Quelques contributions à la statistique des modèles dynamiques aléatoires

Amélie Rosier

Université Paris Nanterre

200 av. de la République
92001 Nanterre Cedex
www.parisnanterre.fr

École Doctorale
Connaissance, langage, modélisation

Unité de recherche
Modélisation Aléatoire de Paris Nanterre

Membre de l'Université Paris Lumières

Amélie ROSIER

Quelques contributions à la statistique des modèles dynamiques aléatoires

Thèse présentée et soutenue publiquement le 10 novembre 2023
en vue de l'obtention du doctorat de Mathématiques appliquées et applications mathématiques
de l'Université Paris Nanterre
sous la direction de Nicolas MARIE (Université Paris Nanterre)
et de Pierre ALQUIER (ESSEC Asia-Pacific)

Jury :

Rapporteur :	M. Mathieu Rosenbaum	Professeur : Ecole Polytechnique
Rapporteur :	M. Joseph Rynkiewicz	Maître de conférences : Université Paris 1
Membre du jury :	M. Nicolas MARIE	Maître de conférences : Université Paris Nanterre
Membre du jury :	M. Pierre Alquier	Professeur : Essec Asia-Pacific
Membre du jury :	Mme Fabienne Comte	Professeure : Université Paris Descartes
Membre du jury :	Mme Adeline Leclercq-Samson	Professeure : Université Grenoble Alpes
Membre du jury :	M. Vincent Rivoirard	Professeur : Université Paris Dauphine
Membre invité :	Mme Emilie Lebarbier	Professeure : Université Paris Nanterre

Remerciements

En ce moment d'accomplissement, je tiens à exprimer ma profonde gratitude envers toutes les personnes qui ont contribué à la réalisation de ce travail de recherche. Le chemin de la thèse a été une aventure exaltante et exigeante, et je suis reconnaissante pour le soutien et l'encouragement que j'ai reçus tout au long de cette période.

Tout d'abord, je souhaite exprimer ma reconnaissance à mes deux directeurs de thèse, Nicolas Marie et Pierre Alquier. Leur expertise, leur dévouement et leur mentorat exceptionnel m'ont guidée tout au long de ce voyage académique. Merci Nicolas, de m'avoir initiée à la recherche et d'avoir éveillé mon intérêt pour ce domaine. Ta passion pour le métier d'enseignant-chercheur m'a toujours profondément inspirée. Depuis les enseignements que tu donnais à l'ESME et que j'ai suivi en tant qu'élève, jusqu'à cette soutenance finale, ta connaissance et ton soutien ont été le fondement sur lequel j'ai construit ma carrière académique. Malgré mes craintes du départ et les nombreuses remises en question que j'ai eues au cours de ma thèse, tu n'as jamais douté de moi ni cessé de m'encourager et je te serais éternellement reconnaissante pour tout ce que tu as fais pour moi. Je n'oublie pas non plus nos longues discussions qui m'ont permises d'avancer et de prendre un peu plus d'assurance au fur et à mesure des mois et des années parcourus. Pierre, un immense merci à toi également, pour ta présence permanente durant ces trois années, même de l'autre côté de la planète. Tes nombreux encouragements, tes conseils avisés et les discussions stimulantes que nous avons eues tous les trois avec Nicolas ont été plus que précieux. Je suis honorée d'avoir eu la chance de travailler avec toi et d'avoir appris à tes côtés. J'espère sincèrement que nos collaborations ne s'arrêteront pas là.

Je tiens également à exprimer ma gratitude envers toutes les personnes avec lesquelles j'ai eu le privilège de collaborer ou de rédiger des articles au cours de ma thèse. Je remercie tout d'abord Fabienne Comte, professeure à l'Université Paris Descartes, qui a été également une des premières personnes à me faire découvrir le monde de la recherche en m'encadrant lors de mon stage de master 2. Je garde un très bon souvenir de cette première expérience et c'est un honneur de la compter aujourd'hui parmi les membres de mon jury de thèse. Un immense merci à Emilie Lebarbier, professeure à l'Université Paris Nanterre, qui m'a fait découvrir un autre aspect des statistiques sur lequel je prend aujourd'hui énormément de plaisir à travailler. Son ouverture d'esprit, ses nombreux conseils et les enseignements qu'elle m'a transmise à travers nos différentes collaborations ont, sans le moindre doute, joué un rôle précieux dans l'accomplissement de cette thèse. Je remercie également Vincent Brault, maître de conférence à l'Université Grenoble Alpes, avec qui j'ai la chance de travailler depuis quelques mois et qui m'a apportée son aide précieuse plus d'une fois.

