^*(-u)q^*(-v).
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{E}[V_T(Y_k, Y_{k+1})|X_1, .., X_{n+1}] &= \frac{1}{4\pi^2} \int e^{iX_k u + iX_{k+1} v} q^*(-u)q^*(-v) \frac{T^*(u,v)}{q^*(-u)q^*(-v)} dudv \\
&= \frac{1}{4\pi^2} \int e^{iX_k u + iX_{k+1} v} T^*(u,v)dudv = T(X_k, X_{k+1}).
\end{aligned}$$

4.7.2 Sketch of proof of Theorem 4.2:

Let $m \in \mathcal{M}_n$. The definitions of $\hat{F}_m$ and $\hat{m}$ lead to the inequality

$$\Gamma_n(\hat{F}_{\hat{M}}) + \mathrm{Pen}(\hat{M}) \leq \Gamma_n(F_m) + \mathrm{Pen}(m). \tag{4.12}$$

Let

$$Z_n(T) = (1/n) \sum_{i=1}^{n} \left\{ V_T(Y_i, Y_{i+1}) - \int T(x,y)F(x,y)dxdy \right\}. \tag{4.13}$$

It is easy to see that

$$\Gamma_n(\hat{F}_{\hat{M}}) - \Gamma_n(F_m) = \|\hat{F}_{\hat{M}} - F\|^2 - \|F_m - F\|^2 - 2Z_n(\hat{F}_{\hat{M}} - F_m)$$

so that (4.12) becomes

$$\begin{aligned} \|\hat{F}_{\hat{M}} - F\|^2 &\leq \|F_m - F\|^2 + 2Z_n(\hat{F}_{\hat{M}} - F_m) + \mathrm{Pen}(m) - \mathrm{Pen}(\hat{M}) \\ &\leq \|F_m - F\|^2 + 2\|\hat{F}_{\hat{M}} - F_m\| \sup_{T \in B(m,\hat{M})} Z_n(T) + \mathrm{Pen}(m) - \mathrm{Pen}(\hat{M}) \end{aligned}$$

where $B(m, m') = \{T \in \mathbb{S}_m + \mathbb{S}_{m'}, \quad \|T\| = 1\}$. The main step of the proof is then to control the term $\sup_{T \in B(m,\hat{M})} Z_n(T)$,

To deal with the supremum of the empirical process $Z_n(T)$, we will use an inequality of Talagrand stated in Lemma 4.5 (Section 4.7.8). This inequality is very powerful but can be applied only to sum of independent random variables. That is why we split $Z_n(T)$ into two processes.

$$Z_n(T) = Z_{n,1}(T) + Z_{n,2}(T)$$

with

$$\begin{cases} Z_{n,1}(T) = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \{V_T(Y_i, Y_{i+1}) - \mathbb{E}[V_T(Y_i, Y_{i+1})|X_1, ..., X_{n+1}]\}, \\ Z_{n,2}(T) = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \left\{ T(X_i, X_{i+1}) - \int T(x,y)F(x,y)dxdy \right\}. \end{cases} \tag{4.14}$$

For the first process $Z_{n,1}(T)$, we return to independent variables remarking that, conditionally to $X_1, \ldots, X_{n+1}$, the variables (Y_{2i-1}, Y_{2i}) are independent (see Proposition 4.1).

For the other processes, we use the mixing assumption H2b to build auxiliary variables X_i^* which are approximation of the X_i's and which constitute independent clusters of variables (see Proposition 4.2).

4.7.3 Detailed proof of Theorem 4.2

First, we introduce some auxiliary variables whose existence is ensured by Assumption H2b of mixing. In the case of arithmetical mixing, since $\theta > 8$, there exists a real c such that $0 < c < 1/2$ and $c\theta > 4$. We set in this case $q_n = \frac{1}{2}\lfloor n^c \rfloor$. In the case of geometrical mixing, we set $q_n = \frac{1}{2}\lfloor c \log(n) \rfloor$ where c is a real larger than $4/\theta$.

For the sake of simplicity, we suppose that $n + 1 = 4p_n q_n$, with p_n an integer. Let for $i = 1, \ldots, (n+1)/2$, $V_i = (X_{2i-1}, X_{2i})$ and for $l = 0, \ldots, p_n - 1$, $A_l = (V_{2lq_n+1}, ..., V_{(2l+1)q_n})$, $B_l = (V_{(2l+1)q_n+1}, ..., V_{(2l+2)q_n})$. As in Viennet (1997), by using Berbee's coupling Lemma,

we can build a sequence (A_l^*) such that

$$\begin{cases} A_l \text{ and } A_l^* \text{ have the same distribution,} \\ A_l^* \text{ and } A_{l'}^* \text{ are independent if } l \neq l', \\ P(A_l \neq A_l^*) \leq \beta_{2q_n}. \end{cases}$$

In the same way, we build (B_l^*) and we define for any $l \in \{0, \ldots, p_n - 1\}$, $A_l^* = (V_{2lq_n+1}^*, ..., V_{(2l+1)q_n}^*)$, $B_l^* = (V_{(2l+1)q_n+1}^*, ..., V_{(2l+2)q_n}^*)$ so that the sequence $(V_1^*, \ldots, V_{n/2}^*)$ and then the sequence $(X_1^*, \ldots, X_n^*)$ are well defined. We can now define

$$\Omega^* = \{\forall i, 1 \leq i \leq n+1 \quad X_i = X_i^*\}.$$

Then we split the risk into two terms:

$$\mathbb{E}(\|\tilde{F} - F\|^2) = \mathbb{E}(\|\tilde{F} - F\|^2 \mathbb{1}_{\Omega^*}) + \mathbb{E}(\|\tilde{F} - F\|^2 \mathbb{1}_{\Omega^{*c}}).$$

To pursue the proof, we observe that for all T, T'

$$\Gamma_n(T) - \Gamma_n(T') = \|T - F\|^2 - \|T' - F\|^2 - 2Z_n(T - T')$$

where $Z_n(T)$ is defined by (4.13). Let us fix $m \in \mathbb{M}_n$ and denote by F_m the orthogonal projection of F on $\mathbb{S}_m$. Since $\Gamma_n(\tilde{F}) + \text{Pen}(\hat{M}) \leq \Gamma_n(F_m) + \text{Pen}(m)$, we have

$$\begin{aligned} \|\tilde{F} - F\|^2 &\leq \|F_m - F\|^2 + 2Z_n(\tilde{F} - F_m) + \text{Pen}(m) - \text{Pen}(\hat{M}) \\ &\leq \|F_m - F\|^2 + 2\|\tilde{F} - F_m\| \sup_{T \in B(m,\hat{M})} Z_n(T) + \text{Pen}(m) - \text{Pen}(\hat{M}) \end{aligned}$$

where, for all m, m', $B(m, m') = \{T \in \mathbb{S}_m + \mathbb{S}_{m'}, \quad \|T\| = 1\}$. Then, using inequality $2xy \leq x^2/4 + 4y^2$,

$$\|\tilde{F} - F\|^2 \leq \|F_m - F\|^2 + \frac{1}{4}\|\tilde{F} - F_m\|^2 + 4 \sup_{T \in B(m,\hat{M})} Z_n^2(T) + \text{Pen}(m) - \text{Pen}(\hat{M}). \quad (4.15)$$

Using Lemma 4.2, $Z_n(T)$ can be split into two terms :

$$Z_n(T) = Z_{n,1}(T) + Z_{n,2}(T)$$

with $Z_{n,1}(T)$ and $Z_{n,2}(T)$ defined by (4.14). Now let $P_1(.,.)$ be a function such that for all m, m',

$$16P_1(m, m') \leq \text{Pen}(m) + \text{Pen}(m'). \quad (4.16)$$

Then (4.15) becomes

$$\begin{aligned} \|\tilde{F} - F\|^2 \leq \|F_m - F\|^2 + \frac{1}{2}(\|\tilde{F} - F\|^2 + \|F - F_m\|^2) + 2\text{Pen}(m) \\ +8[\sup_{T \in B(m,\hat{M})} Z_{n,1}^2(T) - P_1(m, \hat{M})] + 8[\sup_{T \in B(m,\hat{M})} Z_{n,2}^2(T) - P_1(m, \hat{M})] \end{aligned}$$

which gives, by introducing a function $P_2(.,.)$,

$$\frac{1}{2}\|\tilde{F}-F\|^2\mathbb{1}_{\Omega^*} \leq \frac{3}{2}\|F_m-F\|^2 + 2\mathrm{Pen}(m) + 8\sum_{m'\in\mathbb{M}_n}[\sup_{T\in B(m,m')} Z_{n,1}^2(T) - P_1(m,m')]_+$$
$$+8\sum_{m'\in\mathbb{M}_n}[\sup_{T\in B(m,m')} Z_{n,2}^2(T) - P_2(m,m')]_+\mathbb{1}_{\Omega^*} + 8\sum_{m'\in\mathbb{M}_n}[P_2(m,m')-P_1(m,m')].$$

We now use the following propositions:

Proposition 4.1 *Let $P_1(m,m') = C(q)(\pi m'')^{[s-(1-s)_+]_+}\Delta^2(m'')/n$ where $\Delta(m)$ is defined in (4.6) and $m'' = \max(m,m')$ and $C(q)$ is a constant. Then, under assumptions of Theorem 4.2, there exists a positive constant C such that*

$$\sum_{m'\in\mathbb{M}_n}\mathbb{E}\left(\left[\sup_{T\in B(m,m')} Z_{n,1}^2(T) - P_1(m,m')\right]_+\right) \leq \frac{C}{n}. \tag{4.17}$$

Proposition 4.2 *Let $P_2(m,m') = 96(\sum_k \beta_{2k})m''/n$ where $m'' = \max(m,m')$. Then, under assumptions of Theorem 4.2, there exists a positive constant C such that*

$$\sum_{m'\in\mathbb{M}_n}\mathbb{E}\left(\left[\sup_{T\in B(m,m')} Z_{n,2}^2(T) - P_2(m,m')\right]_+\mathbb{1}_{\Omega^*}\right) \leq \frac{C}{n}. \tag{4.18}$$

The definitions of the functions $P_1(m,m')$ and $P_2(m,m')$ given in Propositions 4.1 and 4.2 imply that there exists m_0 such that $\forall m' > m_0 \quad P_1(m,m') \geq P_2(m,m')$. (If $s = 0 = \gamma$ (case of a null noise), it would be wrong and the penalty would then be $P_2(m,m')$ instead of $P_1(m,m')$). Then

$$\sum_{m'\in\mathbb{M}_n}[P_2(m,m')-P_1(m,m')] \leq \sum_{m'\leq m_0} P_2(m,m') \leq \frac{C(m_0)}{n}. \tag{4.19}$$

Since $m''\Delta^2(m'') \leq m\Delta^2(m) + m'\Delta^2(m')$, condition (4.16) is verified with

$$\mathrm{Pen}(m) = 16C(q)(\pi m)^{[s-(1-s)_+]_+}\frac{\Delta^2(m)}{n}.$$

And finally, combining (4.19) and Propositions 4.1 and 4.2,

$$\mathbb{E}(\|\tilde{F}-F\|^2\mathbb{1}_{\Omega^*}) \leq 4(\|F_m-F\|^2 + \mathrm{Pen}(m)) + \frac{C}{n}.$$

For the term $\mathbb{E}(\|\tilde{F}-F\|^2\mathbb{1}_{\Omega^{*c}})$, recall that

$$\hat{F}_m(x,y) = \sum_{j,k}\hat{A}_{j,k}\varphi_{m,j}(x)\varphi_{m,k}(y) \text{ with } \hat{A}_{j,k} = \frac{1}{n}\sum_{i=1}^n V_{\varphi_{m,j}\otimes\varphi_{m,k}}(Y_i,Y_{i+1}).$$

Thus, for any m in $\mathbb{M}_n$,

$$\begin{aligned}\|\hat{F}_m\|^2 &= \sum_{j,k}\left[\frac{1}{n}\sum_{i=1}^{n} V_{\varphi_{m,j}\otimes\varphi_{m,k}}(Y_i,Y_{i+1})\right]^2 \leq \frac{1}{n^2}\sum_{j,k} n\sum_{i=1}^{n} V^2_{\varphi_{m,j}\otimes\varphi_{m,k}}(Y_i,Y_{i+1}) \\ &\leq \|\sum_{j,k} V^2_{\varphi_{m,j}\otimes\varphi_{m,k}}\|_\infty \leq \|\sum_j v^2_{\varphi_{m,j}}\|^2_\infty \leq \Delta^2(m) \qquad (4.20)\end{aligned}$$

using Lemma 4.3 (see Section 4.7.8). Then $\|\hat{F}_{\hat{M}}\|^2 \leq \Delta^2(\hat{M}) \leq n$ since $\hat{M}$ belongs to $\mathbb{M}_n$. And

$$\mathbb{E}\|\tilde{F}-F\|^2\mathbb{1}_{\Omega^{*c}} \leq \mathbb{E}(2(\|\tilde{F}\|^2+\|F\|^2)\mathbb{1}^{*c}_\Omega) \leq 2(n+\|F\|^2)P(\Omega^{*c}).$$

Using Assumption A4 in the geometric case, $\beta_{2q_n} \leq Me^{-\theta c\log(n)} \leq Mn^{-\theta c}$ and, in the other case, $\beta_{2q_n} \leq M(2q_n)^{-\theta} \leq Mn^{-\theta c}$. Then $P(\Omega^{*c}) \leq 2p_n\beta_{2q_n} \leq nMn^{-c\theta}$. Since $c\theta > 4$, $P(\Omega^{*c}) \leq Mn^{-3}$, which implies $E(\|\tilde{F}-F\|^2\mathbb{1}_{\Omega^{*c}}) \leq C/n^2$.

Finally we obtain

$$\begin{aligned}\mathbb{E}\|\tilde{F}-F\|^2 &\leq \mathbb{E}(\|\tilde{F}-F\|^2\mathbb{1}_{\Omega^*}) + E(\|\tilde{F}-F\|^2\mathbb{1}_{\Omega^{*c}}) \\ &\leq 4(\|F_m-F\|^2+\mathrm{Pen}(m)) + \frac{C}{n}.\end{aligned}$$

This inequality holds for each $m \in \mathbb{M}_n$, so the result is proved. □

4.7.4 Proof of Proposition 4.1

We start by isolating odd terms from even terms to avoid overlaps:

$$Z_{n,1}(T) = \frac{1}{2}Z^o_{n,1}(T) + \frac{1}{2}Z^e_{n,1}(T)$$

with

$$\begin{cases} Z^o_{n,1}(T) = \dfrac{2}{n}\displaystyle\sum_{i=1, i \text{ odd}}^{n} \{V_T(Y_i,Y_{i+1}) - \mathbb{E}_X[V_T(Y_i,Y_{i+1})]\}, \\ Z^e_{n,1}(T) = \dfrac{2}{n}\displaystyle\sum_{i=1, \text{ even}}^{n} \{V_T(Y_i,Y_{i+1}) - \mathbb{E}_X[V_T(Y_i,Y_{i+1})]\}. \end{cases}$$

denoting by $\mathbb{E}_X$ the expectation conditionally to $X_1,\ldots,X_{n+1}$. It is sufficient to deal with the first term, the second one being similar. For each i, let $U_i = (Y_{2i-1}, Y_{2i})$, then

$$Z^o_{n,1}(T) = \frac{1}{n/2}\sum_{i=1}^{n/2}\{V_T(U_i) - \mathbb{E}_X[V_T(U_i)]\}.$$

Let us remark that conditionally to $X_1,\ldots,X_n$, the U_i's are independent. Thus we can use the Talagrand inequality recalled in Lemma 4.5 (see Section 4.7.8). Note that if T

belongs to $\mathbb{S}_m + \mathbb{S}_{m'}$, then T can be written $T_1 + T_2$ where T_1^* has its support in $[-\pi m, \pi m]^2$ and T_2^* has its support in $[-\pi m', \pi m']^2$. Then T belongs to $\mathbb{S}_{m''}$ where m'' is defined by

$$m'' = \max(m, m'). \tag{4.21}$$

Now let us compute M_1, H and v of the Talagrand's inequality.

1. If T belongs to $B(m, m')$,

$$V_T(x,y) = \sum_{j,k} a_{jk} V_{\varphi_{m'',j}\otimes\varphi_{m'',k}}(x,y) = \sum_{j,k} a_{jk} v_{\varphi_{m'',j}}(x) v_{\varphi_{m'',k}}(y).$$

   Thus $|V_T(x,y)|^2 \leq \sum_{j,k} |v_{\varphi_{m'',j}}(x) v_{\varphi_{m'',k}}(y)|^2$. So

$$\sup_{T\in B(m,m')} \|V_T\|_\infty^2 \leq \| \sum_{j,k} |v_{\varphi_{m'',j}}(x) v_{\varphi_{m'',k}}(y)|^2 \|_\infty \leq \| \sum_j |v_{\varphi_{m'',j}}|^2 \|_\infty^2.$$

   By using Lemma 4.3, $M_1 = \Delta(m'')$.

2. To compute H^2, we write

$$\begin{aligned}
\mathbb{E}_X(\sup_{T\in B(m,m')} (Z^o_{n,1})^2(T)) &\leq \mathbb{E}_X(\sum_{j,k} Z^o_{n,1}(\varphi_{m'',j}\otimes\varphi_{m'',k})^2) \\
&\leq \sum_{j,k} \mathrm{Var}_X \left[\frac{2}{n} \sum_{i=1, i \text{ odd}}^{n} v_{\varphi_{m'',j}}(Y_i) v_{\varphi_{m'',k}}(Y_{i+1}) \right] \\
&\leq \sum_{j,k} \frac{4}{n^2} \sum_{i=1, i \text{ odd}}^{n} \mathrm{Var}_X [v_{\varphi_{m'',j}}(Y_i) v_{\varphi_{m'',k}}(Y_{i+1})]
\end{aligned}$$

   since, conditionally to $X_1, \ldots, X_{n+1}$, the U_i's are independent. And then

$$\begin{aligned}
\mathbb{E}_X(\sup_{T\in B(m,m')} Z^{o2}_{n,1}(T)) &\leq \sum_{j,k} \frac{4}{n^2} \sum_{i=1, i \text{ odd}}^{n} \mathbb{E}_X [v^2_{\varphi_{m'',j}}(Y_i) v^2_{\varphi_{m'',k}}(Y_{i+1})] \\
&\leq \frac{4}{n^2} \sum_{i=1, i \text{ odd}}^{n} \| \sum_j |v_{\varphi_{m'',j}}|^2 \|_\infty \| \sum_k |v_{\varphi_{m'',k}}|^2 \|_\infty \leq \frac{2\Delta(m'')^2}{n}.
\end{aligned}$$

   So we set $H = \sqrt{2}\Delta(m'')/\sqrt{n}$.

3. We still have to find v. On the one hand

$$\begin{aligned}
\mathrm{Var}_X[V_T(Y_k, Y_{k+1})] &\leq \mathbb{E}_X[(\sum_{j,k} a_{jk} v_{\varphi_{m'',j}}(Y_k) v_{\varphi_{m'',k}}(Y_{k+1}))^2] \\
&\leq \sum_{j,k} a_{jk}^2 \| \sum_j |v_{\varphi_{m'',j}}|^2 \|_\infty \| \sum_k |v_{\varphi_{m'',k}}|^2 \|_\infty
\end{aligned}$$

and so $v \geq \Delta(m'')^2$. On the other hand

$$\begin{aligned}
&\mathrm{Var}_X[V_T(Y_k, Y_{k+1})] \\
&\leq \sum_{j_1,k_1}\sum_{j_2,k_2} a_{j_1k_1}a_{j_2k_2}\mathbb{E}_X[v_{\varphi_{m''},j_1}v_{\varphi_{m''},j_2}(Y_k)v_{\varphi_{m''},k_1}v_{\varphi_{m''},k_2}(Y_{k+1})] \\
&\leq \sum_{j,k} a_{jk}^2 \sqrt{\sum_{j_1,k_1}\sum_{j_2,k_2}\mathbb{E}_X^2[v_{\varphi_{m''},j_1}v_{\varphi_{m''},j_2}(Y_k)v_{\varphi_{m''},k_1}v_{\varphi_{m''},k_2}(Y_{k+1})]} \\
&\leq \sum_{j,k} a_{jk}^2 \sqrt{\sum_{j_1,j_2}\mathbb{E}_X^2[v_{\varphi_{m''},j_1}v_{\varphi_{m''},j_2}(Y_k)]\sum_{k_1,k_2}\mathbb{E}_X^2[v_{\varphi_{m''},k_1}v_{\varphi_{m''},k_2}(Y_{k+1})]}, \quad (4.22)
\end{aligned}$$

using conditional independence. Now we use Lemma 4.3 to compute

$$\begin{aligned}
&\mathbb{E}_X[v_{\varphi_{m''},j_1}v_{\varphi_{m''},j_2}(Y_k)] = \int (v_{\varphi_{m''},j_1}v_{\varphi_{m''},j_2})(X_k+x)q(x)dx \\
&= \frac{m''}{4\pi^2}\int\int_{-\pi}^{\pi}\frac{e^{-ij_1v}e^{i(x+X_k)vm''}}{q^*(-vm'')}dv\int_{-\pi}^{\pi}\frac{e^{-ij_2u}e^{i(x+X_k)um''}}{q^*(-um'')}duq(x)dx \\
&= \frac{m''}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\frac{e^{-ij_1v-ij_2u}e^{iX_k(u+v)m''}}{q^*(-vm'')q^*(-um'')}\int e^{ix(u+v)m''}q(x)dxdudv.
\end{aligned}$$

If we set $W(u,v) = m''e^{iX_k(u+v)m''}q^*(-(u+v)m'')/[q^*(-vm'')q^*(-um'')]$, then $\mathbb{E}_X[v_{\varphi_{m''},j_1}v_{\varphi_{m''},j_2}(Y_k)]$ is the Fourier coefficient with order (j_1,j_2) of W. Using Parseval's formula

$$\begin{aligned}
\sum_{j_1,j_2}\mathbb{E}_X^2[v_{\varphi_{m''},j_1}v_{\varphi_{m''},j_2}(Y_k)] &= \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}|W(u,v)|^2dudv \\
&= \frac{m''^2}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\left|\frac{q^*(-(u+v)m'')}{q^*(-vm'')q^*(-um'')}\right|^2 dudv.
\end{aligned}$$

Now we apply the Schwarz inequality:

$$\begin{aligned}
&\sum_{j_1,j_2}\mathbb{E}_X^2[v_{\varphi_{m''},j_1}(Y_k)] \\
&\leq \frac{m''^2}{4\pi^2}\sqrt{\iint\frac{|q^*(-(u+v)m'')|^2}{|q^*(-um'')|^4}dudv}\sqrt{\iint\frac{|q^*(-(u+v)m'')|^2}{|q^*(-vm'')|^4}dudv} \\
&\leq \frac{m''}{4\pi^2}\int_{-\pi}^{\pi}|q^*(-um'')|^{-4}du\int|q^*(x)|^2dx \leq \frac{\|q\|^2}{2\pi}\int_{-\pi m''}^{\pi m''}|q^*(-u)|^{-4}du.
\end{aligned}$$

We introduce the following notation:

$$\Delta_2(m) = \frac{1}{2\pi}\int_{-\pi m}^{\pi m}|q^*(u)|^{-4}du. \quad (4.23)$$

Hence, coming back to (4.22), $\mathrm{Var}_X[V_T(Y_k, Y_{k+1})] \leq \|T\|^2\|q\|^2\Delta_2(m'')$ which yields $v \geq \|q\|^2\Delta_2(m'')$. Finally we write $v = \min(\|q\|^2\Delta_2(m''), \Delta^2(m''))$.

We can now use Talagrand's inequality (see Lemma 4.5):

$$\begin{aligned}&\mathbb{E}[\sup_{T\in B(m,m')}(Z^o_{n,1})^2(T) - 2(1+2\epsilon)\frac{2\Delta^2(m'')}{n}]_+\\ &\leq \frac{C}{n}\{ve^{-K_1\epsilon\Delta^2(m'')/v} + \frac{\Delta^2(m'')}{nC^2(\epsilon)}e^{-K_2C(\epsilon)\sqrt{\epsilon}\sqrt{n}}\}.\end{aligned}$$

And then, if $P_1(m,m') \geq 4(1+2\epsilon)\Delta^2(m'')/n$,

$$\sum_{m'\in\mathbb{M}_n}\mathbb{E}\left[\sup_{T\in B(m,m')}(Z^o_{n,1})^2(T) - P_1(m,m')\right]_+ \leq \frac{K}{n}\{I(m) + II(m)\}$$

with $I(m) = \sum_{m'\in\mathbb{M}_n} ve^{-K_1\epsilon\Delta^2(m'')/v}$; $II(m) = \sum_{m'\in\mathbb{M}_n}(\Delta^2(m'')/nC^2(\epsilon))e^{-K_2C(\epsilon)\sqrt{\epsilon}\sqrt{n}}$.