Je suis profondément reconnaissante envers l'ESME, l'école d'ingénieurs qui a généreusement financé ma thèse et m'a permise de m'épanouir en tant qu'enseignante durant ces trois dernières années. Je remercie particulièrement Véronique Bonnet, directrice générale de l'ESME, Aude Herry, directrice de la recherche au sein de l'école et Sonia Jeanson, en charge du développement des formations,

pour leur confiance et leurs encouragements. Les ressources, les opportunités et le cadre de cette école ont été essentiels pour la réalisation de cette thèse. De plus, je remercie chaleureusement l'Université Paris Nanterre de m'avoir acceuillie en tant que doctorante et plus particulièrement le laboratoire MODAL'X pour son environnement de recherche stimulant dans lequel j'ai pu évoluer.

Je remercie mes deux rapporteurs de thèse, Mathieu Rosenbaum, professeur à l'Ecole Polytechnique, et Joseph Rynkiewicz, maître de conférences à l'Université Paris 1, pour leurs lectures avisées de mon mémoire, leurs commentaires éclairés et leur expertise. Merci également à Adeline Leclerq-Samson, professeure à l'Université Grenobles Alpes, et Vincent Rivoirard, professeur à l'Université Paris Dauphine, d'avoir accepté, eux-aussi, de faire parti de mon jury de thèse.

Je n'oublie pas non plus mes amis et tout ceux avec qui j'ai pu partager des moments précieux durant ces trois dernières années. Une pensée très spéciale pour Hélène Halconruy, maître de conférences à Télécom SudParis mais avant tout amie, qui m'encourage sans relâche depuis le début de cette aventure. Nos longues balades dans Paris à discuter de tout (et de rien) m'ont remontée le moral plus d'une fois, je te dis donc aujourd'hui un immense merci pour ton soutien irremplaçable. Merci à mon ami Mounir Lahlouh, doctorant à l'ESME, avec qui j'ai partagé mon bureau pendant toutes ces années. Tu a été une magnifique rencontre pendant ma thèse, merci pour ta bonne humeur constante, ta bienveillance et ton humour. Merci également à Maxime Ossonce, enseignant et chercheur à l'ESME, pour ses conseils judicieux et les dicussions pertinentes que j'ai pu partager avec lui et qui m'ont toujours aidé.

Je terminerai ces remerciements en rendant hommage aux personnes les plus importantes dans ma vie : mes parents et mes frères. Merci à ma mère, Véronique, et mon père, Christophe, pour leur amour et leur soutien inconditionnel. Votre présence constante m'a toujours aidée à relever les défis qui se sont mis sur ma route, dont celui de la thèse. Enfin, je tiens à remercier mes deux grands-frères, Alexandre et Adrien, qui m'ont toujours entourée et me donnent, depuis que je suis née, la force de persévérer dans les moments difficiles. Merci du fond du coeur.

Table des matières

Chapitre 1

Introduction

Cette thèse de doctorat aborde deux grandes notions. La première porte sur des modèles de débruitage puis de complétion de séries temporelles multidimensionnelles de grande taille représentées par des matrices. Nous proposons un package R dans lequel une méthode de débruitage existante a été implémentée. En complétion, nous montrons, moyennant une condition sur la matrice de bruit ainsi que sur la composante déterministe des séries temporelles, des bornes de risque pour l'estimateur de la matrice plus fines que celles connues en l'absence de structure de série temporelle. La seconde notion abordée dans cette thèse porte sur l'estimation dans des modèles d'équations différentielles stochastiques (EDS). Nous nous concentrons sur un estimateur non-paramétrique de la fonction de drift d'une EDS dirigée par un mouvement Brownien. Nous proposons plusieurs versions d'un estimateur de Nadaraya-Watson calculées à partir de copies indépendantes du processus de diffusion et établissons des bornes de risque pour ces estimateurs.

Ces dernières décénnies, la modélisation des systèmes dynamiques aléatoires est devenue de plus en plus populaire dans de nombreux domaines, notamment en biologie, en économie, en finance, en ingénierie ou encore en sciences sociales. Cela est en partie dû au développement de puissants outils informatiques qui permettent de collecter et d'analyser de gros volumes de données afin de mieux comprendre et prédire le comportement de systèmes dynamiques complexes. Plusieurs approches sont possibles lorsqu'il s'agit d'étudier le comportement de tels systèmes. Nous en détaillons deux dans cette thèse : nous nous concentrons, d'une part, sur l'étude de séries temporelles multidimensionnelles de grande dimension représentées par des matrices et, d'autre part, sur l'étude des systèmes dynamiques modélisés par des des équations différentielles stochastiques. Nous cherchons à estimer, dans les deux cas, la tendance du processus intervenant le système dynamique étudié. Enfin, il est important de noter que dans les deux cas, l'approche choisie permet de considérer des séries non-stationnaires.

Notre travail est organisé en trois parties distinctes. Nous commencerons par une présentation générale des concepts étudiés au cours de cette thèse et introduirons les différentes problématiques abordées. Ensuite, dans la deuxième partie, nous examinerons de manière plus approfondie les méthodes d'estimation dans

des modèles de matrices de séries temporelles de grande dimension. Enfin, dans la troisième partie, nous nous concentrerons sur un estimateur spécifique, à savoir l'estimateur de Nadaraya-Watson pour la fonction de drift d'une équation différentielle stochastique, calculé à partir de copies indépendantes du processus de diffusion.

1.1 Estimation dans des modèles de séries temporelles de grande dimension

Depuis plusieurs années, les modèles de séries temporelles font l'objet de nombreuses études car ils permettent de mieux comprendre le fonctionnement de nombreux systèmes dynamiques aléatoires (cf. [5, 6]). Depuis les années 1970, les modèles classiquement utilisés sont basés sur l'hypothèse que le comportement d'une série temporelle à un instant donné peut être expliqué par ses valeurs passées, ses erreurs passées ou bien les deux. Par exemple, les modèles ARMA, les modèles GARCH et toutes leurs extensions sont définis de cette manière là (cf. [12]). Nous ne détaillerons pas les équations qui définissent chacun de ces modèles, mais d'une façon générale, une série temporelle univariée peut être modélisée par un processus $(X_t)_{t\in\mathbb{Z}}$ satisfaisant une équation du type

$$F_\theta(X_{t+q}, \dots, X_{t-q}, \eta_{t+p}, \dots, \eta_{t-p}) = 0 \ ; \ t \in \mathbb{Z}, \tag{1.1}$$

où $p, q \in \mathbb{N}$, $\eta = (\eta_t)_{t\in\mathbb{Z}}$ est un processus stationnaire du second ordre, souvent un bruit blanc, et $\mathcal{F} = (F_\theta)_{\theta\in\Theta}$ est une famille de fontions continues de $\mathbb{R}^{2(p+q+1)}$ à valeurs dans $\mathbb{R}$ et indexées dans un ensemble Θ. Un des avantages du modèle (1.1) est que l'on peut choisir $\mathcal{F}$ de sorte à prendre en compte les propriétés connues sur la dynamique du phénomène sous-jacent modélisé. Ces modèles sont notamment utilisés en économie pour modéliser et prédire certaines variables économiques telles que le produit intérieur brut, l'inflation ou encore le chômage (cf. [8]). En santé publique, par exemple, l'incidence des maladies infectieuses telle que la grippe et la détection des épidémies sont également modélisées de cette façon (cf. [9, 10]). Cependant, la majorité des travaux sur les séries temporelles se sont concentrés sur des modèles univariés à l'exception de quelques uns tels que le modèle VCEM ou VARMA (cf. [11, 7]). Bien que très souvent utilisés, ces modèles classiques présentent certaines limites et ne s'appliquent pas à des problématiques plus compliquées. Il est également difficile de contourner la condition de stationnarité sur η. L'approche matricielle est, quant à elle, particulièrement facile à étendre au cadre multidimensionnel et a déjà fait l'objet d'une étude théorique approfondie depuis les dix dernières années pour des séries statistiques non nécessairement temporelles. C'est en particulier la popularité du challenge Netflix, avec les algorithmes de recommandations de films, qui est à l'origine de nombreuses recherches dans ce domaine (cf. [13, 22, 25]). Dans ce contexte, il s'agit d'étudier une matrice dont chaque ligne représente un utilisateur Netflix et chaque colonne correspond à un film de la plateforme, comme illustré dans le tableau ci-dessous.