To bound these terms, we use Lemma 4.4 which yields to

$$v \leq c_3(\pi m'')^{4\gamma+\min(1-s,2-2s)}e^{4b(\pi m'')^s} \text{ and } \frac{\Delta^2(m'')}{v} \geq c_4(\pi m'')^{(1-s)_+}$$

where c_3 and c_4 depend only on k_0, k_1, γ and s. Therefore,

$$\begin{aligned}I(m) &\leq c_3\sum_{m'\in\mathbb{M}_n}(\pi m'')^{4\gamma+\min(1-s,2-2s)}e^{4b(\pi m'')^s - K_1c_4\epsilon(\pi m'')^{(1-s)_+}}\\ &\leq c_3\sum_{m'\in\mathbb{M}_n}[(\pi m)^{4\gamma+\min(1-s,2-2s)}e^{4b(\pi m)^s}\\ &\quad +(\pi m')^{4\gamma+\min(1-s,2-2s)}e^{4b(\pi m')^s}]e^{-\frac{K_1c_4\epsilon}{2}[(\pi m)^{(1-s)_+}+(\pi m')^{(1-s)_+}]}\\ &\leq c_3(\pi m)^{4\gamma+\min(1-s,2-2s)}e^{4b(\pi m)^s - \frac{K_1c_4\epsilon}{2}(\pi m)^{(1-s)_+}}\sum_{m'\in\mathbb{M}_n}e^{-\frac{K_1c_4\epsilon}{2}(\pi m')^{(1-s)_+}}\\ &\quad +c_3e^{-\frac{K_1c_4\epsilon}{2}(\pi m)^{(1-s)_+}}\sum_{m'\in\mathbb{M}_n}(\pi m')^{4\gamma+\min(1-s,2-2s)}e^{4b(\pi m')^s - \frac{K_1c_4\epsilon}{2}(\pi m')^{(1-s)_+}}.\end{aligned}$$

We have to distinguish three cases

case $s < (1-s)_+ \Leftrightarrow s < 1/2$**.** In this case we choose $\epsilon = 8b/(K_1c_4)$ and then

$$\begin{aligned}I(m) &\leq c_3(\pi m)^{4\gamma+1-s}e^{4b[(\pi m)^s - (\pi m)^{(1-s)}]}\sum_{m'\in\mathbb{M}_n}e^{-K_1c_4(\pi m')^{(1-s)}}\\ &\quad +c_3e^{-K_1c_4(\pi m)^{(1-s)}}\sum_{m'\in\mathbb{M}_n}(\pi m')^{4\gamma+1-s}e^{4b[(\pi m')^s - (\pi m')^{(1-s)}]}\end{aligned}$$

which implies that $I(m)$ is bounded. Moreover the definition of $\mathbb{M}_n$ and Lemma 4.4 give $|\mathbb{M}_n| \leq Cn^\zeta$ with $C > 0$ and $\zeta > 0$. So $II(m) \leq C|\mathbb{M}_n|e^{-K_2'\sqrt{n}}$ is bounded too.

case $s = (1-s)_+ \Leftrightarrow s = 1/2$**.** In this case

$$\begin{aligned} I(m) &\leq c_3(\pi m)^{4\gamma+1/2} e^{(4b-\frac{K_1c_4\epsilon}{2})(\pi m)^{1/2}} \sum_{m' \in \mathbb{M}_n} e^{-\frac{K_1c_4\epsilon}{2}(\pi m')^{1/2}} \\ &\quad + c_3 e^{-\frac{K_1c_4\epsilon}{2}(\pi m)^{1/2}} \sum_{m' \in \mathbb{M}_n} (\pi m')^{4\gamma+1/2} e^{(4b-\frac{K_1c_4\epsilon}{2})(\pi m')^{1/2}}. \end{aligned}$$

We choose ϵ such that $4b - K_1c_4\epsilon/2 = -4b$ so that

$$\begin{aligned} I(m) &\leq c_3(\pi m)^{4\gamma+1/2} e^{-4b(\pi m)^{1/2}} \sum_{m' \in \mathbb{M}_n} e^{-\frac{K_1c_4\epsilon}{2}(\pi m')^{1/2}} \\ &\quad + c_3 e^{-\frac{K_1c_4\epsilon}{2}(\pi m)^{1/2}} \sum_{m' \in \mathbb{M}_n} (\pi m')^{4\gamma+1/2} e^{-4b(\pi m')^{1/2}} \leq C. \end{aligned}$$

The term $II(m)$ is also bounded since ϵ is a constant.

case $s > (1-s)_+ \Leftrightarrow s > 1/2$**.** Here we choose ϵ such that

$$4b(\pi m'')^s - K_1c_4\epsilon(\pi m'')^{(1-s)_+}/2 = -4b(\pi m'')^s$$

so that

$$\begin{aligned} I(m) &\leq c_3(\pi m)^{4\gamma+\min(1-s,2-2s)} e^{-4b(\pi m)^s} \sum_{m' \in \mathbb{M}_n} e^{-\frac{K_1c_4\epsilon}{2}(\pi m')^{(1-s)_+}} \\ &\quad + c_3 e^{-\frac{K_1c_4\epsilon}{2}(\pi m)^{(1-s)_+}} \sum_{m' \in \mathbb{M}_n} (\pi m')^{4\gamma+\min(1-s,2-2s)} e^{-4b(\pi m')^s} \leq C. \end{aligned}$$

Moreover

$$II(m) \leq \sum_{m' \in \mathbb{M}_n} \frac{1}{C^2(\epsilon)} e^{-K_2\sqrt{8b/K_1c_4}(\pi m'')^{[s-(1-s)_+]/2}\sqrt{n}} \leq C.$$

In any case $\epsilon = [8b/K_1c_4](\pi m'')^{[s-(1-s)_+]_+}$, so that

$$P_1(m,m') = C(q)(\pi m'')^{[s-(1-s)_+]_+}\Delta^2(m'')/n$$

where $C(q)$ is a constant depending only on k_0, k_1, b, γ, s.

□

4.7.5 Proof of Proposition 4.2

We split $Z_{n,2}(T)$ into two terms :

$$Z_{n,2}(T) = \frac{1}{2}Z^o_{n,2}(T) + \frac{1}{2}Z^e_{n,2}(T)$$

with

$$\begin{cases} Z^o_{n,2}(T) = \dfrac{2}{n} \displaystyle\sum_{i=1, i \text{ odd}}^{n} \left\{ T(X_i, X_{i+1}) - \int T(x,y) F(x,y) dxdy \right\}, \\ Z^e_{n,2}(T) = \dfrac{2}{n} \displaystyle\sum_{i=1, i \text{ even}}^{n} \left\{ T(X_i, X_{i+1}) - \int T(x,y) F(x,y) dxdy \right\}. \end{cases}$$

We bound $\mathbb{E}\left(\left[\sup_{T \in B(m,m')} (Z^o_{n,2})^2(T) - P_2(m,m') \right]_+ \mathbb{1}_{\Omega^*} \right)$. The second term can be bounded in the same way. We write $Z^o_{n,2}(T) = (2/n) \sum_{i=1}^{n/2} \{T(V_i) - \mathbb{E}[T(V_i)]\}$ with $V_i = (X_{2i-1}, X_{2i})$. In order to use Lemma 4.5, we introduce

$$Z^{o*}_{n,2}(T) = \frac{1}{2}\nu_{n,1}(T) + \frac{1}{2}\nu_{n,2}(T)$$

where

$$\begin{cases} \nu_{n,1}(T) = \dfrac{1}{p_n} \displaystyle\sum_{l=0}^{p_n - 1} \frac{1}{q_n} \sum_{i=2lq_n+1}^{(2l+1)q_n} \{T(V^*_i) - \mathbb{E}[T(V^*_i)]\}, \\ \nu_{n,2}(T) = \dfrac{1}{p_n} \displaystyle\sum_{l=0}^{p_n - 1} \frac{1}{q_n} \sum_{i=(2l+1)q_n+1}^{(2l+2)q_n} \{T(V^*_i) - \mathbb{E}[T(V^*_i)]\}. \end{cases}$$

Since $X_i = X^*_i$ on Ω^*, we can replace $Z^o_{n,2}$ by $Z^{o*}_{n,2}$. This leads us to bound

$$\mathbb{E}\left(\left[\sup_{T \in B(m,m')} \nu^2_{n,1}(T) - P_2(m,m') \right]_+ \mathbb{1}_{\Omega^*} \right).$$

So we compute the bounds M_1, H and v of Lemma 4.5.

1. If T belongs to $\mathbb{S}_{m''}$, $|T(x,y)|^2 \leq \sum_{j,k} a^2_{j,k} \sum_{j,k} \varphi^2_{m'',j}(x) \varphi^2_{m'',k}(y)$ and so

$$\|T\|_\infty \leq \|T\| \| \sum_j \varphi^2_{m'',j} \|_\infty \leq \|T\| m'',$$

using the first point of Lemma 4.3. Then $\| 1/q_n \sum_{i=2lq_n+1}^{(2l+1)q_n} T \|_\infty \leq \|T\| m''$ and $M_1 = m''$.

2. Let us compute H^2.

$$\sup_{T \in B(m,m')} \nu^2_{n,1}(T) \leq \sum_{j,k} \nu^2_{n,1}(\varphi_{m'',j} \otimes \varphi_{m'',k})$$

Then, by taking the expectation,

$$\mathbb{E}\left(\sup_{T\in B(m,m')}\nu_{n,1}^2(T)\right) \leq \sum_{j,k}\frac{1}{p_n^2}\mathrm{Var}(\sum_{l=0}^{p_n-1}\frac{1}{q_n}\sum_{i=2lq_n+1}^{(2l+1)q_n}\varphi_{m'',j}\otimes\varphi_{m'',k}(V_i^*))$$
$$\leq \sum_{j,k}\frac{1}{p_n^2}\sum_{l=0}^{p_n-1}\mathrm{Var}(\frac{1}{q_n}\sum_{i=2lq_n+1}^{(2l+1)q_n}\varphi_{m'',j}\otimes\varphi_{m'',k}(V_i^*)),$$

by using independence of the A_l^*. Lemma 4.6 then gives

$$\mathbb{E}\left(\sup_{T\in B(m,m')}\nu_{n,1}^2(T)\right) \leq \frac{4}{p_nq_n}\|\sum_{j,k}(\varphi_{m'',j}\otimes\varphi_{m'',k})^2\|_\infty\sum\beta_{2k} \leq \frac{16}{n}(\sum\beta_{2k})m''^2$$

Finally $H = 4\sqrt{\sum\beta_{2k}}m''/\sqrt{n}$.

3. v remains to be calculated. If T belongs to $B(m')$, using Lemma 4.6

$$\mathrm{Var}[\frac{1}{q_n}\sum_{i=2lq_n+1}^{(2l+1)q_n}T(V_i^*)] \leq \frac{4}{q_n}\mathbb{E}[T^2(V_1)b(V_1)]$$
$$\leq \frac{4}{q_n}\|T\|_\infty\sqrt{\mathbb{E}[T^2(V_1)]}\sqrt{\mathbb{E}[b^2(V_1)]} \leq \frac{4}{q_n}\|T\|_\infty\sqrt{\|F\|_\infty}\sqrt{2\sum(k+1)\beta_{2k}}$$

and so $v = 4\|F\|_\infty^{1/2}\sqrt{2\sum(k+1)\beta_{2k}}m''/q_n$.

By writing Talagrand's inequality (Lemma 4.5) with $\epsilon = 1$, we obtain

$$\mathbb{E}\left(\left[\sup_{T\in B(m,m')}(\nu_{n,1})^2(T) - 6\frac{16}{n}(\sum\beta_{2k})m''^2\right]_+\mathbb{1}_{\Omega^*}\right) \leq \frac{K}{n}\left\{m''e^{-K_1m''} + \frac{m''q_n^2}{n}e^{-K_2\sqrt{n}/q_n}\right\}.$$

Then by summation over m'

$$\sum_{m'\in\mathbb{M}_n}\mathbb{E}\left([\sup_{T\in B(m,m')}(\nu_{n,1})^2(T) - \frac{96}{n}(\sum\beta_{2k})m''^2]_+\mathbb{1}_{\Omega^*}\right)$$
$$\leq \frac{K}{n}\{\sum_{m'\in\mathbb{M}_n}m''e^{-K_1m''} + \sum_{m'\in\mathbb{M}_n}m''n^{2c-1}e^{-K_2n^{1/2-c}}\} \leq \frac{C}{n}$$

since $c < 1/2$. In the same way, we obtain

$$\sum_{m'\in\mathbb{M}_n}\mathbb{E}\left([\sup_{T\in B(m,m')}(\nu_{n,2})^2(T) - \frac{96}{n}(\sum\beta_{2k})m''^2]_+\mathbb{1}_{\Omega^*}\right) \leq \frac{C}{n},$$

which yields

$$\sum_{m'\in\mathbb{M}_n}\mathbb{E}\left([\sup_{T\in B(m,m')}(Z^o_{n,2})^2(T)-P_2(m,m')]_+\mathbb{1}_{\Omega^*}\right)\leq\frac{C}{n}$$

with $P_2(m,m')=96(\sum\beta_{2k})m''^2/n$.

□

4.7.6 Proof of Corollary 4.2

Let us compute the bias term. Since $F^*_m=F^*\mathbb{1}_{[-\pi m,\pi m]^2}$,

$$\begin{aligned}\|F-F_m\|^2&=\frac{1}{4\pi^2}\iint_{([-\pi m,\pi m]^2)^c}|F^*(u,v)|^2dudv\\&\leq\frac{1}{4\pi^2}\iint_{[-\pi m,\pi m]^c\times\mathbb{R}}|F^*(u,v)|^2dudv+\frac{1}{4\pi^2}\iint_{\mathbb{R}\times[-\pi m,\pi m]^c}|F^*(u,v)|^2dudv\end{aligned}$$

But

$$\iint_{[-\pi m,\pi m]^c\times\mathbb{R}}|F^*(u,v)|^2dudv\leq L((\pi m)^2+1)^{-\Delta}e^{-2A(\pi m)^R}.$$

Thus $\|F-F_m\|^2=O((\pi m)^{-2\Delta}e^{-2A(\pi m)^R})$ and

$$\mathbb{E}\|F-\tilde F\|^2\leq C'\inf_{m\in\mathbb{M}_n}\left\{(\pi m)^{-2\Delta}e^{-2A(\pi m)^R}+(\pi m)^{[s-(1-s)_+]_++4\gamma+2-2s}\frac{e^{4b(\pi m)^s}}{n}\right\}+\frac{C}{n}.$$

Next the bias-variance trade-off is performed similarly to the chapter 3.

□

4.7.7 Proof of Theorem 4.3

Let

$$E_n=\{\|f-\tilde f\|_\infty\leq f_0/2\}.$$

On E_n and for $x\in A_1$, $\tilde f(x)=\tilde f(x)-f(x)+f(x)\geq f_0/2$. Since $\tilde F$ belongs to $\mathbb{S}_{\hat M}$, using (4.2), $\|\tilde F\|_\infty\leq\hat M\|\tilde F\|$. Now (4.20) gives $\|\tilde F\|\leq\Delta(\hat M)$ so that

$$\|\tilde F\|_\infty\leq\hat M\Delta(\hat M).$$

Since $\hat M$ belongs to $\mathbb{M}_n$, $\Delta(\hat M)\leq\sqrt{n}$ and Lemma 4.4 gives $\hat M\leq C\Delta(\hat M)^{1/(2\gamma+1)}$ if $s=0$ or $\hat M\leq C(\log\Delta(\hat M))^{1/s}$ otherwise. So, for n large enough, $(2/f_0)\|\tilde F\|_\infty\leq n$ and $\tilde\Pi(x,y)=\tilde F(x,y)/\tilde f(x)$.

For all $(x,y) \in A$,

$$\begin{aligned}|\tilde{\Pi}(x,y) - \Pi(x,y)|^2 &\leq \left|\frac{\tilde{F}(x,y) - \tilde{f}(x)\Pi(x,y)}{\tilde{f}(x)}\right|^2 \mathbb{1}_{E_n} + (|\tilde{\Pi}(x,y)| + |\Pi(x,y|)^2 \mathbb{1}_{E_n^c} \\ &\leq \frac{|\tilde{F}(x,y) - F(x,y) + \Pi(x,y)(f(x) - \tilde{f}(x))|^2}{f_0^2/4} \\ &\quad +2(\|\tilde{\Pi}\|_\infty^2 + |\Pi(x,y)|^2) 1_{E_n^c}.\end{aligned}$$

Since $\int_{A_2} \Pi^2(x,y)dy \leq \|\Pi\|_{A,\infty} \int_{A_2} \Pi(x,y)dy \leq \|\Pi\|_{A,\infty}$ for all $x \in A_1$,

$$\mathbb{E}\|\Pi - \tilde{\Pi}\|_A^2 \leq \frac{8}{f_0^2}[\mathbb{E}\|F - \tilde{F}\|^2 + \|\Pi\|_{A,\infty}\mathbb{E}\|f - \tilde{f}\|^2] + 2|A_1|(|A_2|n^2 + \|\Pi\|_{A,\infty})P(E_n^c).$$

We still have to prove that $P(E_n^c) \leq Cn^{-3}$. Given that $\|f - \tilde{f}\|_\infty \leq \|f - f_{\hat{m}}\|_\infty + \|f_{\hat{m}} - \hat{f}_{\hat{m}}\|_\infty$ we obtain

$$P(E_n^c) \leq P(\|f - f_{\hat{m}}\|_\infty > f_0/4) + P(\|f_{\hat{m}} - \hat{f}_{\hat{m}}\|_\infty > f_0/4).$$

Let us now prove that if f belongs to $\mathcal{A}_{\delta,r,a}(l)$, $\|f - f_m\|_\infty = O(m^{1/2-\delta-r/2}e^{-a(\pi m)^r})$. Since $f_m^* = f^* \mathbb{1}_{[-\pi m, \pi m]}$ and using the inverse Fourier transform,

$$\|f - f_m\|_\infty \leq \frac{1}{2\pi} \int_{|u| \geq \pi m} |f^*(u)|du.$$

If $r > 0$, let $0 < \alpha < a$. By considering that function $x \mapsto (x^2+1)^{\delta/2} e^{(a-\alpha)|x|^r}$ is increasing and using the Schwarz inequality, we obtain

$$\|f - f_m\|_\infty \leq \frac{1}{2\pi}((\pi m)^2 + 1)^{-\delta/2} e^{(\alpha - a)(\pi m)^r} \sqrt{l} \sqrt{\int_{|u| \geq \pi m} e^{-2\alpha|u|^r} du}.$$

But $\int_{|u| \geq \pi m} e^{-2\alpha|u|^r} \leq C(\pi m)^{1-r} e^{-2\alpha(\pi m)^r}$ and then

$$\|f - f_m\|_\infty \leq \frac{\sqrt{Cl}}{2\pi}((\pi m)^2+1)^{-\delta/2} e^{(\alpha-a)(\pi m)^r} (\pi m)^{1/2-r/2} e^{-\alpha(\pi m)^r} = O(m^{1/2-\delta-r/2} e^{-a(\pi m)^r}).$$

If $r = 0$, we use the increasing fonction $x \mapsto (x^2+1)^{(\delta - \delta')/2}$ with $\delta' < \delta$ and we obtain $\|f - f_m\|_\infty = O(m^{1/2-\delta})$. Thus, since $\hat{m} \geq \log\log n$, $\|f - f_{\hat{m}}\|_\infty \to 0$ and for n large enough $P(\|f - f_{\hat{m}}\|_\infty > f_0/4) = 0$. Next

$$P(\|f_{\hat{m}} - \hat{f}_{\hat{m}}\|_\infty > f_0/4) \leq P(\Omega^{*c}) + P\left(\|f_{\hat{m}} - \hat{f}_{\hat{m}}\| \mathbb{1}_{\Omega^*} > \frac{f_0}{4\sqrt{\hat{m}}}\right).$$

Since $c\theta > 4$, $P(\Omega^{*c}) \leq Mn^{1-c\theta} \leq Mn^{-3}$. We still have to prove that

$$P\left(\|f_{\hat{m}} - \hat{f}_{\hat{m}}\| \mathbb{1}_{\Omega^*} > \frac{f_0}{4\sqrt{\hat{m}}}\right) \leq \frac{C}{n^3}.$$

First, we observe that

$$\|f_{\hat{m}} - \hat{f}_{\hat{m}}\|^2 = \sum_{j\in\mathbb{Z}} \left(\frac{1}{n}\sum_{i=1}^{n} v_{\varphi_{\hat{m}j}}(Y_i) - \mathbb{E}[v_{\varphi_{\hat{m}j}}(Y_i)] \right)^2 = \sup_{t\in B_{\hat{m}}} \nu_n^2(t)$$

where $\nu_n(t) = \frac{1}{n}\sum_{i=1}^{n} v_t(Y_i) - \mathbb{E}[v_t(Y_i)]$, $B_m = \{t \in S_m, \|t\| \leq 1\}$.

Then

$$P\left(\|f_{\hat{m}} - \hat{f}_{\hat{m}}\| \mathbb{1}_{\Omega^*} > \frac{f_0}{4\sqrt{\hat{m}}} \right) = P\left(\sup_{t\in B_{\hat{m}}} |\nu_n(t)| \mathbb{1}_{\Omega^*} > \frac{f_0}{4\sqrt{\hat{m}}} \right).$$

As previously, we split $\nu_n(t)$ into two terms

$$\nu_n(t) = \frac{1}{2p_n}\sum_{l=0}^{p_n-1}\frac{1}{q_n}\sum_{i=2lq_n+1}^{(2l+1)q_n} v_t(Y_i) - \mathbb{E}[v_t(Y_i)] + \frac{1}{2p_n}\sum_{l=0}^{p_n-1}\frac{1}{q_n}\sum_{i=(2l+1)q_n+1}^{(2l+2)q_n} v_t(Y_i) - \mathbb{E}[v_t(Y_i)]$$

and it is sufficient to study

$$P\left(\sup_{t\in B_{\hat{m}}} \left| \frac{1}{p_n}\sum_{l=0}^{p_n}\frac{1}{q_n}\sum_{i=2lq_n+1}^{(2l+1)q_n} v_t(Y_i^*) - \mathbb{E}[v_t(Y_i^*)] \right| > \frac{f_0}{4\sqrt{\hat{m}}} \right).$$

We bound this term by the sum

$$\sum_{m\in\mathcal{M}_n} P\left(\sup_{t\in B_m} \left| \frac{1}{p_n}\sum_{l=0}^{p_n}\frac{1}{q_n}\sum_{i=2lq_n+1}^{(2l+1)q_n} v_t(Y_i^*) - \mathbb{E}[v_t(Y_i^*)] \right| > \frac{f_0}{4\sqrt{m}} \right)$$

and we use inequality (4.24) in proof of Lemma 4.5 with $\eta = 1$ and $\lambda = \dfrac{f_0}{8\sqrt{m}}$:

$$P\left(\sup_{t\in B_m} \left| \frac{1}{p_n}\sum_{l=0}^{p_n}\frac{1}{q_n}\sum_{i=2lq_n+1}^{(2l+1)q_n} v_t(Y_i^*) - \mathbb{E}[v_t(Y_i^*)] \right| > 2H + \lambda \right) \leq \exp\left(-Kp_n \min\left(\frac{\lambda^2}{v}, \frac{\lambda}{M_1} \right) \right)$$

Here, we compute

$$M_1 = \sqrt{\Delta(m)}; \quad H^2 = 8\sum_k \beta_k \frac{\Delta(m)}{n}; \quad v = 4\sum_k \beta_k \frac{\Delta(m)}{q_n}.$$

Thus

$$P\left(\sup_{t\in B_m} |\nu_n(t)| > 2H + \frac{f_0}{8\sqrt{m}} \right) \leq 2\exp\left(-K' \min\left(\frac{n}{m\Delta(m)}, \frac{p_n}{\sqrt{m\Delta(m)}} \right) \right).$$

Now we use the assumption $\forall m \quad m\Delta(m) \leq n/(\log n)^2$. For n large enough, $2H = 4\sqrt{2\sum_k \beta_k}\sqrt{\Delta(m)}/\sqrt{n} \leq f_0/(8\sqrt{m})$. So

$$\sum_{m\in\mathcal{M}_n} P\left(\sup_{t\in B_m}|\nu_n(t)| > \frac{f_0}{4\sqrt{m}}\right) \leq 2|\mathcal{M}_n|\exp\left(-K'\min\left((\log n)^2, n^{1/2-c}\log n\right)\right) \leq \frac{C}{n^3}.$$

□

4.7.8 Technical Lemmas

Lemma 4.3 *For each $m \in \mathcal{M}_n$*

1. $\|\sum_j \varphi_{m,j}^2\|_\infty = m$
2. $v_{\varphi_{m,j}}(x) = \sqrt{m}/(2\pi)\int_{-\pi}^{\pi} e^{-ijv}e^{ixvm}[q^*(-vm)]^{-1}dv$
3. $\|\sum_j |v_{\varphi_{m,j}}|^2\|_\infty = \Delta(m)$

where $\Delta(m)$ is defined in (4.6).

Proof of Lemma 4.3: First we remark that

$$\begin{aligned}\varphi_{m,j}^*(u) &= \int e^{-ixu}\sqrt{m}\varphi(mx-j)dx \\ &= \frac{1}{\sqrt{m}}e^{-iju/m}\int e^{-ixu/m}\varphi(x)dx = \frac{1}{\sqrt{m}}e^{-iju/m}\varphi^*(\frac{u}{m}).\end{aligned}$$

Thus, using the inverse Fourier transform

$$\varphi_{m,j}(x) = \frac{1}{2\pi}\int e^{iux}\frac{1}{\sqrt{m}}e^{-iju/m}\varphi^*(\frac{u}{m})du = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-ijv}\sqrt{m}e^{ixvm}dv.$$

The Parseval equality yields $\sum_j \varphi_{m,j}^2(x) = 1/2\pi\int_{-\pi}^{\pi}|\sqrt{m}e^{ixvm}|^2dv = m$. The first point is proved. Now we compute $v_{\varphi_{m,j}}(x)$

$$\begin{aligned}v_{\varphi_{m,j}}(x) &= \frac{1}{2\pi}\int e^{ixu}\frac{\varphi_{m,j}^*(u)}{q^*(-u)}du = \frac{1}{2\pi}\int e^{ixu}\frac{1}{\sqrt{m}}e^{-iju/m}\varphi^*(\frac{u}{m})\frac{du}{q^*(-u)} \\ &= \frac{\sqrt{m}}{2\pi}\int e^{-ijv}e^{ixvm}\frac{\varphi^*(v)}{q^*(-vm)}dv.\end{aligned}$$

But $\varphi^*(v) = \mathbb{1}_{[-\pi,\pi]}(v)$ and thus the second point is proved. Moreover $v_{\varphi_{m,j}}(x)$ can be seen as a Fourier coefficient. Parseval's formula then gives

$$\sum_j |v_{\varphi_{m,j}}(x)|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\sqrt{m}e^{ixvm}\frac{1}{q^*(-vm)}\right|^2 dv = \frac{m}{2\pi}\int |q^*(-vm)|^{-2}\,dv.$$

Therefore $\|\sum_j |v_{\varphi_{m,j}}|^2\|_\infty = 1/2\pi\int_{-\pi m}^{\pi m}|q^*(-u)|^{-2}du = \Delta(m)$. □

Lemma 4.4 *If q verifies $|q^*(x)| \geq k_0(x^2+1)^{-\gamma/2}\exp(-b|x|^s)$, then*

1. $\Delta(m) \leq c_1(\pi m)^{2\gamma+1-s}e^{2b(\pi m)^s}$,
2. $\Delta_2(m) \leq c_2(\pi m)^{4\gamma+1-s}e^{4b(\pi m)^s}$.

Moreover if $|q^*(x)| \leq k_1(x^2+1)^{-\gamma/2}\exp(-b|x|^s)$, then $\Delta(m) \geq c_1'(\pi m)^{2\gamma+1-s}e^{2b(\pi m)^s}$.

The proof of this result is omitted. It is obtained by distinguishing the cases $s > 2\gamma+1$ and $s \leq 2\gamma+1$ and with standard evaluations of integrals.

Lemma 4.5 *Let $T_1, \ldots, T_n$ be independent random variables and*

$$\nu_n(r) = (1/n)\sum_{i=1}^{n}[r(T_i) - \mathbb{E}(r(T_i))],$$

for r belonging to a countable class $\mathcal{R}$ of measurable functions. Then, for $\epsilon > 0$,

$$\mathbb{E}[\sup_{r\in\mathcal{R}}|\nu_n(r)|^2 - 2(1+2\epsilon)H^2]_+ \leq C\left(\frac{v}{n}e^{-K_1\epsilon\frac{nH^2}{v}} + \frac{M_1^2}{n^2C^2(\epsilon)}e^{-K_2C(\epsilon)\sqrt{\epsilon}\frac{nH}{M_1}}\right)$$

with $K_1 = 1/6$, $K_2 = 1/(21\sqrt{2})$, $C(\epsilon) = \sqrt{1+\epsilon}-1$ and C a universal constant and where

$$\sup_{r\in\mathcal{R}}\|r\|_\infty \leq M_1, \quad \mathbb{E}\left(\sup_{r\in\mathcal{R}}|\nu_n(r)|\right) \leq H, \quad \sup_{r\in\mathcal{R}}\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(r(T_i)) \leq v.$$

Usual density arguments allow to use this result with non-countable class of functions $\mathcal{R}$.

Proof of Lemma 4.5: We apply the Talagrand concentration inequality given in Klein and Rio (2005) to the functions $s^i(x) = r(x) - \mathbb{E}(r(T_i))$ and we obtain

$$P(\sup_{r\in\mathcal{R}}|\nu_n(r)| \geq H+\lambda) \leq \exp\left(-\frac{n\lambda^2}{2(v+4HM_1)+6M_1\lambda}\right).$$

Then we modify this inequality following Birgé and Massart (1998) Corollary 2 p.354. It gives

$$P(\sup_{r\in\mathcal{R}}|\nu_n(r)| \geq (1+\eta)H+\lambda) \leq \exp\left(-\frac{n}{3}\min\left(\frac{\lambda^2}{2v}, \frac{\min(\eta,1)\lambda}{7M_1}\right)\right). \tag{4.24}$$

To conclude we set $\eta = \sqrt{1+\epsilon}-1$ and we use the formula $\mathbb{E}[X]_+ = \int_0^\infty P(X \geq t)dt$ with $X = \sup_{r\in\mathcal{R}}|\nu_n(r)|^2 - 2(1+2\epsilon)H^2$. □

Lemma 4.6 *(Viennet (1997) Theorem 2.1 and Lemma 4.2) Let (T_i) a strictly stationary process with β-mixing coefficients β_k. Then there exists a function b such that*

$$\mathbb{E}[b(T_1)] \leq \sum_k \beta_k \quad \text{and} \quad \mathbb{E}[b^2(T_1)] \leq 2\sum_k (k+1)\beta_k$$

and for all function ψ (such that $\mathbb{E}[\psi^2(T_1)] < \infty$) and for all N

$$\mathrm{Var}(\sum_{i=1}^{N}\psi(T_i)) \leq 4N\mathbb{E}[\psi^2(T_1)b(T_1)].$$

Chapitre 5

Estimation de la densité de transition par contraste moindres carrés

5.1 Introduction

In this chapter we consider the following additive hidden Markov model:

$$Y_i = X_i + \varepsilon_i \qquad i = 1, \dots, n+1 \tag{5.1}$$

with $(X_i)_{i\geq 1}$ a real-valued Markov chain, $(\varepsilon_i)_{i\geq 1}$ a sequence of independent and identically distributed variables and

$$(X_i)_{i\geq 1} \text{ and } (\varepsilon_i)_{i\geq 1} \text{ independent.} \tag{5.2}$$

Only the variables $Y_1, ..., Y_{n+1}$ are observed. Besides its initial distribution, the chain $(X_i)_{i\geq 1}$ is characterized by its transition, i.e. the distribution of X_{i+1} knowing X_i. We assume that this transition has a density Π, defined by $\Pi(x, y)dy = P(X_{i+1} \in dy | X_i = x)$, and it is this quantity that we want to estimate.

This model belongs to the class of hidden Markov models. For a general reference on these models, we refer to Cappé *et al.* (2005). Here, we study a simple model of HMM where the noise is additive.

This model is also similar to the so-called convolution model (for which the aim is to estimate the density of $(X_i)_{i\geq 1}$). As proceeded for this model, we use extensively the Fourier transform. The restrictions on the error distribution and the rate of convergence obtained for our estimator are also of the same kind. Related works include Stefanski (1990), Fan (1993), Masry (1993) (for the multivariate case), Pensky and Vidakovic (1999), Comte *et al.* (2006b).

The estimation of the transition density of a hidden Markov chain is studied by Clémençon (2003). His estimator is based on thresholding of a wavelet-vaguelette decomposition. The drawback of this estimator is that it does not achieve the minimax rate because of a logarithmic loss. In the chapter 4, we described an estimation procedure by quotient of an estimator of the joint density F and an estimator of the stationary density f. The minimax rate is reached by this estimator if we assume that f and Πf have the regularity α. But this smoothness condition on f raises a problem. Indeed Clémençon (2000) exhibits an example where the stationary density f is not continuous, whereas the transition density Π is constant. It shows that f can be much less regular than Π. Our aim is then to find an estimator of the transition density which does not have these disadvantages.

To estimate Π, we use an original contrast inspired by the mean square contrast. The first idea is to connect our problem with the regression model. For any function g, we can write

$$g(X_{i+1}) = \left(\int \Pi(., y) g(y) dy \right) (X_i) + \eta_{i+1}$$

where $\eta_{i+1} = g(X_{i+1}) - \mathbb{E}[g(X_{i+1})|X_i]$. Then, for all function g, we can consider $\int \Pi g$ as a regression function. The mean square contrast to estimate this regression function, if the

X_i were known, should be $(1/n)\sum_{i=1}^{n}[t^2(X_i) - 2t(X_i)g(X_{i+1})]$. If $\int g^2 = 1$, this contrast can be written $(1/n)\sum_{i=1}^{n}[\int T^2(X_i, y)dy - 2T(X_i, X_{i+1})]$ by setting $T(x, y) = t(x)g(y)$ i.e. T such that $\int T(x, y)g(y)dy = t(x)$. It is this contrast that is used in the second chapter but in our case, only the $Y_1, \ldots, Y_{n+1}$ are known. So we introduce in this paper two operators Q and V such that $\mathbb{E}[Q_{T^2}(Y_i)|X_i] = \int T^2(X_i, y)dy$ and $\mathbb{E}[V_T(Y_i, Y_{i+1})|X_i, X_{i+1}] = T(X_i, X_{i+1})$. It leads to the following contrast:

$$\gamma_n(T) = \frac{1}{n}\sum_{i=1}^{n}[Q_{T^2}(Y_i) - 2V_T(Y_i, Y_{i+1})]. \tag{5.3}$$

A collection of estimators is then defined by minimization of this contrast on wavelet spaces. Indeed, wavelets have many useful properties and in particular they can have a compact support and can be regular enough to balance the smoothness of the noise. A general reference on the subject is the book of Meyer (1990).

A method of model selection inspired by Barron *et al.* (1999) and based on contrast (5.3) is used to build an adaptive estimator. A data driven choice of model is performed via the minimization of a penalized criterion. The chosen model is the one which minimizes the empirical risk added to a penalty function. In most cases in estimation of mixing processes, a mixing term appears in this penalty. In the same way, some unknown terms derived from the dependence between the X_i appears in the thresholding constant used to define the estimator of Clémençon (2003). Here a conditioning argument allows to lead us back to independent variables and thus to avoid such a mixing term in the penalty. Our penalty contains only known quantities or terms that can be estimated and is then computable.

For an ordinary smooth noise with regularity γ, the rate of convergence $n^{-\alpha/(2\alpha+4\gamma+2)}$ is obtained if the transition Π is supposed to belong to a Besov space with regularity α. Our estimator is then better than those of Clémençon which achieves only the rate $(\log(n)/n)^{\alpha/(2\alpha+4\gamma+2)}$. Moreover this rate is obtained without supposing known the regularity α of f, our estimator is then adaptive.

This chapter is organized as follows. In Section 5.2 we present the model and the assumptions. Section 5.3 is devoted to the definitions of the contrast and of the estimator. The main result and a sketch of proof are to be found in Section 5.4. Numerical illustration through simulated examples is reported in Section 5.5. The detailed proofs are gathered in Section 5.6.

5.2 Study framework

We consider the model defined by (5.1) and (5.2) where $(X_i)_{i\geq 1}$ is an irreducible and positive recurrent Markov chain with values in the real line $\mathbb{R}$. We assume that $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed random variables with known distribution.

The purpose of this chapter is to estimate the transition density Π of the hidden chain from the data $Y_1, \ldots, Y_{n+1}$. We estimate Π on a compact set $A = A_1 \times A_2$ only.

As in the previous chapter, the Fourier transform t^* of t is defined by

$$t^*(u) = \int e^{-ixu} t(x) dx.$$

And for a bivariate function T,

$$T^*(u, v) = \iint e^{-ixu - iyv} T(x, y) dx dy.$$

Assumptions on the noise

The density of ε_i is denoted by q and is assumed to be known. We assume that the Fourier transform of q never vanishes and that q is ordinary smooth. More precisely the assumption on the error density is the following.

H5 q is uniformly bounded and there exist $\gamma > 0$ and $k_0 > 0$ such that $\forall x \in \mathbb{R}\ |q^*(x)| \geq k_0(x^2+1)^{-\gamma/2}$.

This assumption restrains the regularity class of the noise. Among the so-called ordinary smooth noises, we can cite the Laplace distribution, the exponential distribution and all the Gamma or symmetric Gamma distributions. A noise follows a Gamma distribution with scale parameter λ and shape parameter ζ if $q(x) = \lambda^\zeta x^{\zeta-1} e^{-\lambda x}/\Gamma(\zeta)$ for $x > 0$ with Γ the classic Gamma function, and then

$$|q^*(x)| = \left(1 + \frac{x^2}{\lambda^2}\right)^{-\zeta/2}.$$

So q is bounded if $\zeta \geq 1$ and verifies H5 with $\gamma = \zeta$. The case $\zeta = 1$ corresponds to an exponential distribution and if $\lambda = 1/2$, $\zeta = p/2$, it is a chi-square $\chi(p)$. A Laplace noise is defined in the following way

$$q(x) = \frac{\lambda}{2} e^{-\lambda|x-\mu|} \text{ and } |q^*(x)| = \frac{\lambda^2}{x^2 + \lambda^2}$$

Then H5 is satisfied with $\gamma = 2$. More generally, we can define the symmetric gamma distribution with density $q(x) = \lambda^\zeta |x|^{\zeta-1} e^{-\lambda|x|}/(2\Gamma(\zeta))$. The characteristic function is then

$$q^*(x) = \left(1 + \frac{x^2}{\lambda^2}\right)^{-\zeta/2} \cos\left(2\zeta \arctan\left(\frac{x}{\lambda + \sqrt{x^2+\lambda^2}}\right)\right)$$

so that H5 is verified with $\gamma = \zeta + 1$ if ζ is an odd integer and $\gamma = \zeta$ otherwise.

Remark 5.1 *We have to notice that the Gaussian noise does not verify Assumption H5. Indeed, an exponential decreasing of the Fourier transform of the error density is more difficult to control and a supersmooth noise makes the denoising more difficult. For that reason, many authors have considered only ordinary smooth noise, one can cite among others Butucea (2004), Koo and Lee (1998) or Youndjé and Wells (2002). The method used in this chapter does not allow to deal with supersmooth noise. Indeed, it requires a basis more regular than the noise and also with compact support (because of Assumption H4), which is impossible.*

Assumptions on the chain

The assumptions on the hidden Markov chain are the following.

H1 The chain is irreducible, positive recurrent and stationary.

H2b The process (X_k) is geometrically β-mixing ($\beta_q \leq e^{-\theta q}$), or arithmetically β-mixing ($\beta_q \leq q^{-\theta}$ with $\theta > 8$) where

$$\beta_q = \int \|P^q(x,.) - \mu\|_{TV} f(x)dx$$

with $P^q(x,.)$ the distribution of X_{i+q} given $X_i = x$ and μ the stationary distribution and $\|.\|_{TV}$ the total variation distance.

H3a The stationary density f verifies $\|f\|_{\infty,A_1} := \sup_{x\in A_1} |f(x)| < \infty$

H3e The transition density Π is integrable on A.

H4 There exists a positive real f_0 such that, for all x in $A1$, $f(x) \geq f_0$.

Assumption H2b implies that the process (Y_k) is β-mixing with β-mixing coefficients smaller than those of (X_k). Assumption H4 is common (but restrictive) and is crucial to control the empirical processes brought into play. Many processes verify Assumptions H1–H4, as autoregressive processes, diffusions or ARCH processes. These examples are detailed in Chapter 1.

5.3 Estimation procedure

5.3.1 Projection spaces

Here we describe the projection that we use in this paper to estimate the transition Π. We will consider an increasing sequence of spaces, indexed by m, to construct a collection of estimators. For the sake of simplicity, we assume in this section that $A = [0,1]^2$.

We use a compactly supported wavelet basis on the interval $[0,1]$, described in Cohen *et al.* (1993). The construction furnishes a set of functions (ϕ_k) for $k = 0, \dots, 2^J - 1$ with J a fixed level, and for all $j > J$ a set of functions $(\psi_{jk}), k = 0, \dots, 2^j - 1$. The collection of these functions forms a complete orthonormal system on $[0,1]$. Then , for u in $L^2([0,1])$, we can write

$$u = \sum_{k=0}^{2^J-1} b_k \phi_k + \sum_{j>J} \sum_{k=0}^{2^j-1} a_{jk} \psi_{jk}.$$

Actually

$$\phi_k(x) = \begin{cases} 2^{J/2}\phi^0(2^J x - k) & \text{if } k = 0, \dots, N-1 \\ 2^{J/2}\phi(2^J x - k) & \text{if } k = N, \dots, 2^J - N - 1 \\ 2^{J/2}\phi^1(2^J x - k) & \text{if } k = 2^J - N, \dots, 2^J - 1 \end{cases}$$

where ϕ is a Daubechies father wavelet with support $[-N+1, N]$ and ϕ^0, ϕ^1 are edge wavelets explicitely constructed in Cohen *et al.* (1993). The functions ϕ_{jk} have support $[(k-N+1)/2^J, (k+N)/2^J] \cap [0,1]$. For r a positive real, N is chosen large enough so that ϕ has regularity r (it is possible since it is a property of the Daubechies wavelets that the smoothness of ϕ increases linearly in N). We choose J such that $2^J \geq 2N$ so that the two edges do not interact (no overlap between ϕ^0 and ϕ^1). The construction ensures that ϕ^0 and ϕ^1 are also of regularity r. In the same way, for each level j, the ψ_{jk}'s are dilatations and translations of functions ψ, ψ^0 and ψ^1 with regularity r.

Now we construct a wavelet basis of $L^2([0,1]^2)$ by the tensorial product method (see Meyer (1990) Chapter 3 Section 3). The father wavelet is $\phi \otimes \phi$ and the mother wavelets are $\phi \otimes \psi$, $\psi \otimes \phi$, $\psi \otimes \psi$. A function T in $L^2([0,1]^2)$ can then be written

$$\begin{aligned} T(x,y) = \sum_{k=0}^{2^J-1} \sum_{l=0}^{2^J-1} b_{kl} \phi_k(x) \phi_l(y) + \sum_{j>J} \sum_{k=0}^{2^j-1} \sum_{l=0}^{2^j-1} (a^{(1)}_{jkl} \phi_{jk}(x) \psi_{jl}(y) \\ + a^{(2)}_{jkl} \psi_{jk}(x) \phi_{jl}(y) + a^{(3)}_{jkl} \psi_{jk}(x) \psi_{jl}(y)). \end{aligned}$$

For the sake of simplicity, we adopt the following notation

$$T(x,y) = \sum_{j \geq J} \sum_{(k,l) \in \Lambda_j} a_{jkl} \varphi_{jk}(x) \varphi_{jl}(y).$$

where $\varphi_{jk} = 2^{j/2}\varphi(2^j x - k)$ with $\varphi = \phi, \phi^0, \phi^1, \psi, \psi^0$ or ψ^1 according to the values of j and k. For $j > J$, Λ_j is a set with cardinal 3.2^{2j} and Λ_J is a set with cardinal 2^{2J}. In the sequel we will use the following property of φ deriving from the regularity of the initial Daubechies wavelet: there exists a positive constant k_3 such that

$$\forall u \in \mathbb{R} \quad |\varphi^*(u)| \leq k_3 (u^2 + 1)^{-r/2} \tag{5.4}$$

Now , for $m \geq J$, we can consider the space

$$\mathbb{S}_m = \{T : \mathbb{R}^2 \to \mathbb{R}, \quad T(x,y) = \sum_{j=J}^{m} \sum_{(k,l)\in\Lambda_j} a_{jkl}\varphi_{jk}(x)\varphi_{jl}(y)\}.$$

Note that the functions in $\mathbb{S}_m$ are all supported in the interval $[0,1]^2$. The dimension of the space $\mathbb{S}_m$ is $D_m^2 = 2^{2J} + 3\sum_{j=J+1}^{m} 2^{2j} \in [2^{2m}, 2^{2m+2}]$. We denote by $\mathcal{S}$ the space S_{m_0} with the greatest dimension $D_{m_0}^2 = \mathcal{D}^2$ smaller than $n^{1/(4\gamma+2)}$. It is the maximal space that we will consider. The spaces $\mathbb{S}_m$ have the following properties:

P1 $m' \leq m \Rightarrow \mathbb{S}_{m'} \subset \mathbb{S}_m$

P2 $\| \sum_{jkl} a_{jkl}\varphi_{jk} \otimes \varphi_{jl}\|^2 = \sum_{jkl} a_{jkl}^2$.

This property derives from the orthonormality of the basis.

Now, for all function $t : \mathbb{R} \mapsto \mathbb{R}$, let v_t be the inverse Fourier transform of $t^*/q^*(-.)$, i.e.

$$v_t(x) = \frac{1}{2\pi}\int e^{ixu}\frac{t^*(u)}{q^*(-u)}du.$$

This operator is introduced because it verifies $\mathbb{E}[v_t(Y_k)|X_k] = t(X_k)$ for all function t. We can write the following lemma :

Lemma 5.1 *If $r > \gamma + 2$, there exists $\Phi_1 > 0$ such that*

P3 $\| \sum_{j=J}^{m} \sum_k \varphi_{jk}^2\|_\infty \leq \Phi_1 D_m$

P4 $\| \sum_k |v_{\varphi_{jk}}|^2\|_\infty \leq \Phi_1 (2^j)^{2\gamma+2}$

P5 $\sum_k \|v_{\varphi_{jk}}\|^2 \leq \Phi_1 (2^j)^{2\gamma+1}$

P6 $\| \sum_{kk'} |v_{\varphi_{jk}\varphi_{jk'}}|^2\|_\infty \leq \Phi_1 (2^j)^{2\gamma+3}$

P7 $\sum_{kk'} \int |v_{\varphi_{jk}\varphi_{jk'}}|^2 \leq \Phi_1 (2^j)^{2\gamma+2}$

This lemma is proved in Section 5.6.

5.3.2 Construction of a contrast

We will estimate the transition density of the Markov chain by minimizing a contrast. This section is devoted to the definition of this contrast. We explain here how we can obtain it by considering first the case without noise.

First step: if the X_i's were observed

We present here an heuristic to understand why we choose the contrast, by assuming that the (X_i) are known.

For all function g, the definition of the transition density implies $\mathbb{E}[g(X_{i+1})|X_i] = \int \Pi(X_i, y)g(y)dy$ so that we can write

$$g(X_{i+1}) = \left(\int \Pi(., y)g(y)dy\right)(X_i) + \eta_i$$

where $\eta_i = g(X_{i+1}) - \mathbb{E}[g(X_{i+1})|X_i]$ is a centered process. We recognize then a regression model. A contrast to estimate $\int \Pi(., y)g(y)dy$ is then

$$\gamma_n(u) = \frac{1}{n}\sum_{i=1}^{n}[u^2(X_i) - 2u(X_i)g(X_{i+1})].$$

It is the classic mean square contrast to estimate a regression function. But we want to estimate $\Pi(., y)$ and not only $\int \Pi(., y)g(y)dy$.

A first approach could be to use a basis function (φ_λ) of a subspace of $L^2(A)$ and then to estimate $\int \Pi(., y)\varphi_\lambda(y)dy$ for each λ. This heuristic is detailed in the second chapter. A second approach is to observe that if $\int g^2 = 1$ and $T(x, y) = u(x)g(y)$, then $u(.) = \int T(., y)g(y)dy$. So if $u(.) = \int T(., y)g(y)dy$ estimates $\int \Pi(., y)g(y)dy$, we can hope that T estimates Π. Since $\int T^2(., y)dy = u^2(.)$, the contrast becomes

$$\gamma_n(T) = \frac{1}{n}\sum_{i=1}^{n}[\int T^2(X_i, y)dy - 2T(X_i, X_{i+1})]$$

It is the contrast studied in the second chapter and it allows a good estimation of Π in the case where the Markov chain is observed. We can observe that

$$\mathbb{E}\gamma_n(T) = \int T^2(x, y)f(x)dxdy - 2\int T(x, y)f(x)\Pi(x, y)dxdy = \|T - \Pi\|_f^2 - \|\Pi\|_f^2$$

where f is the density of (X_i) and

$$\|T\|_f = \left(\int T^2(x, y)f(x)dxdy\right)^{1/2}.$$

Then this contrast is an empirical counterpart of the distance $\|T - \Pi\|_f$.