A la position (i, j) de la matrice, on retrouve la note rentrée par l'utilisateur i sur le film j. Dans ce cas, les algorithmes de complétion permettent de prédire les entrées manquantes de la matrice en estimant les notes que chaque utilisateur pourrait mettre aux films non visionnés. Cela permet ainsi aux algorithmes de recommandation de proposer, selon le profil de l'utilisateur, des films qu'il est susceptible d'apprécier.

	Film 1	Film 2	Film 3	$\cdots$	Film T
Utilisateur 1	9	1	3	$\cdots$	8
Utilisateur 2	?	6	?	$\cdots$	7
Utilisateur 3	9	1	?	$\cdots$	8
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
Utilisateur d	8	?	3	$\cdots$	?

Les premiers articles théoriques sur le sujet traitent de méthodes de factorisation et de complétion de matrices de faible rang ayant des entrées indépendantes et identiquement distribuées (cf. [14, 15, 16, 17]), ce qui est notamment le cas avec les algorithmes de recommandation. L'objectif, dans ce cas, est d'estimer une matrice donnée par un produit de deux matrices de rang inférieur, dont les lignes de l'une d'entre elles peuvent être interprétées comme des facteurs latents et ainsi capturer l'information essentielle dans les données. En d'autres termes, si une matrice est de grande dimension, l'hypothèse de faible rang suggère qu'elle peut être représentée comme une combinaison linéaire de vecteurs de base de plus petite dimension. L'approche matricielle présente ainsi un avantage certain car cette hypothèse de faible rang permet de réduire la dimension du problème et donc la complexité de l'analyse liée à la manipulation de matrice de grande taille. Au-delà des systèmes de recommandation, les méthodes de factorisation de matrices de faible rang s'appliquent avec succès dans des domaines variés tels que l'estimation de données manquantes (cf. [26]), l'analyse de données génomiques (cf. [27]), le traitement du signal (cf. [28]) ou encore l'analyse de données textuelles (cf. [29]). Cependant, quelque soit le domaine étudié, les méthodes de factorisation et de complétion de matrices existantes ont été developpées dans un contexte où les observations sont indépendantes les unes des autres, ce qui n'est pas le cas lorsque l'on étudie des séries temporelles. Par ailleurs, aucune justification théorique n'a été fournie sur le fait qu'elles peuvent être utilisées pour des observations dépendantes. L'objet de nos travaux était donc d'étendre ces méthodes existantes à l'analyse de séries temporelles de grande dimension, un cas particulier de données dépendantes. En effet, dans notre travail, on étudie cette fois des matrices dont chaque ligne représente une série temporelle et chaque colonne correspond à une observation des séries dans le temps :

	Temps 1	Temps 2	Temps 3	$\cdots$	Temps T
Série 1	9	1	3	$\cdots$	8
Série 2	?	6	?	$\cdots$	7
Série 3	9	1	?	$\cdots$	8
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
Série d	8	?	3	$\cdots$	?

L'objectif reste le même, reconstruire la matrice pour laquelle on ne dispose que de certaines entrées, tout en prenant en compte cette fois la difficulté que peut engendrer la structure de séries temporelle dans les lignes et notamment la dépendance qu'il va y avoir entre les colonnes de la matrice.