Second step: the X_i's are unknown, the observations are the Y_i's

The aim of this step is to modify the previous contrast, to take into account that the X_i's are not observed. To do this, we use the same technique as in the convolution problem

(see Comte *et al.* (2006b)). Let us denote by F_X the density of (X_i, X_{i+1}) and F_Y the density of (Y_i, Y_{i+1}). We remark that $F_Y = F_X * (q \otimes q)$ and $F_Y^* = F_X^*(q^* \otimes q^*)$ and then

$$\mathbb{E}[T(X_i, X_{i+1})] = \iint TF_X = \frac{1}{2\pi} \iint T^* \overline{F_X^*} = \frac{1}{2\pi} \iint \frac{T^*}{\overline{q^*} \otimes \overline{q^*}} \overline{F_Y^*}$$

by using the Parseval equality. The idea is then to define $V_T^* = T^*/(\overline{q^*} \otimes \overline{q^*})$ so that

$$\mathbb{E}[T(X_i, X_{i+1})] = \frac{1}{2\pi} \iint V_T^* \overline{F_Y^*} = \iint V_T F_Y = \mathbb{E}[V_T(Y_i, Y_{i+1})].$$

Then we will replace the term $T(X_i, X_{i+1})$ in the contrast by $V_T(Y_i, Y_{i+1})$. In the same way, we find an operator Q to replace the term $\int T^2(X_i, y)dy$. More precisely, for all function T, let V_T be the inverse Fourier transform of $T^*/(q^* \otimes q^*)(-.)$, i.e.

$$V_T(x, y) = \frac{1}{4\pi^2} \iint e^{ixu+iyv} \frac{T^*(u, v)}{q^*(-u)q^*(-v)} dudv.$$

Let Q_T be the inverse Fourier transform of $T^*(., 0)/(q^*)(-.)$, i.e.

$$Q_T(x) = \frac{1}{2\pi} \int e^{ixu} \frac{T^*(u, 0)}{q^*(-u)} du.$$

V and Q have been chosen so that the following lemma holds.

Lemma 5.2 *For all* $k \in \{1, \ldots, n+1\}$

1. $\mathbb{E}[V_T(Y_k, Y_{k+1})|X_1, ..., X_{n+1}] = T(X_k, X_{k+1})$
2. $\mathbb{E}[V_T(Y_k, Y_{k+1})] = \iint T(x, y)\Pi(x, y)f(x)dxdy$
3. $\mathbb{E}[Q_T(Y_k)|X_1, ..., X_{n+1}] = \int T(X_k, y)dy$
4. $\mathbb{E}[Q_T(Y_k)] = \iint T(x, y)f(x)dxdy$

Points *1* and *3* are proved in Section 5.6, the other assertions are immediate consequences. Notice that V and Q are strongly linked with v. In particular $V_{s\otimes t}(x, y) = v_s(x)v_t(y)$ and $Q_{s\otimes t}(x) = v_s(x) \int t(y)dy$.

By using the operators V and Q, we define now the contrast, depending only on the observations $Y_1, \ldots, Y_{n+1}$:

$$\gamma_n(T) = \frac{1}{n} \sum_{k=1}^{n} [Q_{T^2}(Y_k) - 2V_T(Y_k, Y_{k+1})].$$

With Lemma 5.2, we compute $\mathbb{E}(\gamma_n(T)) = \iint T^2(x,y)f(x)dxdy - 2\iint T(x,y)\Pi(x,y)f(x)dxdy = \|T-\Pi\|_f^2 - \|\Pi\|_f^2$. So we want to estimate Π by minimizing γ_n. The definition of the contrast leads to the following "empirical norm":

$$\Psi_n(T) = \frac{1}{n}\sum_{k=1}^{n} Q_{T^2}(Y_k).$$

The term empirical norm is used because $\mathbb{E}\Psi_n(T) = \|T\|_f^2$, but Ψ_n is not a norm in the common sense.

5.3.3 Definition of the estimator

We have to minimize the contrast γ_n to find our estimator. By writing $T = \sum_{j=J}^{m}\sum_{(k,l)\in\Lambda_j} a_{jkl}\varphi_{jk}\otimes\varphi_{jl} = \sum_\lambda a_\lambda\omega_\lambda(x,y)$, we obtain

$$\frac{\partial\gamma_n(T)}{\partial a_{\lambda_0}} = \frac{2}{n}\sum_{i=1}^{n}\left(\sum_\lambda a_\lambda Q_{\omega_\lambda\omega_{\lambda_0}}(Y_i) - V_{\omega_{\lambda_0}}(Y_i,Y_{i+1})\right).$$

And then, by denoting by A_m the vector of the coefficients a_λ of T,

$$\forall\lambda_0 \quad \frac{\partial\gamma_n(t)}{\partial a_{\lambda_0}} = 0 \quad\Longleftrightarrow\quad G_m A_m = Z_m \tag{5.5}$$

where

$$G_m = \left[\frac{1}{n}\sum_{i=1}^{n} Q_{\omega_\lambda\omega_\mu}(Y_i)\right]_{\lambda,\mu}, \quad Z_m = \left[\frac{1}{n}\sum_{i=1}^{n} V_{\omega_\lambda}(Y_i,Y_{i+1})\right]_\lambda$$

But the matrix G_m is not necessarily invertible. That is why we introduce the set

$$\Gamma = \left\{\min \mathrm{Sp}(G_m) \geq \frac{2}{3}f_0\right\} \tag{5.6}$$

where Sp denotes the spectrum, i.e. the set of the eigenvalues of the matrix and f_0 is the lower bound of f on A_1. On Γ, G_m is invertible and γ_n is convex so that the minimization of γ_n is equivalent to Equation (5.5) and admits the solution $A_m = G_m^{-1}Z_m$. Now we can define

$$\hat{\Pi}_m = \begin{cases} \arg\min_{T\in\mathbb{S}_m} \gamma_n(T) & \text{on } \Gamma, \\ 0 & \text{on } \Gamma^c. \end{cases}$$

Remark 5.2 *The construction of $\hat{\Pi}_m$ described here requires the knowledge of f_0. Nevertheless, when f_0 is unknown, we can replace it by an estimator $\hat{f}_0$ defined as the minimum of an estimator of f (for an estimator of the density of a hidden Markov chain, see chapter 4). The result is then unchanged if f is regular enough and the mixing rate strong enough.*

We have then an estimator of Π for all $\mathbb{S}_m$. But we have to choose the best model m to obtain an adaptive estimator, i.e. an estimator which achieves the best rate of convergence whatever the regularity of Π. So we set

$$\hat{m} = \arg \min_{m \in \mathcal{M}_n} \{\gamma_n(\hat{\Pi}_m) + \text{pen}(m)\}$$

where pen is a penalty function to be specified later and

$$\mathcal{M}_n = \{m \geq J, D_m^{4\gamma+2} \leq n\}.$$

Then we can define our definitive estimator:

$$\tilde{\Pi} = \begin{cases} \hat{\Pi}_{\hat{m}} & \text{if } \|\hat{\Pi}_{\hat{m}}\| \leq k_n \quad \text{with } k_n = n^{1/2}, \\ 0 & \text{else.} \end{cases}$$

5.4 Result

5.4.1 Risk and rate of convergence

For a function G and a subspace $\mathbb{S}$, we define

$$d_A(G, \mathbb{S}) = \inf_{T \in \mathbb{S}} \|G - T\|_A.$$

For each estimator $\hat{\Pi}_m$, we have the following decomposition of the risk.

Proposition 5.1 *We consider a Markov chain and a noise satisfying Assumptions H1-H2b-H3a,b-H4-H5 with $\gamma \geq 3/4$. For m fixed in $\mathcal{M}_n$, we consider $\hat{\Pi}_m$ the estimator of the transition density Π previously described. Then there exists $C > 0$ such that*

$$\mathbb{E}\|\hat{\Pi}_m - \Pi\|_A^2 \leq C \left\{ d_A^2(\Pi, S_m) + \frac{D_m^{4\gamma+2}}{n} \right\}$$

We do not prove this proposition because this result is included in those of Theorem 5.1 below, which is proved in Section 5.6.

Now if Π belongs to a Besov space with regularity α, it is a common approximation property of the wavelets spaces that $d_A^2(\Pi, S_m) \leq C D_m^{-2\alpha}$. So, choosing m_1 such that $D_{m_1} = \lfloor n^{1/(2\alpha+4\gamma+2)} \rfloor$, we can obtain the minimum risk

$$\mathbb{E}\|\hat{\Pi}_{m_1} - \Pi\|_A^2 \leq C n^{-\frac{2\alpha}{2\alpha+4\gamma+2}}.$$

But this choice of m_1 is impossible if α is unknown (it is *a priori* the case since Π is unknown). That is why we have built an adaptive estimator via model selection. For our estimator $\tilde{\Pi}$, we can state the following theorem.

Theorem 5.1 *We consider a Markov chain and a noise satisfying Assumptions H1-H2b-H3a,b-H4-H5 with $\gamma > 3/4$. We consider $\tilde{\Pi}$ the estimator of the transition density Π previously described with $r > 2\gamma + 3/2$ and*

$$\mathrm{pen}(m) = K\frac{D_m^{4\gamma+2}}{n} \quad \textit{for some } K > K_0$$

where K_0 is a constant depending on Φ_1, $\|q\|_\infty$ and f_0. Then there exists $C' > 0$ such that

$$\mathbb{E}\|\tilde{\Pi} - \Pi\|_A^2 \leq C \inf_{m\in\mathcal{M}_n}\{d_A^2(\Pi, S_m) + \mathrm{pen}(m)\} + \frac{C'}{n}$$

with $C = \max(2 + 72f_0^{-1}\|f\|_{\infty,A_1}(1 + 2\|\Pi\|_A^2), 12f_0^{-1}(1 + 2\|\Pi\|_A^2))$.

Note that this result is non-asympotic, it is an advantage of the least square method with respect to a quotient method.

All constants on which the penalty depends do not have the same status. The constants Φ_1 and $\|q\|_\infty$ are known, since the wavelets basis and the noise distribution are known. The constant f_0 is unknown but it can be estimated (see Remark 5.2). Then, even if it means replacing f_0 by an estimator $\hat{f}_0$, the penalty is computable, especially since the dependence coefficients of the sequence do not appear at all in the penalty.

The condition $\gamma > 3/4$ is due to an additional term of order $D_m^{2\gamma+7/2}/n$ (coming from the term $(1/n)\sum_{i=1}^n Q_{T^2}(Y_i)$ in the contrast) inside the penalty. If $\gamma > 3/4$, then $2\gamma + 7/2 < 4\gamma + 2$ and $D_m^{4\gamma+2}/n$ is the dominant term. If $\gamma = 3/4$, the result is still true but the constant in the penalty also depends on $\|\Pi\|_A$. In the other cases the estimation is possible but the term $D_m^{2\gamma+4}/n$ is not negligible any more and the order of the variance (and consequently the rate of convergence) must be changed. This constraint $\gamma > 3/4$ is not restrictive since γ must be larger than $1/2$ in order that q be square integrable. Moreover in the case of a Gamma noise, q is not bounded if $\gamma < 1$.

We can now evaluate the rate of convergence of our estimator.

Corollary 5.1 *We suppose that the restriction of Π to A belongs to the Besov space $B_{2,\infty}^\alpha(A)$ with $r > \alpha - 1$. Then, under the assumptions of Theorem 5.1,*

$$\mathbb{E}\|\tilde{\Pi} - \Pi\|_A^2 = O(n^{-\frac{2\alpha}{2\alpha+4\gamma+2}}).$$

There is to our knowledge no lower bound for our precise estimation problem and such a study is beyond the scope of the present work. But Clémençon (2003) proves the optimality of the rate $n^{-\frac{2\alpha}{2\alpha+4\gamma+2}}$ in the case where f belongs to $B_{2,\infty}^\alpha(\mathbb{R})$ and $f\Pi$ belongs to $B_{2,\infty}^\alpha(\mathbb{R}^2)$.

Nevertheless we remark that we obtain then the same rate of convergence with $\tilde{\Pi}$ than those obtained with $\hat{\Pi}_{m_1}$ where $D_{m_1} = \lfloor n^{1/(2+4\gamma+2\alpha)}\rfloor$, but whitout requiring the knowledge of α. That is why we can assert that this estimator is adaptive. Moreover

our estimator is better than the one of Clémençon (2003), which achieves only the rate $(\log(n)/n)^{\frac{2\alpha}{2\alpha+4\gamma+2}}$. It is also an improvement of the result of the chapter 4 because this rate is obtained without requiring that f has a regularity α.

5.4.2 Sketch of proof of Theorem 5.1

We give in this section a sketch of proof of Theorem 5.1.

Let $m \in \mathcal{M}_n$. We denote by Π_m the orthogonal projection of Π on $\mathbb{S}_m$. We have the following bias-variance decomposition

$$\mathbb{E}\|\tilde{\Pi} - \Pi\|_A^2 = \mathbb{E}\|\tilde{\Pi} - \Pi_m\|_A^2 + \|\Pi_m - \Pi\|_A^2$$

The term $\|\tilde{\Pi} - \Pi_m\|_A^2$ can be written in the following way

$$\begin{aligned}\|\tilde{\Pi} - \Pi_m\|_A^2 &= \|\tilde{\Pi} - \Pi_m\|_A^2 \mathbb{1}_{\{\|\hat{\Pi}_{\hat{m}}\| \leq k_n\}} + \|\tilde{\Pi} - \Pi_m\|_A^2 \mathbb{1}_{\{\|\hat{\Pi}_{\hat{m}}\| > k_n\}} \\ &\leq \|\hat{\Pi}_{\hat{m}} - \Pi_m\|_A^2 + \|\Pi_m\|_A^2 \mathbb{1}_{\{\|\hat{\Pi}_{\hat{m}}\| > k_n\}}\end{aligned}$$

since $\tilde{\Pi} = 0$ on the set $\{\|\hat{\Pi}_{\hat{m}}\| > k_n\}$ and $\tilde{\Pi} = \hat{\Pi}_{\hat{m}}$ on the complementary. The term $\|\Pi_m\|_A^2 \mathbb{1}_{\{\|\hat{\Pi}_{\hat{m}}\| > k_n\}}$ is easy to deal with, the main term is $\|\hat{\Pi}_{\hat{m}} - \Pi_m\|_A^2$. But, on Γ, the definitions of $\hat{\Pi}_m$ and $\hat{m}$ lead to the inequality

$$\gamma_n(\hat{\Pi}_{\hat{m}}) + \text{pen}(\hat{m}) \leq \gamma_n(\Pi_m) + \text{pen}(m). \tag{5.7}$$

Letting $Z_{n,m}(T) = \frac{1}{n}\sum_{k=1}^n [V_T(Y_k, Y_{k+1}) - Q_{T\Pi_m}(Y_k)]$, a fast computation gives

$$\gamma_n(\hat{\Pi}_{\hat{m}}) - \gamma_n(\Pi_m) = \Psi_n(\hat{\Pi}_{\hat{m}} - \Pi_m) - 2Z_{n,m}(\hat{\Pi}_{\hat{m}} - \Pi_m)$$

so that (5.7) becomes

$$\begin{aligned}\Psi_n(\hat{\Pi}_{\hat{m}} - \Pi_m) &\leq 2Z_{n,m}(\hat{\Pi}_{\hat{m}} - \Pi_m) + \text{pen}(m) - \text{pen}(\hat{m}) \\ &\leq 2\|\hat{\Pi}_{\hat{m}} - \Pi_m\|_f \sup_{T \in B_f(m,\hat{m})} Z_{n,m}(T) + \text{pen}(m) - \text{pen}(\hat{m})\end{aligned}$$

where $B_f(m, \hat{m}) = \{T \in S_m + S_{\hat{m}}, \quad \|T\|_f = 1\}$. The main steps of the proof are then

1. to control the term $\sup_{T \in B_f(m,\hat{m})} Z_{n,m}(T)$,
2. to link the empirical "norm" Ψ_n with the L^2 norm $\|.\|_A$.

• To deal with the supremum of the empirical process $Z_{n,m}(T)$, we will use an inequality of Talagrand stated in Lemma 5.5 (Section 5.6.8). This inequality is very powerful but can be applied only to sum of independent random variables. That is why we split $Z_{n,m}(T)$ into three processes plus a bias term.

$$Z_{n,m}(T) = Z_n^{(1)}(T) - Z_n^{(2)}(T) + Z_n^{(3)}(T) + \iint T(x,y)(\Pi - \Pi_m)(x,y)f(x)dxdy$$

with

$$\begin{cases} Z_n^{(1)}(T) = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} V_T(Y_k, Y_{k+1}) - \mathbb{E}[V_T(Y_k, Y_{k+1})|X_1, \ldots, X_{n+1}] \\ Z_n^{(2)}(T) = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} Q_{T\Pi_m}(Y_k) - \mathbb{E}[Q_{T\Pi_m}(Y_k)] \\ Z_n^{(3)}(T) = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} T(X_k, X_{k+1}) - \mathbb{E}[T(X_k, X_{k+1})] \end{cases}$$

For the first process $Z_n^{(1)}$, we get back to independent variables by remarking that, conditionally to $X_1, \ldots, X_{n+1}$, the couples $(Y_{2i-1}, Y_{2i}), i = 1, ..., (n+1)/2$, are independent (see Proposition 5.3).

For the other processes, we use the mixing assumption H2b to build auxiliary variables X_i^* which are approximation of the X_i's and which constitute independent clusters of variables (see Proposition 5.4).

- To pass from Ψ_n to the L^2 norm, we introduce the following set

$$\Delta = \{\forall T \in \mathcal{S} \quad \|T\|_f^2 \leq \frac{3}{2}\Psi_n(T)\}$$

We can easily prove (see Section 5.6.3) that $\Delta \subset \Gamma$. Then,

$$\|\hat{\Pi}_{\hat{m}} - \Pi_m\|_A \mathbb{1}_\Delta \leq \frac{3}{2} f_0^{-1} \Psi_n(\hat{\Pi}_{\hat{m}} - \Pi_m)\mathbb{1}_\Gamma$$

It remains to prove that $P(\Delta^c) = P(\exists T \in \mathcal{S}, \Psi_n(T) < (2/3)\mathbb{E}[\Psi_n(T)])$ is small enough. It is done in Proposition 5.2.

5.5 Simulations

To illustrate the method, we compute our estimator $\tilde{\Pi}$ for different Markov processes with known transition density. The estimation procedure contains several Fourier transforms, which can seem heavy, but the computation of $v_{\varphi_{jk}}$ for all the basis functions can be done beforehand. Next, to compute $\tilde{\Pi}$ from data $Y_1, \ldots, Y_{n+1}$, we use the following steps (see Section 5.3.3):

- For each m, compute matrices G_m and Z_m,
- Deduce the matrix A_m,

- Select the $\hat{m}$ which minimize $\gamma_n(\hat{\Pi}_m) + \text{pen}(m) = -{}^t A_m Z_m + \text{pen}(m)$,
- Compute $\tilde{\Pi}$ using matrix $A_{\hat{m}}$.

We consider the different kinds of Markov chains described in Section 1.5 of Chapter 1. We consider two different noises:

Laplace noise In this case, the density of ε_i is given by

$$q(x) = \frac{\lambda}{2} e^{-\lambda|x|}; \quad q^*(x) = \frac{\lambda^2}{\lambda^2 + x^2}; \quad \lambda = 5.$$

The smoothness parameter is $\gamma = 2$ and we choose

$$\text{pen}(m) = \frac{1}{n} \left(\frac{\lambda}{2}\right)^2 \left(\frac{D_m}{4}\right)^{10}.$$

n	50	100	250	500	1000	noise
AR(i)	0.579	0.407	0.270	0.230	0.209	Lapl
	0.599	0.480	0.313	0.272	0.245	Gauss
AR(ii)	0.389	0.294	0.195	0.155	0.139	Lapl
	0.339	0.304	0.280	0.273	0.271	Gauss
$\sqrt{\text{CIR}}$	0.171	0.138	0.123	0.118	0.111	Lapl
	0.199	0.169	0.150	0.142	0.139	Gauss
CIR(iii)	0.420	0.345	0.237	0.195	0.175	Lapl
	0.337	0.302	0.276	0.245	0.209	Gauss
CIR(iv)	0.525	0.403	0.337	0.304	0.292	Lapl
	0.369	0.345	0.344	0.327	0.321	Gauss
ARCH	0.312	0.287	0.261	0.185	0.150	Lapl
	0.337	0.319	0.296	0.290	0.183	Gauss

Table 5.1: MISE $\mathbb{E}\|\Pi - \tilde{\Pi}\|^2$ averaged over $N = 200$ samples.

Gaussian noise In this case, the density of ε_i is given by

$$q(x) = \frac{1}{\lambda\sqrt{2\pi}} e^{-\frac{x^2}{2\lambda^2}}; \quad q^*(x) = e^{-\frac{\lambda^2 x^2}{2}}; \quad \lambda = 0.3.$$

This noise does not verify Assumption H1 but it is interesting to see if this assumption is also necessary for practical purposes. Given the exponential regularity of this noise, we consider the following penalty

$$\text{pen}(m) = \frac{5}{n} \exp(\lambda^2 D_m^2).$$

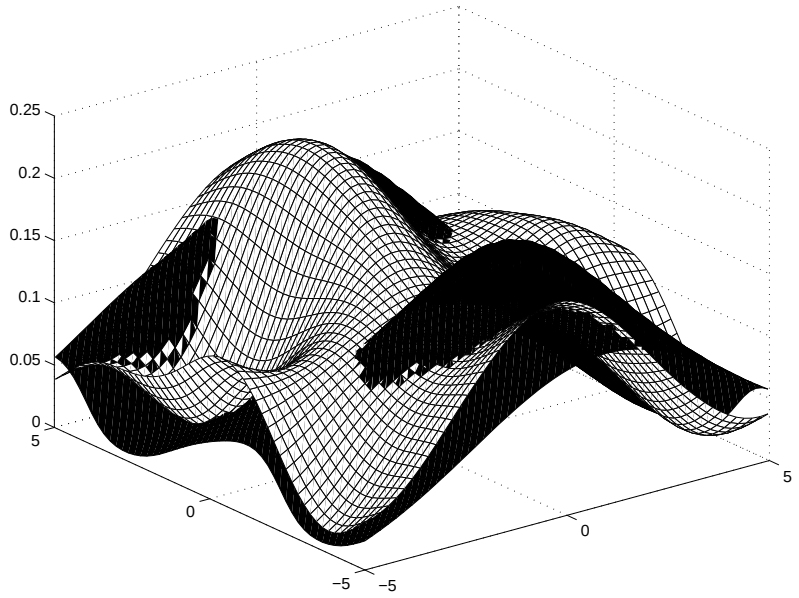


Figure 5.1: True Π (black) and estimator $\tilde{\Pi}$ (white) for process ARCH observed through a Laplace noise, $n = 500$

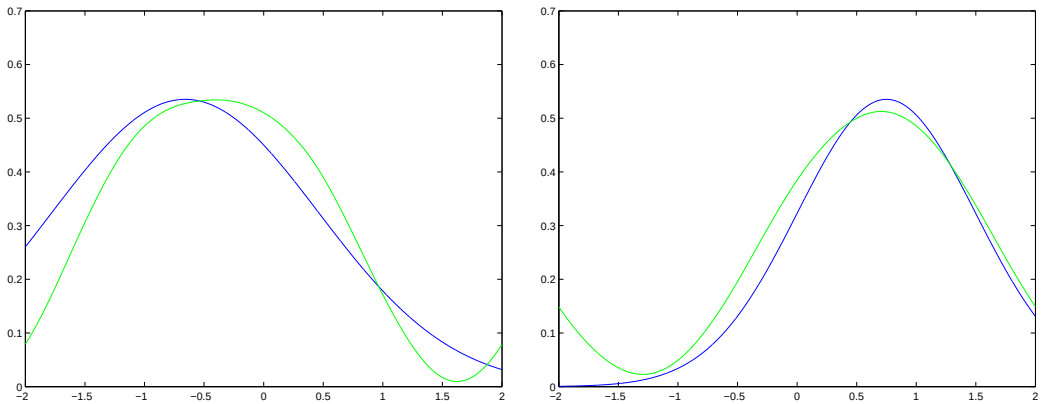


Figure 5.2: Sections for the process AR(i) observed through a Laplace noise, $n = 500$

Table 5.1 presents the L^2 risk of our estimator of the transition density for the 6 Markov chains and the 2 noises. These results can be compared with those of Chapter 2 (Table 2.2) where the processes were directly observed, i.e. without noise. The risk values are then higher in our case, but with the same order, which is satisfactory. It is noticeable that the estimation works pretty much the same with the Gaussian noise, but with a slower decrease of the risk. It is a classical phenomenon in deconvolution problems, since the Gaussian noise is much more regular than the Laplace noise.

Figure 5.5 allows to visualize the result for the process ARCH observed through a Laplace noise: the surfaces $z = \Pi(x, y)$ and $z = \tilde{\Pi}(x, y)$ are presented. We also give figures of sections of this kind of surfaces. We can see on Figure 5.5 the curves $z = \Pi(x, -0.44)$ versus $z = \tilde{\Pi}(x, -0.44)$ and the curves $z = \Pi(1.12, y)$ versus $z = \tilde{\Pi}(1.12, y)$ for the process AR(i). Generally, for a multidimensional estimation, the mixed control of the directions does not allow to do as well as a classical one-dimensionnal function estimation. Nevertheless the curves are here very close.

5.6 Detailed proofs

5.6.1 Proofs of Lemma 5.1

In these proofs, we shall use the following formula, directly deduced from the definition of φ_{jk}

$$(\varphi_{jk})^*(u) = 2^{-j/2} e^{iuk/2^j} \varphi^*(u/2^j).$$

Proof of Property P3

It is known that, for all (a_k),

$$\| \sum_k a_k \varphi_{jk} \|_\infty \leq C(\varphi) 2^{j/2} |(a_k)|_\infty.$$

Therefore

$$| \sum_k \varphi_{jk}^2(x)| \leq C(\varphi) 2^{j/2} \|\varphi_{jk}\|_\infty \leq C'(\varphi) 2^j \tag{5.8}$$

and $| \sum_{jk} \varphi_{jk}^2(x)| \leq \sum_{j=J}^m C'(\varphi) 2^j \leq 2C'(\varphi) D_m$. Then the property holds if $\Phi_1 \geq 2C'(\varphi)$.

Proof of Property P4

According to the definition of v_t,

$$\begin{aligned} |v_{\varphi_{jk}}(x)| &\leq \frac{1}{2\pi} \int \frac{|\varphi_{jk}^*(u)|}{|q^*(-u)|} du \leq \frac{1}{2\pi} \int 2^{-j/2} \frac{|\varphi^*(u/2^j)|}{|q^*(-u)|} du \\ &\leq \frac{2^{j/2}}{2\pi} \int \frac{|\varphi^*(v)|}{|q^*(-v2^j)|} dv, \end{aligned}$$

by change of variable. Next, it follows from Assumption H5 that

$$\begin{aligned} |v_{\varphi_{jk}}(x)| &\leq \frac{2^{j/2}}{2\pi}\int |\varphi^*(v)|k_0^{-1}((v2^j)^2+1)^{\gamma/2}dv \\ &\leq \frac{k_0^{-1}}{2\pi}(2^j)^{\gamma+1/2}\int |\varphi^*(v)|(v^2+1)^{\gamma/2}dv \leq C_{1,\gamma}\frac{k_0^{-1}}{2\pi}(2^j)^{\gamma+1/2}, \end{aligned}$$

using Lemma 5.4 since $r > \gamma + 1$. Then for all x,

$$\sum_k |v_{\varphi_{jk}}(x)|^2 \leq 3.2^j C_{1,\gamma}^2 \frac{k_0^{-2}}{4\pi^2}(2^j)^{2\gamma+1} \leq 3C_{1,\gamma}^2\frac{k_0^{-2}}{4\pi^2}(2^j)^{2\gamma+2}$$

that establishes P4 with $\Phi_1 \geq 3C_{1,\gamma}^2 k_0^{-2}/(4\pi^2)$.

Proof of Property P5

To prove P5, we apply the Parseval equality:

$$\begin{aligned} \int |v_{\varphi_{jk}}|^2 &= \frac{1}{2\pi}\int |v^*_{\varphi_{jk}}|^2 = \frac{1}{2\pi}\int \frac{2^{-j}|\varphi^*(u/2^j)|^2}{|q^*(-u)|^2}du \\ &= \frac{1}{2\pi}\int \frac{|\varphi^*(v)|^2}{|q^*(-v2^j)|^2}dv. \end{aligned}$$

Assumption H5 then gives

$$\begin{aligned} \int |v_{\varphi_{jk}}|^2 &\leq \frac{1}{2\pi}\int |\varphi^*(v)|^2 k_0^{-2}((v2^j)^2+1)^{\gamma}dv \\ &\leq \frac{k_0^{-2}}{2\pi}(2^j)^{2\gamma}\int |\varphi^*(v)|^2(v^2+1)^{\gamma}dv \leq C_{2,2\gamma}\frac{k_0^{-2}}{2\pi}(2^j)^{2\gamma}, \end{aligned}$$

because $2r > 2\gamma + 1$. And finally

$$\sum_k \int |v_{\varphi_{jk}}|^2 \leq 3.2^j C_{2,2\gamma}\frac{k_0^{-2}}{2\pi}(2^j)^{2\gamma}$$

Then P5 holds with $\Phi_1 \geq 3C_{2,2\gamma}k_0^{-2}/(2\pi)$.

Proof of Property P6

We begin with computing $|v_{\varphi_{jk}\varphi_{jk'}}(x)|$ by using the fact that $(\varphi_{jk}\varphi_{jk'})^*$ is equal to the convolution product $\varphi^*_{jk} * \varphi^*_{jk'}$.

$$\begin{aligned} |v_{\varphi_{jk}\varphi_{jk'}}(x)| &\leq \frac{1}{2\pi}\int \frac{|(\varphi_{jk}\varphi_{jk'})^*|(u)}{|q^*(-u)|}du \leq \frac{1}{2\pi}\iint \frac{|\varphi^*_{jk}(v)||\varphi^*_{jk'}(u-v)|}{|q^*(-u)|}dudv \\ &\leq \frac{1}{2\pi}2^{-j}\iint \frac{|\varphi^*(v/2^j)||\varphi^*((u-v)/2^j)|}{|q^*(-u)|}dudv. \end{aligned}$$

Now a change of variables gives:

$$\begin{aligned} |v_{\varphi_{jk}\varphi_{jk'}}(x)| &\leq \frac{1}{2\pi}\iint 2^j \frac{|\varphi^*(y)||\varphi^*(x-y)|}{|q^*(-2^j x)|}dxdy \\ &\leq \frac{k_0^{-1}}{2\pi}2^j(2^j)^\gamma \iint |\varphi^*(y)\varphi^*(x-y)|(x^2+1)^{\gamma/2}dxdy. \end{aligned}$$

Then Lemma 5.4 shows that

$$\begin{aligned} |v_{\varphi_{jk}\varphi_{jk'}}(x)| &\leq \frac{k_0^{-1}}{2\pi}(2^j)^{\gamma+1}C_r\Big[\int_{|x|>1}|x|^{1-r}(x^2+1)^{\gamma/2}dx \\ &\quad + \int_{|x|\leq 1}(x^2+1)^{\gamma/2}dx\Big]. \end{aligned}$$

Then, since $r > \gamma + 2$, there exists $C > 0$ such that

$$|v_{\varphi_{jk}\varphi_{jk'}}(x)| \leq C(2^j)^{\gamma+1}.$$

Now we observe that φ_{jk} and $\varphi_{jk'}$ have disjoint supports if $k + N \leq k' - N + 1$ or $k' + N \leq k - N + 1$. Then

$$\sum_{k,k'}|v_{\varphi_{jk}\varphi_{jk'}}(x)|^2 \leq \sum_{k'}\sum_{k=k'-2N+2}^{k'+2N-2} C^2(2^j)^{2\gamma+2} \leq 3.2^j(4N-3)C^2(2^j)^{2\gamma+2}.$$

Hence, if $\Phi_1 \geq 3(4N-3)C^2$, P6 is proved.

Proof of Property P7

Applying Parseval's equality,

$$\begin{aligned} \int |v_{\varphi_{jk}\varphi_{jk'}}|^2 &= \frac{1}{2\pi}\int \frac{|(\varphi_{jk}\varphi_{jk'})^*|^2(u)}{|q^*(-u)|^2}du \\ &= \frac{2^j}{2\pi}\int \frac{|(\varphi_{jk}\varphi_{jk'})^*|^2(2^j v)}{|q^*(-2^j v)|^2}dv. \end{aligned}$$

But

$$\begin{aligned}|(\varphi_{jk}\varphi_{jk'})^*(2^j v)| &\leq \int |\varphi^*_{jk}(z)||\varphi^*_{jk'}(2^j v - z)|dz \\ &\leq 2^{-j}\int |\varphi^*(z/2^j)||\varphi^*((2^j v - z)/2^j)|dz \\ &\leq \int |\varphi^*(y)||\varphi^*(v-y)|dy.\end{aligned}$$

We use Lemma 5.4 to write

$$|(\varphi_{jk}\varphi_{jk'})^*(2^j v)| \leq C_r\left[|v|^{1-r}\mathbb{1}_{|v|>1} + \mathbb{1}_{|v|\leq 1}\right]. \tag{5.9}$$

Then, it follows that

$$\begin{aligned}&\int |v_{\varphi_{jk}\varphi_{jk'}}|^2 \leq \frac{2^j}{2\pi}\int \frac{C_r^2(|v|^{2(1-r)}\mathbb{1}_{|v|>1} + \mathbb{1}_{|v|\leq 1})}{|q^*(-2^j v)|^2}dv \\ &\leq \frac{k_0^{-2}2^j}{2\pi}C_r^2\int (|v|^{2(1-r)}\mathbb{1}_{|v|>1} + \mathbb{1}_{|v|\leq 1})((2^j v)^2+1)^\gamma dv \\ &\leq \frac{k_0^{-2}2^j}{2\pi}C_r^2 2^{2j\gamma}\left(\int_{|v|>1}|v|^{2(1-r)}(v^2+1)^\gamma dv + \int_{|v|\leq 1}(v^2+1)^\gamma dv\right) \\ &\leq C(2^j)^{2\gamma+1}.\end{aligned}$$

We now sum this quantity for all k, k' by taking account of the superposition of the supports.

$$\begin{aligned}&\sum_{k,k'}\int |v_{\varphi_{jk}\varphi_{jk'}}|^2 = \sum_{k'}\sum_{k=k'-2N+2}^{k'+2N-2}\int |v_{\varphi_{jk}\varphi_{jk'}}|^2 \\ &\leq 3.2^j(4N-3)C(2^j)^{2\gamma+1} \leq 3C(4N-3)(2^j)^{2\gamma+2}\end{aligned}$$

which concludes the proof, if $\Phi_1 \geq 3C(4N-3)$.

□

5.6.2 Proof of Lemma 5.2

We have already proved the first two points in Chapter 4 Section 4.7.1.

We proceed in a similar way for Q. Since $Q_T(Y_k) = 1/2\pi\int e^{iY_k u}T^*(u,0)/q^*(-u)du$, then

$$\mathbb{E}[Q_T(Y_k)|X_1,\ldots,X_{n+1}] = \frac{1}{2\pi}\int \mathbb{E}[e^{iY_k u}|X_1,\ldots,X_{n+1}]\frac{T^*(u,0)}{q^*(-u)}du.$$

By using the independence between (X_i) and (ε_i), we compute

$$\mathbb{E}[e^{iY_k u}|X_1,\ldots,X_{n+1}] = \mathbb{E}[e^{iX_k u}e^{i\varepsilon_k u}|X_1,\ldots,X_{n+1}] = e^{iX_k u}\mathbb{E}[e^{i\varepsilon_k u}] = e^{iX_k u}q^*(-u).$$

Thus

$$\mathbb{E}[Q_T(Y_k)|X_1,\dots,X_{n+1}] = \frac{1}{2\pi}\int e^{iX_k u}q^*(-u)\frac{T^*(u,0)}{q^*(-u)}du = \frac{1}{2\pi}\int e^{iX_k u}T^*(u,0)du.$$

By denoting by T_y the function $x \mapsto T_y(x) = T(x,y)$, we obtain

$$T^*(u,0) = \iint e^{-ixu}T_y(x)dxdy = \int T_y^*(u)dy$$

and then

$$\begin{aligned}\frac{1}{2\pi}\int e^{iX_k u}T^*(u,0)du &= \frac{1}{2\pi}\iint e^{iX_k u}T_y^*(u)dydu \\ &= \int T_y(X_k)dy = \int T(X_k,y)dy.\end{aligned} \tag{5.10}$$

□

5.6.3 Proof of Theorem 5.1

We start with introducing some auxiliary variables whose existence is ensured by Assumption H2b of mixing. In the case of arithmetical mixing, since $\theta > 8$, there exists a real c such that $0 < c < 3/8$ and $c\theta > 3$. We set in this case $q_n = \lfloor n^c \rfloor$. In the case of geometrical mixing, we set $q_n = \lfloor c\log(n) \rfloor$ where c is a real larger than $3/\theta$.

For the sake of simplicity, we suppose that $n+1 = 2p_nq_n$, with p_n an integer. Let for $l = 0,\dots,p_n-1$, $A_l = (X_{2lq_n+1},...,X_{(2l+1)q_n})$, $B_l = (X_{(2l+1)q_n+1},...,X_{(2l+2)q_n})$. As in Viennet (1997), by using Berbee's coupling Lemma, we can build a sequence (A_l^*) such that

$$\begin{cases} A_l \text{ and } A_l^* \text{ have the same distribution,} \\ A_l^* \text{ and } A_{l'}^* \text{ are independent if } l \neq l', \\ P(A_l \neq A_l^*) \leq \beta_{q_n}. \end{cases} \tag{5.11}$$

In the same way, we build (B_l^*) and we define for any $l \in \{0,\dots,p_n-1\}$,
$A_l^* = (X^*_{2lq_n+1},...,X^*_{(2l+1)q_n})$, $B_l^* = (X^*_{(2l+1)q_n+1},...,X^*_{(2l+2)q_n})$ so that the sequence $(X_1^*,\dots,X_n^*)$ is well defined. We can now define

$$\Omega_X^* = \{\forall i, 1 \leq i \leq n+1 \quad X_i = X_i^*\}.$$

Let us recall that $\mathcal{S}$ is the space $\mathbb{S}_m$ with maximal dimension $\mathcal{D}^2 \leq n^{\frac{1}{4\gamma+2}}$. We now adopt the notations

$$\Delta = \{\forall T \in \mathcal{S} \quad \|T\|_f^2 \leq \frac{3}{2}\Psi_n(T)\}; \qquad \Omega = \Delta \cap \Omega_X^*.$$

Let us fix $m \in \mathcal{M}_n$. We denote by Π_m the orthogonal projection of Π on $\mathbb{S}_m$. Then we have the decomposition

$$\begin{aligned}\mathbb{E}\|\tilde{\Pi}-\Pi\|_A^2 &\leq 2\mathbb{E}\left(\|\tilde{\Pi}-\Pi_m\|^2\mathbb{1}_\Omega\mathbb{1}_{\|\hat{\Pi}_{\hat{m}}\|\leq k_n}\right)+2\mathbb{E}\left(\|\tilde{\Pi}-\Pi_m\|^2\mathbb{1}_\Omega\mathbb{1}_{\|\hat{\Pi}_{\hat{m}}\|>k_n}\right)\\ &\quad +2\mathbb{E}\left(\|\tilde{\Pi}-\Pi_m\|^2\mathbb{1}_{\Omega^c}\right)+2\|\Pi_m-\Pi\|_A^2\\ &\leq 2\mathbb{E}\left(\|\hat{\Pi}_{\hat{m}}-\Pi_m\|^2\mathbb{1}_\Omega\right)+2\|\Pi_m\|^2\mathbb{E}\left(\mathbb{1}_\Omega\mathbb{1}_{\|\hat{\Pi}_{\hat{m}}\|>k_n}\right)\\ &\quad +2\mathbb{E}\left([2\|\tilde{\Pi}\|^2+2\|\Pi_m\|^2]\mathbb{1}_{\Omega^c}\right)+2\|\Pi_m-\Pi\|_A^2.\end{aligned}$$

Now, using the Markov inequality and the definition of $\tilde{\Pi}$,

$$\begin{aligned}\mathbb{E}\|\tilde{\Pi}-\Pi\|_A^2 \leq 2\mathbb{E}\left(\|\hat{\Pi}_{\hat{m}}-\Pi_m\|^2\mathbb{1}_\Omega\right)+2\|\Pi\|_A^2\frac{\mathbb{E}(\|\hat{\Pi}_{\hat{m}}\|^2\mathbb{1}_\Omega)}{k_n^2}\\ +4(k_n^2+\|\Pi\|_A^2)\mathbb{E}\left(\mathbb{1}_{\Omega^c}\right)+2\|\Pi_m-\Pi\|_A^2.\end{aligned}$$

But $\mathbb{E}(\|\hat{\Pi}_{\hat{m}}\|^2\mathbb{1}_\Omega) \leq 2\mathbb{E}(\|\hat{\Pi}_{\hat{m}}-\Pi_m\|^2\mathbb{1}_\Omega)+2\|\Pi_m\|^2$ and $k_n=\sqrt{n}$, so

$$\begin{aligned}\mathbb{E}\|\tilde{\Pi}-\Pi\|_A^2 \leq 2\mathbb{E}\left(\|\hat{\Pi}_{\hat{m}}-\Pi_m\|^2\mathbb{1}_\Omega\right)(1+2\|\Pi\|_A^2)+\frac{4\|\Pi\|_A^4}{n}\\ +4(n+\|\Pi\|_A^2)P(\Omega^c)+2\|\Pi_m-\Pi\|_A^2.\end{aligned}$$

We now state the following proposition :

Proposition 5.2 *There exists $C_0>0$ such that*

$$P(\Omega^c) \leq \frac{C_0}{n^2}.$$

Hence

$$\begin{aligned}\mathbb{E}\|\tilde{\Pi}-\Pi\|_A^2 \leq 2\|\Pi_m-\Pi\|_A^2+2\mathbb{E}\left(\|\hat{\Pi}_{\hat{m}}-\Pi_m\|^2\mathbb{1}_\Omega\right)(1+2\|\Pi\|_A^2)\\ +\frac{4}{n}(\|\Pi\|_A^4+C_0(1+\|\Pi\|_A^2)).\end{aligned} \tag{5.12}$$

Now we have to bound $\mathbb{E}\left(\|\hat{\Pi}_{\hat{m}}-\Pi_m\|^2\mathbb{1}_\Omega\right)$. The estimators $\hat{\Pi}_m$ are defined by minimization of the contrast on a set Γ defined in (5.6). Let us prove that this set Γ includes Ω. More precisely, we prove that $\Delta \subset \Gamma$. For $T=\sum_\lambda a_\lambda\omega_\lambda \in \mathbb{S}_m$, the vector $A_m=(a_\lambda)$ of its coefficients in the basis $(\omega_\lambda(x,y))$ verifies $\Psi_n(T)={}^tA_mG_mA_m$. Then, on Δ,

$${}^tA_mG_mA_m \geq \frac{2}{3}\|T\|_f^2 \geq \frac{2}{3}f_0\|T\|^2.$$

Now, using P2, $\|T\|^2 = {}^tA_mA_m$ and then ${}^tA_mG_mA_m \geq (2/3)f_0{}^tA_mA_m$. If λ_0 is an eigenvalue of G_m, there exists $A_m \neq 0$ such that $G_mA_m = \lambda_0 A_m$ and then ${}^tA_mG_mA_m = \lambda_0{}^tA_mA_m$. Then, on Δ,

$$\lambda_0{}^tA_mA_m \geq \frac{2}{3}f_0{}^tA_mA_m.$$

Consequently $\lambda_0 \geq (2/3)f_0$. So $\Delta \subset \Gamma$ and $\hat{\Pi}_{\hat{m}}$ minimizes the contrast on Δ.

We now observe that, for all functions T, S

$$\gamma_n(T) - \gamma_n(S) = \Psi_n(T-S) - \frac{2}{n}\sum_{k=1}^{n}[V_{(T-S)}(Y_k, Y_{k+1}) - Q_{(T-S)S}(Y_k)].$$

Then, since on Δ, $\gamma_n(\hat{\Pi}_{\hat{m}}) + \text{pen}(\hat{m}) \leq \gamma_n(\Pi_m) + \text{pen}(m)$,

$$\begin{aligned}
\Psi_n(\hat{\Pi}_{\hat{m}} - \Pi_m) &\leq \frac{2}{n}\sum_{k=1}^{n}[V_{(\hat{\Pi}_{\hat{m}}-\Pi_m)}(Y_k, Y_{k+1}) - Q_{(\hat{\Pi}_{\hat{m}}-\Pi_m)\Pi_m}(Y_k)] \\
&\quad + \text{pen}(m) - \text{pen}(\hat{m}) \\
&\leq 2Z_{n,m}(\hat{\Pi}_{\hat{m}} - \Pi_m) + \text{pen}(m) - \text{pen}(\hat{m}) \\
&\leq 2\|\hat{\Pi}_{\hat{m}} - \Pi_m\|_f \sup_{T \in B_f(m,\hat{m})} Z_{n,m}(T) + \text{pen}(m) - \text{pen}(\hat{m})
\end{aligned}$$

where

$$Z_{n,m}(T) = \frac{1}{n}\sum_{k=1}^{n}[V_T(Y_k, Y_{k+1}) - Q_{T\Pi_m}(Y_k)]$$

and, for all m', $B_f(m,m') = \{T \in S_m + S_{m'}, \quad \|T\|_f = 1\}$. Now let $p(.,.)$ be a function such that for all m, m', $12p(m,m') \leq \text{pen}(m) + \text{pen}(m')$. Then

$$\Psi_n(\hat{\Pi}_{\hat{m}} - \Pi_m) \leq \frac{1}{3}\|\hat{\Pi}_{\hat{m}} - \Pi_m\|_f^2 + 3[\sup_{T \in B_f(m,\hat{m})} Z_{n,m}^2(T) - 4p(m,\hat{m})] + 2\text{pen}(m).$$

So, using the definition of $\Delta \supset \Omega$,

$$\begin{aligned}
&\|\hat{\Pi}_{\hat{m}} - \Pi_m\|_f^2 \mathbb{1}_\Omega \leq \frac{3}{2}\Psi_n(\hat{\Pi}_{\hat{m}} - \Pi_m)\mathbb{1}_\Omega \\
&\leq \frac{1}{2}\|\hat{\Pi}_{\hat{m}} - \Pi_m\|_f^2 \mathbb{1}_\Omega + \frac{9}{2}\sum_{m' \in \mathcal{M}_n}[\sup_{T \in B_f(m,m')} Z_{n,m}^2(T) - 4p(m,m')]\mathbb{1}_\Omega + 3\text{pen}(m)
\end{aligned}$$

Thus

$$\frac{1}{2}\|\hat{\Pi}_{\hat{m}} - \Pi_m\|_f^2 \mathbb{1}_\Omega \leq \frac{9}{2}\sum_{m' \in \mathcal{M}_n}[\sup_{T \in B_f(m,m')} Z_{n,m}^2(T) - 4p(m,m')]\mathbb{1}_\Omega + 3\text{pen}(m).$$

And using Assumption H4,

$$\|\hat{\Pi}_{\hat{m}} - \Pi_m\|_A^2 \mathbb{1}_\Omega \leq 9 f_0^{-1} \sum_{m' \in \mathcal{M}_n} [\sup_{T \in B_f(m,m')} Z_{n,m}^2(T) - 4p(m,m')] \mathbb{1}_\Omega + 6 f_0^{-1} \text{pen}(m). \quad (5.13)$$

Now, by denoting $\mathbb{E}_X$ the expectation conditionally to $X_1, \ldots, X_{n+1}$, the process $Z_{n,m}(T)$ can be split in the following way :

$$Z_{n,m}(T) = Z_n^{(1)}(T) - Z_n^{(2)}(T) + Z_n^{(3)}(T) + \iint T(x,y)(\Pi - \Pi_m)(x,y) f(x) dx dy$$

with

$$\begin{cases} Z_n^{(1)}(T) = \dfrac{1}{n} \displaystyle\sum_{k=1}^{n} V_T(Y_k, Y_{k+1}) - \mathbb{E}_X[V_T(Y_k, Y_{k+1})] \\ Z_n^{(2)}(T) = \dfrac{1}{n} \displaystyle\sum_{k=1}^{n} Q_{T\Pi_m}(Y_k) - \mathbb{E}[Q_{T\Pi_m}(Y_k)] \\ Z_n^{(3)}(T) = \dfrac{1}{n} \displaystyle\sum_{k=1}^{n} T(X_k, X_{k+1}) - \mathbb{E}[T(X_k, X_{k+1})] \end{cases}$$

Then, by introducing functions $p_1(.,.)$, $p_2(.,.)$ and $p_3(.,.)$

$$\begin{aligned} &\sup_{T \in B_f(m,m')} Z_{n,m}^2(T) - 4p(m,m') \leq 4 \sup_{T \in B_f(m,m')} (Z_n^{(1)}(T)^2 - p_1(m,m')) \\ &+4 \sup_{T \in B_f(m,m')} (Z_n^{(2)}(T)^2 - p_2(m,m')) + 4 \sup_{T \in B_f(m,m')} (Z_n^{(3)}(T)^2 - p_3(m,m')) \\ &+4((p_1 + p_2 + p_3)(m,m') - p(m,m')) + 4 \sup_{T \in B_f(m,m')} \|(\Pi - \Pi_m)\mathbb{1}_A\|_f^2 \|T\|_f^2 \end{aligned}$$

We now employ the following propositions.