1.1.1 Débruitage de séries temporelles de grande dimension

Avant de rentrer dans les détails de la complétion, abordons un des principaux problèmes que l'on peut rencontrer lorsqu'on étudie des séries temporelles : le débruitage. Notons $\mathbf{M}$ la matrice observée de taille $d \times T$ et dont chaque ligne est une série temporelle. Le but du débruitage est d'estimer la matrice $\mathbf{\Theta}^0$ dans le modèle suivant

$$\mathbf{M} = \mathbf{\Theta}^0 + \varepsilon \tag{1.2}$$

où $\mathbf{\Theta}^0$ et ε sont des matrices de taille $d \times T$. On réécrit ici la matrice $\mathbf{M}$ comme la somme d'une matrice deterministe $\mathbf{\Theta}^0$ de faible rang $k \in \mathbb{N}^*$ ($k \ll d \wedge T$) et d'une matrice aléatoire ε. De plus, en utilisant les méthodes de factorisation, la matrice $\mathbf{\Theta}^0$ peut s'écrire comme le produit $\mathbf{UV}$ de deux matrices $\mathbf{U} \in \mathcal{M}_{d,k}(\mathbb{R})$ et $\mathbf{V} \in \mathcal{M}_{k,T}(\mathbb{R})$. La matrice d'erreur ε, quant à elle, a des lignes centrées, indépendantes, identiquement distribuées, sous-gaussiennes et a pour matrice de covariance Σ_ε, prenant ainsi en compte la dépendance temporelle entre les différentes colonnes du bruit. La matrice $\mathbf{\Theta}^0$ est estimée en résolvant le problème des moindres carrés suivant :

$$\widehat{\mathbf{\Theta}} \in \arg \min_{\mathbf{U} \in \mathcal{M}_{d,k}(\mathbb{R}), \mathbf{V} \in \mathcal{M}_{k,T}(\mathbb{R})} \|\mathbf{M} - \mathbf{UV}\|_{\mathcal{F}}^2 \tag{1.3}$$

où, pour toute matrice $\mathbf{A}$, $\|\mathbf{A}\|_{\mathcal{F}}^2 := \text{trace}(\mathbf{AA}^*)$. De plus, du fait de la présence de séries temporelles dans $\mathbf{M}$, les auteurs de [24] ont montré que l'estimateur de $\mathbf{\Theta}^0$ pouvait être ajusté pour prendre en compte une éventuelle périodicité ou régularité dans les tendances des séries étudiées. Pour ce faire, ils ont réécrit la matrice $\mathbf{\Theta}^0$ de la façon suivante : $\mathbf{\Theta}^0 := \mathbf{T}^0\mathbf{\Lambda}$ où $\mathbf{T}^0 := \mathbf{UV}$ avec $\mathbf{U} \in \mathcal{M}_{d,k}(\mathbb{R})$, $\mathbf{V} \in \mathcal{M}_{k,\tau}(\mathbb{R})$ et $\mathbf{\Lambda}$ est une matrice connue de taille $\tau \times T$, ($\tau \leqslant T$), reflétant les propriétés liées à la structure temporelle des données. Plus précisemment, deux cas ont été étudiés dans leurs travaux, chacun d'entre eux impliquant une définition différente pour la matrice $\mathbf{\Lambda}$:

- Si les tendances des séries temporelles sont **τ-périodiques**, alors,

  $$\mathbf{\Lambda} = (\mathbf{I}_\tau | \dots | \mathbf{I}_\tau)$$

  où $\mathbf{I}_\tau$ désigne la matrice identité d'ordre τ.

- Si les tendances séries temporelles sont **décomposables dans la base de Fourier**, alors,

  $$\mathbf{\Lambda} = \left(\varphi_l \left(\frac{t}{T} \right) \right)_{(l,t) \in \{1,\dots,\tau\} \times \{1,\dots,T\}}$$

où τ est impair et $(\varphi_1, \ldots, \varphi_\tau)$ est la base trigonométrique définie, pour tous $x \in [0,1]$ et $m \in \{1, \ldots, (\tau-1)/2\}$, par

$$\varphi_l(x) := \begin{cases} 1 & \text{si} \quad l = 1 \\ \sqrt{2}\cos(2\pi m x) & \text{si} \quad l = 2m \\ \sqrt{2}\sin(2\pi m x) & \text{si} \quad l = 2m+1. \end{cases}$$