Proposition 5.3 *Let $p_1(m,m') = K_1 D_{m''}^{4\gamma+2}/n$ where $m'' = \max(m,m')$. Then, if $r > 2\gamma + 1/2$, there exists a positive constant C_1 such that*

$$\sum_{m' \in \mathcal{M}_n} \mathbb{E}\left(\left[\sup_{T \in B_f(m,m')} Z_n^{(1)}(T)^2 - p_1(m,m') \right]_+ \right) \leq \frac{C_1}{n}.$$

Proposition 5.4 *Let $p_2(m,m') = p_2^{(1)}(m,m') + p_2^{(2)}(m,m')$ with $p_2^{(1)}(m,m') = K_2 \|\Pi\|_A^2 D_{m''}^{2\gamma+7/2}/n$ and $p_2^{(2)}(m,m') = K_2 \|\Pi\|_A^2 (\sum_k \beta_k) D_{m''}^3/n$ where $m'' = \max(m,m')$. Then, if $r > 2\gamma + 3/2$, there exists a positive constant C_2 such that*

$$\sum_{m' \in \mathcal{M}_n} \mathbb{E}\left(\left[\sup_{T \in B_f(m,m')} Z_n^{(2)}(T)^2 - p_2(m,m') \right]_+ \mathbb{1}_\Omega \right) \leq \frac{C_2}{n}.$$

Proposition 5.5 *Let* $p_3(m,m') = K_3 \sum_k \beta_{2k} D^2_{m''}/n$ *where* $m'' = \max(m,m')$. *Then, there exists a positive constant* C_3 *such that*

$$\sum_{m' \in \mathcal{M}_n} \mathbb{E}\left(\left[\sup_{T \in B_f(m,m')} Z_n^{(3)}(T)^2 - p_3(m,m') \right]_+ \mathbb{1}_\Omega \right) \leq \frac{C_3}{n}.$$

The two first propositions are proved in Sections 5.6.6 and 5.6.7. The last proposition is proved in the chapter 4 (for another basis but only the property P3 $\| \sum_{jk} \varphi_{jk}^2 \|_\infty \leq \Phi_1 D_m$ is used).

Then we get

$$\sum_{m' \in \mathcal{M}_n} \mathbb{E}\left(\left[\sup_{T \in B_f(m,m')} Z^2_{n,m}(T) - 4p(m,m') \right] \mathbb{1}_\Omega \right) \leq 4\frac{C_1 + C_2 + C_3}{n}$$
$$+4\|(\Pi - \Pi_m)\mathbb{1}_A\|_f^2 + 4 \sum_{m' \in \mathcal{M}_n} ((p_1 + p_2 + p_3)(m,m') - p_1(m,m')).$$

But, if $\gamma > 3/4$, $4\gamma + 2 > 2\gamma + 7/2$ and there exists m_2 such that for all $m' > m_2$, $p_1(m,m') > p_2(m,m') + p_3(m,m')$. That implies that

$$\sum_{m' \in \mathcal{M}_n} (p_1(m,m') + p_2(m,m') + p_3(m,m') - 2p_1(m,m'))$$
$$\leq \sum_{m' \leq m_2} (p_2(m,m') + p_3(m,m') - p_1(m,m')) \leq \frac{C(m_2)}{n}.$$

Thus in the case $\gamma > 3/4$, we choose $p = 2p_1$ and

$$\sum_{m' \in \mathcal{M}_n} \mathbb{E}\left(\left[\sup_{T \in B(m')} Z^2_{n,m}(T) - 4p(m,m') \right] \mathbb{1}_\Omega \right) \leq 4\frac{C_1 + C_2 + C_3 + C(m_2)}{n} \qquad (5.14)$$
$$+4\|f\|_{\infty,A_1}\|\Pi - \Pi_m\|_A^2.$$

If $\gamma = 3/4$, we choose $p = 2(p_1 + p_2^{(1)})$. Since there exists m_2 such that for all $m' > m_2$, $p_1(m,m') + p_2^{(1)}(m,m') > p_2^{(2)}(m,m') + p_3(m,m')$, we can write

$$\sum_{m' \in \mathcal{M}_n} (p_1(m,m') + p_2(m,m') + p_3(m,m') - 2p(m,m'))$$
$$\leq \sum_{m' \leq m_2} (p_2^{(2)}(m,m') + p_3(m,m') - p_1(m,m') - p_2^{(1)}(m,m')) \leq \frac{C(m_2)}{n}$$

and (5.14) holds.

Finally, combining (5.12), (5.13) and (5.14), we obtain

$$\begin{aligned}\mathbb{E}\|\tilde{\Pi}-\Pi\|_A^2 &\leq 2\|\Pi_m-\Pi\|_A^2+\frac{4}{n}(\|\Pi\|_A^4+C_0(1+\|\Pi\|_A^2))\\ &+2(1+2\|\Pi\|_A^2)9f_0^{-1}\Big[4\frac{C_1+C_2+C_3+C(m_2)}{n}+4\|f\|_{\infty,A_1}\|\Pi-\Pi_m\|_A^2\Big]\\ &+2(1+2\|\Pi\|_A^2)6f_0^{-1}\mathrm{pen}(m).\end{aligned}$$

Then, by letting $C=\max(2+72f_0^{-1}\|f\|_{\infty,A_1}(1+2\|\Pi\|_A^2),12f_0^{-1}(1+2\|\Pi\|_A^2))$,

$$\mathbb{E}\|\tilde{\Pi}-\Pi\|_A^2 \leq C\inf_{m\in\mathcal{M}_n}(\|\Pi_m-\Pi\|_A^2+\mathrm{pen}(m))+\frac{C'}{n}$$

We still have to check that $12p(m,m')\leq \mathrm{pen}(m)+\mathrm{pen}(m')$. But, if $\gamma>3/4$,

$$12p(m,m')=24K_1\frac{D_{m''}^{4\gamma+2}}{n}=24K_1\frac{\dim(S_m+S_{m'})^{4\gamma+2}}{n}\leq \mathrm{pen}(m)+\mathrm{pen}(m')$$

with $\mathrm{pen}(m)\geq 24K_1D_m^{4\gamma+2}/n$. And if $\gamma=3/4$,

$$12p(m,m')=24(K_1+K_2\|\Pi\|_A^2)\frac{D_{m''}^5}{n}\leq \mathrm{pen}(m)+\mathrm{pen}(m')$$

with $\mathrm{pen}(m)\geq 24(K_1+K_2\|\Pi\|_A^2)D_m^{4\gamma+2}/n$.

□

5.6.4 Proof of Corollary 5.1

It follows from Meyer (1990) Chapter 6, Section 10 that Π belongs to $B_{2,\infty}^{\alpha}$if and only if $\sup_{j\geq J}2^{2j\alpha}(\sum_{k,l}|a_{jkl}|^2)^{1/2}<\infty$ with $a_{jkl}=\int\Pi(x,y)\varphi_{jk}(x)\varphi_{jl}(y)dxdy$. Then

$$d_A^2(\Pi,S_m)=\sum_{j>m}\sum_{k,l}|a_{jkl}|^2\leq C\sum_{j>m}2^{-4j\alpha}\leq C'D_m^{-2\alpha}$$

Since $d_A^2(\Pi,S_m)=O(D_m^{-2\alpha})$, Theorem 5.1 becomes

$$\mathbb{E}\|\tilde{\Pi}-\Pi\|_A^2\leq C''\inf_{m\in\mathcal{M}_n}\{D_m^{-2\alpha}+\frac{D_m^{4\gamma+2}}{n}\}.$$

with C'' a positive constant. By setting D_{m_1} the integer part of $n^{1/(4\gamma+2\alpha+2)}$, then

$$\mathbb{E}\|\tilde{\Pi}-\Pi\|_A^2\leq C''\{D_{m_1}^{-2\alpha}+\frac{D_{m_1}^{4\gamma+2}}{n}\}=O(n^{-\frac{2\alpha}{4\gamma+2\alpha+2}}).$$

□

5.6.5 Proof of Proposition 5.2

We first remark that $P(\Omega_\rho^{*c}) \leq P(\Omega_X^{*c}) + P(\Delta^c \cap \Omega_X^*)$. In the geometric case $\beta_{q_n} \leq e^{-\theta c \log(n)} \leq n^{-\theta c}$ and in the other case $\beta_{q_n} \leq (q_n)^{-\theta} \leq n^{-\theta c}$. Then

$$P(\Omega_X^{*c}) \leq 2p_n\beta_{q_n} \leq n^{1-c\theta}.$$

But, $c\theta > 3$ and so $P(\Omega_X^{*c}) \leq n^{-2}$.

We still have to bound $P(\Delta^c \cap \Omega_X^*)$. To do this, we observe that if $\omega \in \Delta^c$, then there exists T in $\mathcal{S}$ such that $\|T\|_f^2 > (3/2)\Psi_n(T)$ and then $\|T\|_f^2 > (3/2)\mathbb{E}_X\Psi_n(T)$. But $\mathbb{E}_X\Psi_n(T) = \frac{1}{n}\sum_{k=1}^n \int T^2(X_k, y)dy$. So $P(\Delta^c \cap \Omega_X^*) \leq P(\Delta'^c \cap \Omega_X^*)$ with

$$\Delta' = \{\forall T \in \mathcal{S} \quad \|T\|_f^2 \leq \frac{3}{2}\frac{1}{n}\sum_{k=1}^n \int T^2(X_k, y)dy\}.$$

Let us remark that $(1/n)\sum_{k=1}^n \int T^2(X_k, y)dy - \|T\|_f^2 = \nu_n(T^2)$ with

$$\nu_n(T) = \frac{1}{n}\sum_{i=1}^n \int [T(X_i, y) - \mathbb{E}(T(X_i, y))]dy.$$

Hence

$$P(\Delta'^c \cap \Omega_X^*) \leq P(\sup_{T\in\mathcal{B}} |\nu_n(T^2)|\mathbb{1}_{\Omega_X^*} > 1/3)$$

with $\mathcal{B} = \{T \in \mathcal{S} \quad \|T\|_f = 1\}$.

A function T in $\mathcal{S}$ can be written $T(x, y) = \sum_{j=J}^{m_0}\sum_{kl} a_{jkl}\varphi_{jk}(x)\varphi_{jl}(y)$ where m_0 is such that $\mathcal{S} = \mathbb{S}_{m_0}$. Then

$$\nu_n(T^2)\mathbb{1}_{\Omega_X^*} = \sum_{jkk'l} a_{jkl}a_{jk'l}\bar{\nu}_n(\varphi_{jk}\varphi_{jk'})$$

where

$$\bar{\nu}_n(u) = \frac{1}{n}\sum_{i=1}^n [u(X_i^*) - \mathbb{E}(u(X_i^*))]. \tag{5.15}$$

Let $b_{jk} = (\sum_l a_{jkl}^2)^{1/2}$, then $|\nu_n(T^2)| \leq \sum_{jkk'} b_{jk}b_{jk'}|\bar{\nu}_n(\varphi_{jk}\varphi_{jk'})|$ and, if $T \in \mathcal{B}$, $\sum_{jk} b_{jk}^2 = \sum_{jkl} a_{jkl}^2 = \|T\|^2 \leq f_0^{-1}$

Thus,

$$\sup_{T\in\mathcal{B}} |\nu_n(T^2)| \leq f_0^{-1} \sup_{\sum b_{jk}^2=1} \sum_{jkk'} b_{jk}b_{jk'}|\bar{\nu}_n(\varphi_{jk}\varphi_{jk'})|.$$

For the sake of simplicity, we denote $\lambda = (j, k)$ and $\lambda' = (j, k')$ so that

$$\sup_{T\in\mathcal{B}} |\nu_n(T^2)| \leq f_0^{-1} \sup_{\sum b_\lambda^2=1} \sum_{\lambda\lambda'} b_\lambda b_{\lambda'}|\bar{\nu}_n(\varphi_\lambda\varphi_{\lambda'})|.$$

Lemma 5.3 *Let $B_{\lambda,\lambda'} = \|\varphi_\lambda \varphi_{\lambda'}\|_\infty$ and $V_{\lambda,\lambda'} = \|\varphi_\lambda \varphi_{\lambda'}\|$. Let, for any symmetric matrix $(A_{\lambda,\lambda'})$*

$$\bar{\rho}(A) = \sup_{\sum a_\lambda^2 = 1} \sum_{\lambda,\lambda'} |a_\lambda a_{\lambda'}| A_{\lambda,\lambda'}$$

and $L(\varphi) = \max\{\bar{\rho}^2(V), \bar{\rho}(B)\}$. Then there exists $\Phi_0 > 0$ such that $L(\varphi) \leq \Phi_0 \mathcal{D}^2$.

This lemma is proved in Baraud *et al.* (2001) for an orthonormal basis verifying $\|\sum_\lambda \varphi_\lambda^2\|_\infty \leq \Phi_0 \mathcal{D}$, that is ensured by property P3.

Now let $x = \dfrac{f_0^2}{24\|f\|_{\infty,A_1} L(\varphi)}$ and

$$D = \left\{ \forall\lambda \forall\lambda' \quad |\bar{\nu}_n(\varphi_\lambda \varphi_{\lambda'})| \leq \left[B_{\lambda,\lambda'} x + V_{\lambda,\lambda'} \sqrt{2\|f\|_{\infty,A_1} x} \right] \right\}.$$

On D:

$$\begin{aligned}
\sup_{T\in\mathcal{B}} |\nu_n(T^2)| &\leq f_0^{-1} \sup_{\sum b_\lambda^2 = 1} \sum_{\lambda,\lambda'} b_\lambda b_{\lambda'} \left[B_{\lambda,\lambda'} x + V_{\lambda,\lambda'} \sqrt{2\|f\|_{\infty,A_1} x} \right] \\
&\leq f_0^{-1} \left[\bar{\rho}(B) x + \bar{\rho}(V) \sqrt{2\|f\|_{\infty,A_1} x} \right] \\
&\leq \frac{f_0}{24\|f\|_{\infty,A_1}} \frac{\bar{\rho}(B)}{L(\varphi)} + \frac{1}{\sqrt{12}} \left(\frac{\bar{\rho}^2(V)}{L(\varphi)} \right)^{1/2} \leq \frac{1}{24} + \frac{1}{2\sqrt{3}} < \frac{1}{3}.
\end{aligned}$$

Then $P\left(\sup_{T\in\mathcal{B}} |\nu_n(T^2)| > 1/3 \right) \leq P(D^c)$. But $\bar{\nu}_n(u) = \bar{\nu}_{n,1}(u)/2 + \bar{\nu}_{n,2}(u)/2$ with

$$\bar{\nu}_{n,s}(u) = \frac{1}{p_n} \sum_{l=0}^{p_n - 1} Y_{l,s}(u) \qquad s = 1, 2$$

with $\begin{cases} Y_{l,1}(u) &= \dfrac{1}{q_n} \sum_{i=2lq_n+1}^{(2l+1)q_n} [u(X_i^*) - \mathbb{E}(u(X_i^*))], \\ Y_{l,2}(u) &= \dfrac{1}{q_n} \sum_{i=2(2l+1)q_n+1}^{(2l+2)q_n} [u(X_i^*) - \mathbb{E}(u(X_i^*))]. \end{cases}$

To bound $P(\bar{\nu}_{n,1}(\varphi_\lambda \varphi_{\lambda'}) \geq B_{\lambda,\lambda'} x + V_{\lambda,\lambda'} \sqrt{2\|f\|_{\infty,A_1} x})$, we will use the Bernstein inequality given in Birgé and Massart (1998). That is why we bound $\mathbb{E}|Y_{l,1}(u)|^p$:

$$\begin{aligned}
\mathbb{E}|Y_{l,1}(u)|^p &\leq \frac{1}{q_n^2} (2\|u\|_\infty)^{p-2} \mathbb{E} \left| \sum_{i=2lq_n+1}^{(2l+1)q_n} [u(X_i^*) - \mathbb{E}(u(X_i^*))] \right|^2 \\
&\leq (2\|u\|_\infty)^{p-2} \frac{1}{q_n^2} \mathbb{E} \left| \sum_{i=2lq_n+1}^{(2l+1)q_n} [u(X_i) - \mathbb{E}(u(X_i))] \right|^2
\end{aligned}$$

since $X_i^* = X_i$ on Ω_X^*. Moreover an elementary convex inequality gives

$$\begin{aligned}\mathbb{E}|Y_{l,1}(u)|^p &\leq (2\|u\|_\infty)^{p-2}\frac{1}{q_n}\sum_{i=2lq_n+1}^{(2l+1)q_n}\mathbb{E}|u(X_i)-\mathbb{E}(u(X_i))|^2 \\ &\leq (2\|u\|_\infty)^{p-2}\int u^2(x)f(x)dx \leq 2^{p-2}(\|u\|_\infty)^{p-2}(\sqrt{\|f\|_{\infty,A_1}}\|u\|)^2.\end{aligned}$$

With $u=\varphi_\lambda\varphi_{\lambda'}$, $\mathbb{E}|Y_{l,1}(\varphi_\lambda\varphi_{\lambda'})|^p \leq 2^{p-2}(B_{\lambda,\lambda'})^{p-2}(\sqrt{\|f\|_{\infty,A_1}}V_{\lambda,\lambda'})^2$. And then

$$P(|\bar{\nu}_{n,s}(\varphi_\lambda\varphi_{\lambda'})| \geq B_{\lambda,\lambda'}x + V_{\lambda,\lambda'}\sqrt{2\|f\|_{\infty,A_1}x}) \leq 2e^{-p_n x}.$$

Let $C = f_0^2[48\|f\|_{\infty,A_1}]^{-1}$, so that $x = 2C/L(\varphi)$. Given that $P(\Delta^c \cap \Omega_X^*) \leq P(D^c) \leq \sum_{\lambda,\lambda'} P\left(|\bar{\nu}_n(\varphi_\lambda\varphi_{\lambda'})| > B_{\lambda,\lambda'}x + V_{\lambda,\lambda'}\sqrt{2\|f\|_{\infty,A_1}x}\right)$,

$$P(\Delta^c \cap \Omega_X^*) \leq 4\mathcal{D}^2\exp\left\{-\frac{2p_nC}{L(\varphi)}\right\} \leq 4n^{1/(2\gamma+1)}\exp\left\{-C\frac{n}{q_nL(\varphi)}\right\}.$$

But $L(\varphi) \leq \Phi_0\mathcal{D}^2 \leq \Phi_0 n^{1/(2\gamma+1)}$ and $q_n \leq n^{1/2}$ so

$$P(\Delta^c \cap \Omega_X^*) \leq 4n^{1/(2\gamma+1)}\exp\left\{-\frac{C}{\Phi_0}n^{\frac{2\gamma-1}{2(2\gamma+1)}}\right\} \leq \frac{C'}{n^2}$$

because $\gamma > 1/2$.

□

5.6.6 Proof of Proposition 5.3

First we need to isolate even terms from odd terms in $Z_n^{(1)}(T)$ to avoid overlaps: $Z_n^{(1)}(T)\mathbb{1}_{\Omega^*} = \frac{1}{2}(Z_n^{(1,1)}(T) + Z_n^{(1,2)}(T))$ with

$$\begin{cases} Z_n^{(1,1)}(T) = \dfrac{1}{n}\displaystyle\sum_{i=1, i \text{ odd}}^{n} V_T(Y_i, Y_{i+1}) - \mathbb{E}_X[V_T(Y_i, Y_{i+1})] \\ Z_n^{(1,2)}(T) = \dfrac{1}{n}\displaystyle\sum_{i=1, i \text{ even}}^{n} V_T(Y_i, Y_{i+1}) - \mathbb{E}_X[V_T(Y_i, Y_{i+1})] \end{cases}$$

It is sufficient to deal with the first term only, as the second one is similar. For each i, let $U_i = (Y_{2i-1}, Y_{2i})$, then

$$Z_n^{(1,1)}(T) = \frac{1}{n/2}\sum_{i=1}^{n/2}\left\{V_T(U_i) - \mathbb{E}_X[V_T(U_i)]\right\}.$$

Notice that conditionally to $X_1, \dots, X_n$, the U_i's are independent. Thus we can use the Talagrand inequality recalled in Lemma 5.5 to bound

$$\mathbb{E}\left(\left[\sup_{T\in B_f(m,m')} Z_n^{(1,1)}(T)^2 - p_1(m,m')\right]_+\right).$$

We first remark that Property P1 entails $B_f(m,m') \subset \mathbb{S}_{m''}$ with $m'' = \max(m,m')$. Then, if T belongs to $B_f(m,m')$,

$$T(x,y) = \sum_{j=J}^{m''} \sum_{kl} a_{jkl}\varphi_{jk}(x)\varphi_{jl}(y)$$

with $\sum_{jkl} a_{jkl}^2 = \|T\|^2 \le f_0^{-1}$.

- Let us bound $\|V_T\|_\infty$ for T in $B_f(m,m')$. If $T(x,y) = \sum_{jkl} a_{jkl}\varphi_{jk}(x)\varphi_{jl}(y)$,

$$|V_T(x,y)|^2 \le \sum_{jkl} a_{jkl}^2 \sum_{jkl} |V_{\varphi_{jk}\otimes\varphi_{jl}}(x,y)|^2.$$

Then, since $V_{s\otimes t}(x,y) = v_s(x)v_t(y)$,

$$\sup_{T\in B_f(m,m')} |V_T(x,y)|^2 \le f_0^{-1} \sum_{jkl} |v_{\varphi_{jk}}(x)v_{\varphi_{jl}}(y)|^2.$$

But, according to Property P4, $\|\sum_k |v_{\varphi_{jk}}|^2\|_\infty \le \Phi_1(2^j)^{2\gamma+2}$. So

$$\sup_{T\in B_f(m,m')} \|V_T\|_\infty^2 \le f_0^{-1}\Phi_1^2 \sum_{j=J}^{m''} (2^j)^{4\gamma+4} \le f_0^{-1}\Phi_1^2 \frac{2^{4\gamma+4}}{2^{4\gamma+4}-1} D_{m''}^{4\gamma+4}$$

and $M_1 = f_0^{-1/2}\Phi_1\sqrt{2^{4\gamma+4}/(2^{4\gamma+4}-1)}D_{m''}^{2\gamma+2}$.

- To compute H^2 we write

$$\begin{aligned}
&\mathbb{E}_X[\sup_{T\in B_f(m,m')} Z_n^{(1,1)}(T)^2] \le f_0^{-1}\sum_{jkl} \mathbb{E}_X[Z_n^{(1,1)}(\varphi_{jk}\otimes\varphi_{jl})^2] \\
&\le f_0^{-1}\sum_{jkl} \mathrm{Var}_X\left(\frac{1}{n}\sum_{i=1,i\text{ odd}}^{n} v_{\varphi_{jk}}(Y_i)v_{\varphi_{jl}}(Y_{i+1})\right) \\
&\le f_0^{-1}\sum_{jkl} \frac{1}{n}\mathrm{Var}_X\left(v_{\varphi_{jk}}(Y_1)v_{\varphi_{jl}}(Y_2)\right) \le \frac{f_0^{-1}}{n}\sum_{jkl} \mathbb{E}_X[|v_{\varphi_{jk}}(Y_1)|^2|v_{\varphi_{jl}}(Y_2)|^2] \quad (5.16)
\end{aligned}$$

Here Var_X denotes the variance conditionally to $X_1, \ldots, X_{n+1}$. Now, for any function G, the following relation holds

$$\begin{aligned}
&\mathbb{E}_X[|G|^2(Y_1, Y_2)] = \mathbb{E}_X[|G|^2(X_1 + \varepsilon_1, X_2 + \varepsilon_2)] \\
&= \iint |G|^2(X_1 + z_1, X_2 + z_2) q(z_1) q(z_2) dz_1 dz_2 \\
&= \iint |G|^2(u_1, u_2) q(u_1 - X_1) q(u_2 - X_2) du_1 du_2 \leq \|q\|_\infty^2 \|G\|^2
\end{aligned}$$

Then, coming back to (5.16),

$$\begin{aligned}
&\mathbb{E}_X \left[\sup_{T \in B_f(m,m')} Z_n^{(1,1)}(T)^2 \right] \leq \frac{f_0^{-1}}{n} \|q\|_\infty^2 \sum_{jkl} \|v_{\varphi_{jk}} \otimes v_{\varphi_{jl}}\|^2 \\
&\leq \frac{f_0^{-1} \|q\|_\infty^2}{n} \sum_j \left(\sum_k \|v_{\varphi_{jk}}\|^2 \right)^2 \leq \frac{\Phi_1^2 f_0^{-1} \|q\|_\infty^2}{n} \sum_{j=J}^{m''} (2^j)^{4\gamma+2},
\end{aligned}$$

using P5. Then $H^2 = \Phi_1^2 f_0^{-1} \|q\|_\infty^2 2^{4\gamma+2} / (2^{4\gamma+2} - 1) \dfrac{D_{m''}^{4\gamma+2}}{n}$.

- We still have to find v. First

$$\mathrm{Var}_X(V_T(Y_i, Y_{i+1})) \leq \mathbb{E}_X |V_T(Y_i, Y_{i+1})|^2 \leq \|q\|_\infty^2 \|V_T\|^2$$

We now observe that $\|V_T\|^2 = \|V_T^*\|^2/(4\pi^2)$ and then

$$\begin{aligned}
\|V_T\|^2 &= \frac{1}{4\pi^2} \iint \left| \frac{T^*(u,v)}{q^*(-u) q^*(-v)} \right|^2 du dv \\
&\leq \frac{1}{4\pi^2} \sqrt{\iint \frac{|T^*(u,v)|^2}{|q^*(-u) q^*(-v)|^4} du dv} \sqrt{\iint |T^*(u,v)|^2 du dv} \\
&\leq \frac{1}{4\pi^2} \sqrt{\sum_{jkl} a_{jkl}^2 \sum_{jkl} \iint \frac{|\varphi_{jk}^*(u) \varphi_{jl}^*(v)|^2}{|q^*(-u) q^*(-v)|^4} du dv} \sqrt{4\pi^2 \|T\|^2}
\end{aligned}$$

For $T \in B_f(m, m')$,

$$\|V_T\|^2 \leq \frac{f_0^{-1/2}}{2\pi} \sqrt{f_0^{-1} \sum_j \sum_{kl} \int \frac{|\varphi_{jk}^*(u)|^2}{|q^*(-u)|^4} du \int \frac{|\varphi_{jl}^*(u)|^2}{|q^*(-u)|^4} du}.$$

But $(\varphi_{jk})^*(u) = 2^{-j/2} e^{iuk/2^j} \varphi^*(u/2^j)$ and then

$$\begin{aligned}
&\int \frac{|\varphi_{jk}^*(u)|^2}{|q^*(-u)|^4} du \leq \int \frac{2^{-j} |\varphi^*(u/2^j)|^2}{|q^*(-u)|^4} du \\
&\leq \int \frac{|\varphi^*(v)|^2}{|q^*(-v 2^j)|^4} dv \leq k_0^{-4} (2^j)^{4\gamma} \int |\varphi^*(v)|^2 (v^2 + 1)^{2\gamma} dv.
\end{aligned}$$

Since $r > 2\gamma + 1/2$, Lemma 5.4 gives

$$\sum_{(k,l)\in\Lambda_j} \int \frac{|\varphi^*_{jk}(u)|^2}{|q^*(-u)|^4}du \int \frac{|\varphi^*_{jl}(u)|^2}{|q^*(-u)|^4}du \leq 3.2^{2j} C^2_{2,4\gamma} k_0^{-8} (2^j)^{8\gamma}.$$

Then

$$\|V_T\|^2 \leq \frac{C_{2,4\gamma} f_0^{-1} k_0^{-4}}{2\pi} \sqrt{\sum_{j=J}^{m''} 3(2^j)^{8\gamma+2}} \leq \frac{C_{2,4\gamma} f_0^{-1} k_0^{-4}}{2\pi} \left(\frac{3.2^{8\gamma+2}}{2^{8\gamma+2}-1}\right)^{1/2} D_{m''}^{4\gamma+1}$$

and $v = \|q\|^2_\infty C_{2,4\gamma} f_0^{-1} k_0^{-4} \sqrt{3.2^{8\gamma+2}} D_{m''}^{4\gamma+1}/(2\pi\sqrt{2^{8\gamma+2}-1})$.

We can now apply inequality (5.19)

$$\begin{aligned}\mathbb{E}[\sup_{T\in B_f(m,m')} |Z_n^{(1,1)}(T)|^2 - 6H^2]_+ &\leq C\left(\frac{v}{n} e^{-k_1 \frac{nH^2}{v}} + \frac{M_1^2}{n^2} e^{-k_2 \frac{nH}{M_1}}\right) \\ &\leq C'\left(\frac{D_{m''}^{4\gamma+1}}{n} e^{-k_1' D_{m''}} + \frac{D_{m''}^{4\gamma+4}}{n^2} e^{-k_2'\sqrt{n}/D_{m''}}\right).\end{aligned}$$

But there exists a positive constant K such that

$$\sum_{m'\in\mathcal{M}_n} D_{m''}^{4\gamma+1} e^{-k_1' D_{m''}} \leq K.$$

Moreover, since $D_{m''} \leq n^{\frac{1}{4\gamma+2}}$, $D_{m''}^{4\gamma+4} e^{-k_2'\sqrt{n}/D_{m''}}/n \leq n^{1/(2\gamma+1)} e^{-k_2' n^{\gamma/(2\gamma+1)}}$ so that

$$\sum_{m'\in\mathcal{M}_n} D_{m''}^{4\gamma+4} e^{-k_2'\sqrt{n}/D_{m''}}/n^2 \leq K'/n.$$

Then, setting $K_1 = 6\Phi_1^2 f_0^{-1} \|q\|^2_\infty 2^{4\gamma+2}/(2^{4\gamma+2}-1)$,

$$\sum_{m'\in\mathcal{M}_n} \mathbb{E}[\sup_{T\in B_f(m,m')} |Z_n^{(1,1)}(T)|^2 - K_1 \frac{D_{m''}^{4\gamma+2}}{n}]_+ \leq \frac{C''}{n}$$

and the proposition is proved.

□

5.6.7 Proof of Proposition 5.4

Since Π_m belongs to $\mathbb{S}_m$, it can be written

$$\Pi_m(x,y) = \sum_{j'=J}^{m} \sum_{(k',l')\in\Lambda_{j'}} b_{j'k'l'} \varphi_{j'k'}(x) \varphi_{j'l'}(y)$$

with $\sum_{j'k'l'} b^2_{j'k'l'} = \|\Pi_m\|^2 \leq \|\Pi\|^2_A$. From the embedding $B_f(m,m') \subset S_{m''}$ (where $m'' = \max(m,m')$), we have, if T belongs to $B_f(m,m')$,

$$T(x,y) = \sum_{j=J}^{m''} \sum_{(k,l)\in\Lambda_j} a_{jkl}\varphi_{jk}(x)\varphi_{jl}(y)$$

with $\sum_{jkl} a^2_{jkl} = \|T\|^2 \leq f_0^{-1}$.

We use the Talagrand inequality (5.19) in Lemma 5.5. But the variables Y_i are not independent. We shall use the following approximation variables

$$\forall 1 \leq i \leq n+1 \qquad Y_i^* = X_i^* + \varepsilon_i.$$

These variables have the same properties (see (5.11)) that the X_i^*'s. More precisely, let, for $l = 0, \ldots, p_n - 1$, $C_l = (Y_{2lq_n+1}, ..., Y_{(2l+1)q_n})$, $D_l = (Y_{(2l+1)q_n+1}, ..., Y_{(2l+2)q_n})$, $C_l^* = (Y^*_{2lq_n+1}, ..., Y^*_{(2l+1)q_n})$, $D_l^* = (Y^*_{(2l+1)q_n+1}, ..., Y^*_{(2l+2)q_n})$. Then, since A_l and A_l^* have the same distribution and the sequences (ε_i) and (X_i) are independent,

$$C_l \text{ and } C_l^* \text{ have the same distribution.}$$

Moreover the construction of A_l^* via Berbee's coupling Lemma implies that

$$C_l^* \text{ and } C_{l'}^* \text{ are independent if } l \neq l'.$$

Now we split $Z_n^{(2)}$ into two terms: $Z_n^{(2)}(T)\mathbb{1}_\Omega = (1/2)Z_n^{(2,1)}(T) + (1/2)Z_n^{(2,2)}(T)$ where

$$\begin{cases} Z_n^{(2,1)}(T) = \dfrac{1}{p_n}\displaystyle\sum_{l=0}^{p_n-1}\frac{1}{q_n}\sum_{i=2lq_n+1}^{(2l+1)q_n} Q_{T\Pi_m}(Y_i^*) - \mathbb{E}[Q_{T\Pi_m}(Y_i^*)] \\ Z_n^{(2,2)}(T) = \dfrac{1}{p_n}\displaystyle\sum_{l=0}^{p_n-1}\frac{1}{q_n}\sum_{i=(2l+1)q_n+1}^{(2l+2)q_n} Q_{T\Pi_m}(Y_i^*) - \mathbb{E}[Q_{T\Pi_m}(Y_i^*)] \end{cases}$$

Then we apply Talagrand's inequality to $Z_n^{(2,1)}(T)$.

• Let us first compute M_1. We have to bound $\|Q_{T\Pi_m}\|_\infty$ for T in $B_f(m,m')$. By linearity of Q

$$Q_{T\Pi_m}(x) = \sum_{jkl} a_{jkl} \sum_{j'k'l'} b_{j'k'l'} Q_{\varphi_{jk}\varphi_{j'k'}\otimes\varphi_{jl}\varphi_{j'l'}}(x)$$

Then, since $Q_{s\otimes t}(x) = v_s(x)\int t(y)dy$, using the Schwarz inequality,

$$\begin{aligned} |Q_{T\Pi_m}(x)|^2 &\leq \sum a^2_{jkl} b^2_{j'k'l'} \sum |v_{\varphi_{jk}\varphi_{j'k'}}(x) \int \varphi_{jl}\varphi_{j'l'}|^2 \\ &\leq f_0^{-1}\|\Pi\|^2_A \sum_{jkk'l} |v_{\varphi_{jk}\varphi_{jk'}}(x)|^2 \end{aligned}$$

since the φ_{jl} are orthonormal. The property P6 then gives

$$\|Q_{T\Pi_m}\|_\infty^2 \leq f_0^{-1}\|\Pi\|_A^2\Phi_1\sum_{j=J}^{m''}(2^j)^{2\gamma+3}2^j$$

so that $M_1 = f_0^{-1/2}\|\Pi\|_A\sqrt{\Phi_1 2^{2\gamma+4}/(2^{2\gamma+4}-1)}D_{m''}^{\gamma+2}$.

• Now, we compute H^2. For $T \in B_f(m,m')$,

$$|Z_n^{(2,1)}(T)|^2 \leq \sum_{jkl} a_{jkl}^2 \sum_{jkl}|Z_n^{(2,1)}(\varphi_{jk}\otimes\varphi_{jl})|^2$$

Thus

$$\begin{aligned}\mathbb{E}\left[\sup_{T\in B_f(m,m')} Z_n^{(2,1)}(T)^2\right] &\leq f_0^{-1}\sum_{jkl}\mathbb{E}\left[Z_n^{(2,1)}(\varphi_{jk}\otimes\varphi_{jl})^2\right]\\ &\leq f_0^{-1}\sum_{jkl}\mathrm{Var}\left[\frac{1}{p_n}\sum_{l=0}^{p_n-1}\frac{1}{q_n}\sum_{i=2lq_n+1}^{(2l+1)q_n} Q_{(\varphi_{jk}\otimes\varphi_{jl})\Pi_m}(Y_i^*)\right].\end{aligned}$$

The variables (C_l^*) are independent and identically distributed so

$$\mathbb{E}\left[\sup_{T\in B_f(m,m')} Z_n^{(2,1)}(T)^2\right] \leq f_0^{-1}\sum_{jkl}\frac{1}{p_n}\mathrm{Var}\left[\frac{1}{q_n}\sum_{i=1}^{q_n} Q_{(\varphi_{jk}\otimes\varphi_{jl})\Pi_m}(Y_i^*)\right].$$

But, on Ω, C_1 and C_1^* have the same distribution, so that

$$\mathrm{Var}\left[\frac{1}{q_n}\sum_{i=1}^{q_n} Q_{(\varphi_{jk}\otimes\varphi_{jl})\Pi_m}(Y_i^*)\right] = \mathrm{Var}\left[\frac{1}{q_n}\sum_{i=1}^{q_n} Q_{(\varphi_{jk}\otimes\varphi_{jl})\Pi_m}(Y_i)\right].$$

But, coming back to the definition of Q_T, for $i_1 \neq i_2$,

$$\begin{aligned}&\mathrm{Cov}(Q_{(\varphi_{jk}\otimes\varphi_{jl})\Pi_m}(Y_{i_1}), Q_{(\varphi_{jk}\otimes\varphi_{jl})\Pi_m}(Y_{i_2}))\\ &= \frac{1}{4\pi^2}\iint \mathbb{E}(e^{iY_{i_1}u}e^{-iY_{i_2}v})\frac{[(\varphi_{jk}\otimes\varphi_{jl})\Pi_m]^*(u,0)}{q^*(-u)}\frac{[(\varphi_{jk}\otimes\varphi_{jl})\Pi_m]^*(-v,0)}{q^*(v)}dudv\\ &= \frac{1}{4\pi^2}\iint \mathbb{E}(e^{iX_{i_1}u}e^{-iX_{i_2}v})[(\varphi_{jk}\otimes\varphi_{jl})\Pi_m]^*(u,0)[(\varphi_{jk}\otimes\varphi_{jl})\Pi_m]^*(-v,0)dudv\end{aligned}$$

since $\mathbb{E}(e^{i\varepsilon_{i_1}u}e^{-i\varepsilon_{i_2}v}) = q^*(-u)q^*(v)$. Now using (5.10),

$$\begin{aligned}&\mathrm{Cov}(Q_{(\varphi_{jk}\otimes\varphi_{jl})\Pi_m}(Y_{i_1}), Q_{(\varphi_{jk}\otimes\varphi_{jl})\Pi_m}(Y_{i_2}))\\ &= \mathrm{Cov}(\int(\varphi_{jk}\otimes\varphi_{jl})\Pi_m(X_{i_1},y)dy, \int(\varphi_{jk}\otimes\varphi_{jl})\Pi_m(X_{i_2},y)dy).\end{aligned}$$

It implies that

$$\mathrm{Var}\left[\frac{1}{q_n}\sum_{i=1}^{q_n} Q_{(\varphi_{jk}\otimes\varphi_{jl})\Pi_m}(Y_i)\right] \leq \frac{1}{q_n^2}\sum_{i=1}^{q_n}\mathrm{Var}[Q_{(\varphi_{jk}\otimes\varphi_{jl})\Pi_m}(Y_i)] + \mathrm{Var}\left[\frac{1}{q_n}\sum_{i=1}^{q_n}\int(\varphi_{jk}\otimes\varphi_{jl})\Pi_m(X_i,y)dy\right]$$

And then

$$\mathbb{E}\left[\sup_{T\in B_f(m,m')} Z_n^{(2,1)}(T)^2\right] \leq f_0^{-1}\sum_{jkl}\frac{1}{p_n q_n}\mathrm{Var}[Q_{(\varphi_{jk}\otimes\varphi_{jl})\Pi_m}(Y_1)] + f_0^{-1}\sum_{jkl}\frac{1}{p_n}\mathrm{Var}\left[\frac{1}{q_n}\sum_{i=1}^{q_n}\int(\varphi_{jk}\otimes\varphi_{jl})\Pi_m(X_i,y)dy\right] \quad (5.17)$$

For the second term in (5.17), we use Lemma 5.6 to write

$$\sum_{jkl}\mathrm{Var}\left[\frac{1}{q_n}\sum_{i=1}^{q_n}\int(\varphi_{jk}\otimes\varphi_{jl})\Pi_m(X_i,y)dy\right] \leq \frac{4\sum_k\beta_k}{q_n}\|\sum_{jkl}|\int(\varphi_{jk}\otimes\varphi_{jl})\Pi_m(X_i,y)dy|^2\|_\infty$$

But

$$\int(\varphi_{jk}\otimes\varphi_{jl})\Pi_m(X_i,y)dy = \sum_{j'k'l'} b_{j'k'l'}\varphi_{jk}\varphi_{j'k'}(X_i)\int\varphi_{jl}\varphi_{j'l'}(y)dy = \sum_{k'} b_{jk'l}\varphi_{jk}\varphi_{jk'}(X_i)$$

so that

$$\sum_{jkl}|\int(\varphi_{jk}\otimes\varphi_{jl})\Pi_m(X_i,y)dy|^2 \leq \|\Pi\|_A^2\sum_{jkk'l}|\varphi_{jk}\varphi_{jk'}(X_i)|^2 \leq \|\Pi\|_A^2\Phi_1^2 D_{m''}^3$$

using property P3. Therefore

$$\sum_{jkl}\mathrm{Var}\left[\frac{1}{q_n}\sum_{i=1}^{q_n}\int(\varphi_{jk}\otimes\varphi_{jl})\Pi_m(X_i,y)dy\right] \leq \frac{4\sum_k\beta_k}{q_n}\|\Pi\|_A^2\Phi_1^2 D_{m''}^3.$$

Then we have bound the second term in (5.17) by $2f_0^{-1}\sum_k\beta_k\|\Pi\|_A^2\Phi_1^2 D_{m''}^3/n$.

For the first term in (5.17), we bound $\sum_{jkl}\mathbb{E}[|Q_{(\varphi_{jk}\otimes\varphi_{jl})\Pi_m}(Y_1)|^2]$:

$$\begin{aligned}\sum_{jkl}\mathbb{E}[|Q_{(\varphi_{jk}\otimes\varphi_{jl})\Pi_m}(Y_1)|^2] &\leq \sum_{j'k'l'} b_{j'k'l'}^2\sum_{jkl}\sum_{j'k'l'}|\int\varphi_{jl}\varphi_{j'l'}|^2\mathbb{E}|v_{\varphi_{jk}\varphi_{j'k'}}(Y_1)|^2 \\ &\leq \|\Pi\|_A^2\sum_{jkk'}2^j\mathbb{E}|v_{\varphi_{jk}\varphi_{jk'}}(Y_1)|^2\end{aligned}$$

But $\mathbb{E}|v_{\varphi_{jk}\varphi_{jk'}}(Y_1)|^2 = \int |v_{\varphi_{jk}\varphi_{jk'}}(x)|^2 p(x)dx$ where p is the density of Y_1. Since $p = q * f$, $|p(x)| \leq \|q\|_\infty$ for all x. Then

$$\mathbb{E}|v_{\varphi_{jk}\varphi_{jk'}}(Y_1)|^2 \leq \|q\|_\infty \int |v_{\varphi_{jk}\varphi_{jk'}}(x)|^2 dx$$

and

$$\sum_{jkl} \mathbb{E}[|Q_{(\varphi_{jk}\otimes\varphi_{jl})\Pi_m}(Y_1)|^2] \leq \|\Pi\|_A^2 \|q\|_\infty \sum_{j=J}^{m''} 2^j \sum_{kk'} \int |v_{\varphi_{jk}\varphi_{jk'}}(x)|^2 dx$$
$$\leq \|\Pi\|_A^2 \|q\|_\infty \Phi_1 \sum_{j=J}^{m''} (2^j)^{2\gamma+3} \leq \|\Pi\|_A^2 \|q\|_\infty \Phi_1 \frac{2^{2\gamma+3}}{2^{2\gamma+3}-1} D_{m''}^{2\gamma+3},$$

applying Property P7. We obtain finally

$$\mathbb{E}\left[\sup_{T\in B_f(m,m')} Z_n^{(2,1)}(T)^2\right] \leq 2f_0^{-1}\|\Pi\|_A^2\|q\|_\infty\Phi_1 \frac{2^{2\gamma+3}}{2^{2\gamma+3}-1}\frac{D_{m''}^{2\gamma+3}}{n}$$
$$+2f_0^{-1}\sum_k \beta_k \|\Pi\|_A^2 \Phi_1^2 \frac{D_{m''}^3}{n}.$$

Since the order of nH^2 has to be larger than the one of v, we choose

$$H^2 = 2f_0^{-1}\|\Pi\|_A^2\Phi_1 \max(\|q\|_\infty 2^{2\gamma+3}/(2^{2\gamma+3}-1), \Phi_1)\left[\frac{D_{m''}^{2\gamma+7/2}}{n} + (\sum_k \beta_k)\frac{D_{m''}^3}{n}\right].$$

- Lastly, using Lemma 5.6 again,

$$\mathrm{Var}(\frac{1}{q_n}\sum_{i=2lq_n+1}^{(2l+1)q_n} Q_{T\Pi_m}^2(Y_i^*)) = \mathrm{Var}(\frac{1}{q_n}\sum_{i=2lq_n+1}^{(2l+1)q_n} Q_{T\Pi_m}^2(Y_i))$$
$$\leq \frac{4}{q_n}\mathbb{E}[|Q_{T\Pi_m}|^2(Y_1)b(Y_1)] \leq \frac{4}{q_n}\|Q_{T\Pi_m}\|_\infty(\mathbb{E}[|Q_{T\Pi_m}|^2(Y_1)])^{1/2}(\mathbb{E}[b^2(Y_1)])^{1/2}$$
$$\leq \frac{4\sqrt{2\sum_k (k+1)\beta_k}}{q_n}\|Q_{T\Pi_m}\|_\infty(\mathbb{E}[|Q_{T\Pi_m}|^2(Y_1)])^{1/2} \tag{5.18}$$

We have already proved that $\|Q_{T\Pi_m}\|_\infty \leq f_0^{-1/2}\|\Pi\|_A\sqrt{\Phi_1 2^{2\gamma+4}/(2^{2\gamma+4}-1)}D_{m''}^{\gamma+2}$.

Now we need a sharp bound on $\mathbb{E}[|Q_{T\Pi_m}(Y_1)|^2]$. We have

$$\mathbb{E}[|Q_{T\Pi_m}(Y_1)|^2] \leq \|q\|_\infty \int |Q_{T\Pi_m}|^2 = \frac{\|q\|_\infty}{2\pi}\int \left|\frac{(T\Pi_m)^*(u,0)}{q^*(-u)}\right|^2 du$$

Then it follows from the Schwarz inequality that

$$\mathbb{E}[|Q_{T\Pi_m}(Y_1)|^2] \leq \frac{\|q\|_\infty}{2\pi}\sqrt{\int \frac{|(T\Pi_m)^*(u,0)|^2}{|q^*(-u)|^4}du}\sqrt{\int |(T\Pi_m)^*(u,0)|^2 du}$$

We will evaluate the two terms under the square roots. First observe that

$$(T\Pi_m)^*(u,0) = \sum_{jkl}\sum_{j'k'l'} a_{jkl}b_{j'k'l'}(\varphi_{jk}\varphi_{j'k'})^*(u)(\varphi_{jl}\varphi_{j'l'})^*(0) = \sum_{jkk'l} a_{jkl}b_{jk'l}(\varphi_{jk}\varphi_{jk'})^*(u)$$

since $(\varphi_{jl}\varphi_{j'l'})^*(0) = \int \varphi_{jl}\varphi_{j'l'} = \mathbb{1}_{j=j',l=l'}$. Then

$$\begin{aligned}
\int |(T\Pi_m)^*(u,0)|^2 du &\leq \sum_{jkk'l} a_{jkl}^2 b_{jk'l}^2 \sum_{jkk'l}\int |(\varphi_{jk}\varphi_{jk'})^*(u)|^2 du \\
&\leq 2\pi f_0^{-1}\|\Pi\|_A^2 \sum_{jkk'l}\int |(\varphi_{jk}\varphi_{jk'})(u)|^2 du \\
&\leq 2\pi f_0^{-1}\|\Pi\|_A^2 \sum_j 2^j \| \sum_{k'} |\varphi_{jk'}|^2\|_\infty \sum_{k=k'-2N+2}^{k'+2N-2}\int |\varphi_{jk'}|^2
\end{aligned}$$

by taking account of the superposition of the supports. Using now (5.8)

$$\begin{aligned}
\int |(T\Pi_m)^*(u,0)|^2 du &\leq 2\pi f_0^{-1}\|\Pi\|_A^2 \sum_j 2^j C'(\varphi) 2^j (4N-3) \\
&\leq 2\pi f_0^{-1}\|\Pi\|_A^2 \Phi_1 (4N-3)\frac{2}{3}D_{m''}^2.
\end{aligned}$$

Now

$$\begin{aligned}
\int \frac{|(T\Pi_m)^*(u,0)|^2}{|q^*(-u)|^4}du &\leq \sum_{jkk'l} a_{jkl}^2 b_{jk'l}^2 \sum_{jkk'l}\int \frac{|(\varphi_{jk}\varphi_{jk'})^*(u)|^2}{|q^*(-u)|^4}du \\
&\leq f_0^{-1}\|\Pi\|_A^2 \sum_{jkk'l}\int 2^j \frac{|(\varphi_{jk}\varphi_{jk'})^*(2^j v)|^2}{|q^*(-2^j v)|^4}dv.
\end{aligned}$$

Hence, inequality (5.9) and Assumption H5 show that

$$\begin{aligned}
\int \frac{|(T\Pi_m)^*(u,0)|^2}{|q^*(-u)|^4}du &\leq f_0^{-1}\|\Pi\|_A^2 \sum_{jkk'l}\int 2^j C_r^2 \left[|v|^{2(1-r)}\mathbb{1}_{|v|>1} + \mathbb{1}_{|v|\leq 1}\right] k_0^{-4}((2^j v)^2+1)^{2\gamma}dv \\
&\leq f_0^{-1}\|\Pi\|_A^2 C_r^2 k_0^{-4} C \sum_{jkk'l}(2^j)^{4\gamma+1}
\end{aligned}$$

with $C = \int \left[|v|^{2(1-r)}\mathbb{1}_{|v|>1} + \mathbb{1}_{|v|\leq 1}\right](v^2+1)^{2\gamma}dv < \infty$ as soon as $r > 2\gamma + 3/2$. Then

$$\begin{aligned}\int \frac{|(T\Pi_m)^*(u,0)|^2}{|q^*(-u)|^4}du &\leq f_0^{-1}\|\Pi\|_A^2 C_r^2 k_0^{-4} C \sum_{j=J}^{m}\sum_{k'l}\sum_{k=k'-2N+2}^{k'+2N-2}(2^j)^{4\gamma+1} \\ &\leq f_0^{-1}\|\Pi\|_A^2 C_r^2 k_0^{-4} C 3(4N-3)\frac{2^{4\gamma+3}}{2^{4\gamma+3}-1}D_{m''}^{4\gamma+3}\end{aligned}$$

Finally

$$\mathbb{E}[|Q_{T\Pi_m}(Y_1)|^2] \leq \frac{\|q\|_\infty}{\pi} f_0^{-1}\|\Pi\|_A^2(4N-3)C_r k_0^{-2}\sqrt{C\frac{2^{4\gamma+3}}{2^{4\gamma+3}-1}\pi\Phi_1}D_{m''}^{2\gamma+5/2}$$

Then (5.18) gives

$$v = \sqrt{2\sum_k (k+1)\beta_k\|q\|_\infty f_0^{-1}\|\Pi\|_A^2 k_0^{-1}C(\gamma,r,N,\Phi_1)}\frac{D_{m''}^{2\gamma+13/4}}{q_n}.$$

Then replacing n by p_n in inequality (5.19) gives

$$\begin{aligned}\mathbb{E}[\sup_{T\in B_f(m,m')}|Z_n^{(2,1)}(T)|^2 - 6H^2]_+ &\leq C\left(\frac{v}{p_n}e^{-k_1\frac{p_n H^2}{v}} + \frac{M_1^2}{p_n^2}e^{-k_2\frac{p_n H}{M_1}}\right) \\ &\leq C'\left(\frac{D_{m''}^{2\gamma+13/4}}{n}e^{-k_1' D_{m''}^{1/4}} + \frac{D_{m''}^{2\gamma+4}q_n^2}{n^2}e^{-k_2'\frac{\sqrt{n}}{q_n D_{m''}^{1/4}}}\right)\end{aligned}$$

where C' and k_1' depend on $r, N, \gamma, \Phi_1, f_0, \|\Pi\|_A, \|q\|_\infty, \sum_k (k+1)\beta_k$ and $\sum_k \beta_k$. But there exists a positive constant K such that

$$\sum_{m'\in\mathcal{M}_n} D_{m''}^{2\gamma+13/4}e^{-k_1' D_{m''}^{1/4}} \leq K.$$

Moreover $D_{m''}^{1/4} \leq n^{1/8}$ and $q_n \leq n^c$ with $c + 1/8 < 1/2$, which involves

$$\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left[\sup_{T\in B_f(m,m')}|Z_n^{(2,1)}(T)|^2 - K_2\|\Pi\|_A^2\left(\frac{D_{m''}^{2\gamma+7/2}}{n} + (\sum_k\beta_k)\frac{D_{m''}^3}{n}\right)\right]_+ \leq \frac{C''}{n}$$

with $K_2 = 12 f_0^{-1}\Phi_1 \max(\|q\|_\infty 2^{2\gamma+3}/(2^{2\gamma+3}-1), \Phi_1)$. Thus, if $p_2(m,m') = p_2^{(1)}(m,m') + p_2^{(2)}(m,m')$ with $p_2^{(1)}(m,m') = K_2\|\Pi\|_A^2 D_{m''}^{2\gamma+7/2}/n$ and $p_2^{(2)}(m,m') = K_2\|\Pi\|_A^2(\sum_k\beta_k)D_{m''}^3/n$, then

$$\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left(\left[\sup_{T\in B_f(m,m')}Z_n^{(2)}(T)^2 - p_2(m,m')\right]_+\mathbb{1}_\Omega\right) \leq \frac{C_2}{n}.$$

□

5.6.8 Technical Lemmas

Lemma 5.4 *If $|\varphi^*(x)| \leq k_3(x^2+1)^{-r/2}$ for all real x then*

- *if s and α are reals such that $sr > \alpha + 1$*

$$\int |\varphi^*(x)|^s (x^2+1)^{\alpha/2} dx \leq C_{s,\alpha} < \infty$$

- *if $r > 1$*

$$\int |\varphi^*(y)\varphi^*(x-y)| dy \leq C_r(|x|^{1-r}\mathbb{1}_{|x|>1} + \mathbb{1}_{|x|\leq 1})$$

Proof of Lemma 5.4:

• For the first point, it is sufficient to observe that the function $(x^2+1)^{(-rs+\alpha)/2}$ is integrable if $-rs + \alpha > -1$.

• By changing the variable ($y = xu$), we get

$$\begin{aligned}
&\int |\varphi^*(y)\varphi^*(x-y)| dy = \int |\varphi^*(xu)\varphi^*(x(1-u))| x du \\
&\leq \int_{|u|>1/3 \text{ and } |1-u|>1/3} k_3|xu|^{-r} k_3 |x(1-u)|^{-r} |x| du \\
&\quad + \int_{|u|\leq 1/3} k_3^2 |x(1-u)|^{-r}|x| du + \int_{|1-u|\leq 1/3} k_3^2 |xu|^{-r}|x| du \\
&\leq k_3^2 3^r |x|^{1-2r} \int_{|u|>1/3} \frac{du}{|u|^r} + k_3^2 |x|^{1-r} \frac{2}{3}\left|\frac{3}{2}\right|^r + k_3^2|x|^{1-r}\frac{2}{3}\left|\frac{3}{2}\right|^r \\
&\leq k_3^2 \left[\frac{2.3^{2r-1}}{r-1}|x|^{1-2r} + 2^{2-r}3^{r-1}|x|^{1-r}\right]
\end{aligned}$$

Thus, if $|x| > 1$, $\int |\varphi^*(y)\varphi^*(x-y)| dy \leq C_r |x|^{1-r}$ and if $|x| \leq 1$, $\int |\varphi^*(y)\varphi^*(x-y)| dy \leq C_r$ with $C_r = k_3^2(2.3^{2r-1}/(r-1) + 2^{2-r}3^{r-1})$.

□

Lemma 5.5 *Let $T_1, \ldots, T_n$ be independent random variables and*

$$\nu_n(r) = (1/n)\sum_{i=1}^n [r(T_i) - \mathbb{E}(r(T_i))],$$

for r belonging to a countable class $\mathcal{R}$ of measurable functions. Then, for $\epsilon > 0$,

$$\mathbb{E}[\sup_{r\in\mathcal{R}} |\nu_n(r)|^2 - 6H^2]_+ \leq C\left(\frac{v}{n}e^{-k_1\frac{nH^2}{v}} + \frac{M_1^2}{n^2}e^{-k_2\frac{nH}{M_1}}\right) \tag{5.19}$$

with $k_1 = 1/6$, $k_2 = 1/(21\sqrt{2})$ and C a universal constant and where

$$\sup_{r\in\mathcal{R}} \|r\|_\infty \leq M_1, \quad \mathbb{E}\left(\sup_{r\in\mathcal{R}} |\nu_n(r)|\right) \leq H, \quad \sup_{r\in\mathcal{R}} \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(r(T_i)) \leq v.$$

Usual density arguments allow to use this result with non-countable class of functions $\mathcal{R}$.

This lemma is proved in Chapter 4 Section 4.7.8.

Lemma 5.6 *(Viennet (1997)) Let (T_i) a strictly stationary process with β-mixing coefficients β_k. Then there exists a function b such that*

$$\mathbb{E}[b(T_1)] \leq \sum_k \beta_k \quad \text{and} \quad \mathbb{E}[b^2(T_1)] \leq 2\sum_k (k+1)\beta_k$$

and for all function ψ (such that $\mathbb{E}[\psi^2(T_1)] < \infty$) and for all N

$$\mathrm{Var}(\sum_{i=1}^{N} \psi(T_i)) \leq 4N\mathbb{E}[|\psi|^2(T_1)b(T_1)].$$

In particular, for functions (ψ_λ), $\sum_\lambda \mathrm{Var}(\sum_{i=1}^N \psi_\lambda(T_i)) \leq 4N(\sum_k \beta_k)\| \sum_\lambda |\psi_\lambda|^2\|_\infty$.

Bibliographie

[1] Ango Nzé, P. (1992). Critères d'ergodicité de quelques modèles à représentation markovienne. *C. R. Acad. Sci. Paris Sér. I Math.*, 315(12), 1301–1304.

[2] Athreya, K. B. and Atuncar, G. S. (1998). Kernel estimation for real-valued Markov chains. *Sankhyā Ser. A*, 60(1), 1–17.

[3] Bakry, D., Milhaud, X., and Vandekerkhove, P. (1997). Statistique de chaînes de Markov cachées à espace d'états fini. Le cas non stationnaire. *C. R. Acad. Sci. Paris Sér. I Math.*, 325(2), 203–206.

[4] Baraud, Y., Comte, F., and Viennet, G. (2001). Adaptive estimation in autoregression or β-mixing regression via model selection. *Ann. Statist.*, 29(3), 839–875.

[5] Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3), 301–413.

[6] Basu, A. K. and Sahoo, D. K. (1998). On Berry-Esseen theorem for nonparametric density estimation in Markov sequences. *Bull. Inform. Cybernet.*, 30(1), 25–39.

[7] Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, 37, 1554–1563.

[8] Bickel, P. J., Ritov, Y., and Rydén, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Ann. Statist.*, 26(4), 1614–1635.

[9] Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete*, 65(2), 181–237.

[10] Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York.

[11] Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves : exponential bounds and rates of convergence. *Bernoulli*, 4(3), 329–375.

[12] Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 38(1), 33–73.

[13] Bosq, D. (1973). Sur l'estimation de la densité d'un processus stationnaire et mélangeant. *C. R. Acad. Sci. Paris Sér. A-B*, 277, A535–A538.

[14] Butucea, C. (2004). Deconvolution of supersmooth densities with smooth noise. *Canad. J. Statist.*, 32(2), 181–192.

[15] Butucea, C. and Matias, C. (2005). Minimax estimation of the noise level and of the deconvolution density in a semiparametric convolution model. *Bernoulli*, 11(2), 309–340.

[16] Butucea, C. and Tsybakov, A. B. (2007). Sharp optimality for density deconvolution with dominating bias. *Theory Probab. Appl.*, 52(1). To appear.

[17] Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in hidden Markov models*. Springer Series in Statistics. Springer, New York.

[18] Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.*, 83(404), 1184–1186.

[19] Chaleyat-Maurel, M. and Genon-Catalot, V. (2006). Computable infinite dimensional filters with applications to discretized diffusions. *Stochastic Process. Appl.*, 116(10), 1447–1467.

[20] Clémençon, S. (1999). *Méthodes d'ondelettes pour la statistique non paramétrique des chaînes de Markov*. PhD thesis, Université Denis Diderot Paris 7.

[21] Clémençon, S. (2000). Adaptive estimation of the transition density of a regular Markov chain. *Math. Methods Statist.*, 9(4), 323–357.

[22] Clémençon, S. (2003). Nonparametric estimation for some specific classes of hidden Markov models. *Preprint Modal'X n° 03-9 : http ://www.u-paris10.fr/65897276/0/fiche__pagelibre/.*

[23] Cohen, A., Daubechies, I., and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.*, 1(1), 54–81.

[24] Comte, F. (2001). Adaptive estimation of the spectrum of a stationary Gaussian sequence. *Bernoulli*, 7(2), 267–298.

[25] Comte, F., Dedecker, J., and Taupin, M.-L. (2006a). Adaptive density deconvolution with dependent inputs. *Preprint MAP5 n° 2006-4 : http ://www.math-info.univ-paris5.fr/map5/publis/titres06.html.*

[26] Comte, F. and Genon-Catalot, V. (2006). Penalized projection estimator for volatility density. *Scand. J. Statist.*, 33(4), 875–895.

[27] Comte, F. and Rozenholc, Y. (2002). Adaptive estimation of mean and volatility functions in (auto-)regressive models. *Stochastic Process. Appl.*, 97(1), 111–145.

[28] Comte, F. and Rozenholc, Y. (2004). A new algorithm for fixed design regression and denoising. *Ann. Inst. Statist. Math.*, 56(3), 449–473.

[29] Comte, F., Rozenholc, Y., and Taupin, M.-L. (2006b). Penalized contrast estimator for adaptive density deconvolution. *Canad. J. Statist.*, 34(3), 431–452.

[30] Dacunha-Castelle, D. and Duflo, M. (1983). *Probabilités et statistiques. Tome 2.* Masson, Paris.

[31] Dalelane, C. (2005). *Data driven kernel choice in nonparametric density estimation.* PhD thesis, Technische Universität Braunschweig. http ://opus.tu-bs.de/opus/volltexte/2005/659/.

[32] DeVore, R. (1998). Nonlinear approximation. *Acta numerica*, pages 51–150.

[33] DeVore, R. and Lorentz, G. (1993). *Constructive approximation.* Springer-Verlag.

[34] Devroye, L. (1989). Consistent deconvolution in density estimation. *Canad. J. Statist.*, 17(2), 235–239.

[35] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2), 508–539.

[36] Doob, J. L. (1953). *Stochastic processes.* John Wiley & Sons Inc.

[37] Dorea, C. C. Y. and Zhao, L. C. (2002). Nonparametric density estimation in hidden Markov models. *Stat. Inference Stoch. Process.*, 5(1), 55–64.

[38] Douc, R., Moulines, É., and Rydén, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Statist.*, 32(5), 2254–2304.

[39] Doukhan, P. (1994). *Mixing. Properties and examples*, volume 85 of *Lecture Notes in Statistics.* Springer-Verlag, New York.

[40] Doukhan, P. and Ghindès, M. (1983). Estimation de la transition de probabilité d'une chaîne de Markov Doëblin-récurrente. Étude du cas du processus autorégressif général d'ordre 1. *Stochastic Process. Appl.*, 15(3), 271–293.

[41] Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, 19(3), 1257–1272.

[42] Fan, J. (1993). Adaptively local one-dimensional subproblems with application to a deconvolution problem. *Ann. Statist.*, 21(2), 600–610.

[43] Genon-Catalot, V., Jeantheau, T., and Larédo, C. (2000). Stochastic volatility models as hidden Markov models and statistical applications. *Bernoulli*, 6(6), 1051–1079.

[44] Gillert, H. and Wartenberg, A. (1984). Density estimation for nonstationary Markov processes. *Math. Operationsforsch. Statist. Ser. Statist.*, 15(2), 263–275.

[45] Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (1998). *Wavelets, approximation, and statistical applications*, volume 129 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

[46] Hernández-Lerma, O., Esparza, S. O., and Duran, B. S. (1988). Recursive nonparametric estimation of nonstationary Markov processes. *Bol. Soc. Mat. Mexicana (2)*, 33(2), 57–69.

[47] Hochmuth, R. (2002). Wavelet characterizations for anisotropic Besov spaces. *Appl. Comput. Harmon. Anal.*, 12(2), 179–208.

[48] Höpfner, R. and Löcherbach, E. (2003). Limit theorems for null recurrent Markov processes. *Mem. Amer. Math. Soc.*, 161(768), vi+92.

[49] Ibragimov, I. A. and Has′minskiĭ, R. Z. (1980). An estimate of the density of a distribution. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)*, 98, 61–85, 161–162, 166. Studies in mathematical statistics, IV.

[50] Jarner, S. F. and Hansen, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stochastic Process. Appl.*, 85(2), 341–361.

[51] Jensen, J. L. and Petersen, N. V. (1999). Asymptotic normality of the maximum likelihood estimator in state space models. *Ann. Statist.*, 27(2), 514–535.

[52] Klein, T. and Rio, E. (2005). Concentration around the mean for maxima of empirical processes. *Ann. Probab.*, 33(3), 1060–1077.

[53] Koo, J.-Y. and Lee, K.-W. (1998). B-spline estimation of regression functions with errors in variable. *Statist. Probab. Lett.*, 40(1), 57–66.

[54] Korostelëv, A. P. and Tsybakov, A. B. (1993). *Minimax theory of image reconstruction*, volume 82 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

[55] Lebarbier, E. (2005). Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, 85, 717–736.

[56] Ledoux, M. (1996). On Talagrand's deviation inequalities for product measures. *ESAIM Probab. Statist.*, 1, 63–87 (electronic).

[57] Lepski, O. V. and Levit, B. Y. (1999). Adaptive nonparametric estimation of smooth multivariate functions. *Math. Methods Statist.*, 8(3), 344–370.

[58] Leroux, B. G. (1992). Maximum-likelihood estimation for hidden Markov models. *Stochastic Process. Appl.*, 40(1), 127–143.

[59] Liebscher, E. (1992). Density estimation for Markov chains. *Statistics*, 23(1), 27–48.

[60] Liu, M. C. and Taylor, R. L. (1989). A consistent nonparametric density estimator for the deconvolution problem. *Canad. J. Statist.*, 17(4), 427–438.

[61] Masry, E. (1993). Strong consistency and rates for deconvolution of multivariate densities of stationary processes. *Stochastic Process. Appl.*, 47(1), 53–74.

[62] Masry, E. and Györfi, L. (1987). Strong consistency and rates for recursive probability density estimators of stationary processes. *J. Multivariate Anal.*, 22(1), 79–93.

[63] Massart, P. (2007). *Concentration Inequalities and Model Selection*. Lecture Notes in Mathematics. Springer.

[64] Matias, C. (2002). Semiparametric deconvolution with unknown noise variance. *ESAIM Probab. Statist.*, 6, 271–292 (electronic). New directions in time series analysis (Luminy, 2001).

[65] Meyer, Y. (1990). *Ondelettes et opérateurs*, volume I. Hermann.

[66] Meyn, S. P. and Tweedie, R. L. (1993). *Markov chains and stochastic stability*. Springer-Verlag, London.

[67] Meyn, S. P. and Tweedie, R. L. (1994). Computable bounds for geometric convergence rates of Markov chains. *Ann. Appl. Probab.*, 4(4), 981–1011.

[68] Mokkadem, A. (1987). Sur un modèle autorégressif non linéaire : ergodicité et ergodicité géométrique. *J. Time Ser. Anal.*, 8(2), 195–204.

[69] Nikol′skiĭ, S. M. (1975). *Approximation of functions of several variables and imbedding theorems*. Springer-Verlag, New York. Translated from the Russian by John M. Danskin, Jr., Die Grundlehren der Mathematischen Wissenschaften, Band 205.

[70] Nummelin, E. (1984). *General irreducible Markov chains and nonnegative operators*. Cambridge University Press, Cambridge.

[71] Nummelin, E. and Tuominen, P. (1982). Geometric ergodicity of Harris recurrent Markov chains with applications to renewal theory. *Stochastic Process. Appl.*, 12(2), 187–202.

[72] Pardoux, E. and Veretennikov, A. Y. (2001). On the Poisson equation and diffusion approximation. I. *Ann. Probab.*, 29(3), 1061–1085.

[73] Pensky, M. and Vidakovic, B. (1999). Adaptive wavelet estimator for nonparametric density deconvolution. *Ann. Statist.*, 27(6), 2033–2053.

[74] Petrov, V. V. (1975). *Sums of independent random variables*. Springer-Verlag, New York.

[75] Prakasa Rao, B. L. S. (1978). Density estimation for Markov processes using delta-sequences. *Ann. Inst. Statist. Math.*, 30(2), 321–328.

[76] Revuz, D. (1984). *Markov chains*, volume 11 of *North-Holland Mathematical Library*. North-Holland Publishing Co., Amsterdam, second edition.

[77] Roberts, G. O. and Rosenthal, J. S. (1998). Markov-chain Monte Carlo : some practical implications of theoretical results. *Canad. J. Statist.*, 26(1), 5–31. With discussion by Hemant Ishwaran and Neal Madras and a rejoinder by the authors.

[78] Rosenblatt, M. (1970). Density estimates and Markov sequences. In *Nonparametric Techniques in Statistical Inference*, pages 199–213. Cambridge Univ. Press, London.

[79] Roussas, G. G. (1969). Nonparametric estimation in Markov processes. *Ann. Inst. Statist. Math.*, 21, 73–87.

[80] Runst, T. and Sickel, W. (1996). *Sobolev spaces of fractional order, Nemytskij operators, and nonlinear partial differential equations*, volume 3 of *de Gruyter Series in Nonlinear Analysis and Applications*. Walter de Gruyter & Co., Berlin.

[81] Stefanski, L. A. (1990). Rates of convergence of some estimators in a class of deconvolution problems. *Statist. Probab. Lett.*, 9(3), 229–235.

[82] Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math.*, 126(3), 505–563.

[83] Tribouley, K. and Viennet, G. (1998). L_p adaptive density estimation in a β mixing framework. *Ann. Inst. H. Poincaré Probab. Statist.*, 34(2), 179–208.

[84] Triebel, H. (1983). *Theory of function spaces*, volume 78 of *Monographs in Mathematics*. Birkhäuser Verlag, Basel.

[85] Tsybakov, A. (2000). On the best rate of adaptive estimation in some inverse problems. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(9), 835–840.

[86] Van Es, B., Spreij, P., and van Zanten, H. (2005). Nonparametric volatility density estimation for discrete time models. *J. Nonparametr. Stat.*, 17(2), 237–251.

[87] Viennet, G. (1996). *Estimation minimax et adaptative dans un cadre absolument régulier*. PhD thesis, Université de Paris-Sud Centre d'Orsay.

[88] Viennet, G. (1997). Inequalities for absolutely regular sequences : application to density estimation. *Probab. Theory Related Fields*, 107(4), 467–492.

[89] Volkonskiĭ, V. A. and Rozanov, Y. A. (1959). Some limit theorems for random functions. I. *Theor. Probability Appl.*, 4, 178–197.

[90] Yakowitz, S. (1989). Nonparametric density and regression estimation for Markov sequences without mixing assumptions. *J. Multivariate Anal.*, 30(1), 124–136.

[91] Youndjé, É. and Wells, M. T. (2002). Least squares cross-validation for the kernel deconvolution density estimator. *C. R. Math. Acad. Sci. Paris*, 334(6), 509–513.

Résumé : Dans cette thèse, on considère une chaîne de Markov (X_i) à espace d'états continu que l'on suppose récurrente positive et stationnaire. L'objectif est d'estimer la densité de transition Π définie par $\Pi(x,y)dy = P(X_{i+1} \in dy | X_i = x)$. On utilise la sélection de modèles pour construire des estimateurs adaptatifs. On se place dans le cadre minimax sur L^2 et l'on s'intéresse aux vitesses de convergence obtenues lorsque la densité de transition est supposée régulière. Le risque intégré de nos estimateurs est majoré grâce au contrôle de processus empiriques par une inégalité de concentration de Talagrand. Dans une première partie, on suppose que la chaîne est directement observée. Deux estimateurs différents sont présentés, l'un par quotient, l'autre minimisant un contraste moindres carrés et prenant également en compte l'anisotropie du problème. Dans une deuxième partie, on aborde le cas d'observations bruitées $Y_1, \ldots, Y_{n+1}$ où $Y_i = X_i + \varepsilon_i$ avec (ε_i) un bruit indépendant de la chaîne (X_i). On généralise à ce cas les deux estimateurs précédents. Des simulations illustrent les performances des estimateurs.

Mots-clefs : estimation adaptative, densité de transition, chaîne de Markov, sélection de modèles, contraste pénalisé, modèle de Markov caché.

Adaptive nonparametric estimation for Markov chains and hidden Markov chains

Abstract : In this thesis, we consider a Markov chain (X_i) with continuous state space which is assumed positive recurrent and stationary. The aim is to estimate the transition density Π defined by $\Pi(X,y)dy = P(X_{i+1} \in dy | X_i = x)$. We use model selection to construct adaptive estimators. We work in the minimax framework on L^2 and we are interested in the rates of convergence obtained when transition density is supposed to be regular. The integrated risk of our estimators is bounded thanks to control of empirical processes by a concentration inequality of Talagrand. In a first part, we suppose that the chain is directly observed. Two different estimators are introduced, one by quotient, the other minimizing a least squares contrast and also taking into account the anisotropy of the problem. In a second part, we treat the case of noisy observations $Y_1, \ldots, Y_{n+1}$ where $Y_i = X_i + \varepsilon_i$ with (ε_i) a noise independent of the chain (X_i). We generalize to this case the two previous estimators. Some simulations illustrate the performances of the estimators.

Keywords : adaptive estimation, transition density, Markov chain, model selection, penalized contrast, hidden Markov model.