Dans le cas où nos données ne présentent aucune structure particulière, $\mathbf{\Theta}^0 := \mathbf{T}^0$, $\tau = T$ et $\mathbf{\Lambda} = \mathbf{I}_T$. En multipliant de part et d'autre l'équation $\mathbf{M} = \mathbf{\Theta}^0 + \varepsilon = \mathbf{T}^0\mathbf{\Lambda} + \epsilon$ par le pseudo-inverse $\mathbf{\Lambda}^+ = \mathbf{\Lambda}^*(\mathbf{\Lambda}\mathbf{\Lambda}^*)^{-1}$, les auteurs de [24] ont obtenu le modèle de débruitage suivant :

$$\underline{\mathbf{M}} = \mathbf{T}^0 + \underline{\varepsilon} \tag{1.4}$$

où $\underline{\mathbf{M}} := \mathbf{M}\mathbf{\Lambda}^+$ et $\underline{\varepsilon} := \varepsilon\mathbf{\Lambda}^+$. Ainsi, un estimateur naturel de $\mathbf{\Theta}^0$ est donné par $\widehat{\mathbf{\Theta}}_{k,\tau} := \widehat{\mathbf{T}}_{k,\tau}\mathbf{\Lambda}$ où

$$\widehat{\mathbf{T}}_{k,\tau} \in \arg\min_{\mathbf{U}\in\mathcal{M}_{d,k}(\mathbb{R}),\mathbf{V}\in\mathcal{M}_{k,\tau}(\mathbb{R})} \|\underline{\mathbf{M}} - \mathbf{U}\mathbf{V}\|_{\mathcal{F}}^2. \tag{1.5}$$

En considérant ce modèle, ils ont réussi à montrer que, sur le plan théorique, la prise en compte de structures temporelles dans les données permettait d'améliorer les bornes de risques existantes pour l'estimateur de la matrice $\mathbf{\Theta}^0$. Les bornes de risque obtenues font notamment intervenir le rang k de la matrice $\mathbf{\Theta}^0$, ainsi que les paramètres liés à la structure temporelle des données (le paramètre τ par exemple, si l'on considère que nos séries sont τ-périodiques). En pratique, ces paramètres peuvent être inconnus et font donc l'objet d'une procédure de séléction de modèle. La sélection conjointe de deux paramètres génère des problèmes numériques auxquels nous nous sommes intéressés dans le premier article présenté dans cette thèse (cf. section **2.1**). Le choix du rang k est une étape cruciale car il affecte directement la qualité de la factorisation et sélectionner un mauvais rang peut entraîner un problème de sur-ajustement ou de sous-ajustement sur des données. Dans la littérature, plusieurs méthodes ont été developpées lorsque l'on cherche à estimer le rang k seulement. Parmi ces méthodes, nous pouvons citer par exemple celles basées sur des critères d'information tels que le critère d'Akaike (AIC) ou le critère d'information bayésien (BIC). L'idée, dans ce cas, est de calculer le critère considéré pour différents rangs et de choisir le rang k pour lequel le critère est minimal (cf. [30, 31, 32]). Une autre méthode que l'on peut envisager consiste à tracer les valeurs singulières de la matrice en fonction du rang et à sélectionner le rang k à partir duquel la courbe commence à s'aplanir ou se stabiliser (cf. [33, 34]). Cette méthode est simple et efficace, mais peut ne pas être optimale dans certains cas et nécessite une analyse approfondie des données. Enfin, les méthodes classiques de validation croisée permettent également de sélectionner le rang ; plusieurs travaux ont déjà été réalisés dans ce domaine (cf. [35, 36]). Toutefois, ces dernières peuvent s'avérer coûteuses en termes de calcul, en particulier pour des données et des matrices de grande dimension. Ces méthodes, non exhaustives, sont utilisées pour la sélection d'un unique paramètre. Dans notre étude, nous nous sommes

intéressés à la sélection conjointe de deux paramètres, ce qui n'est pas une question standard en pratique. La procédure d'estimation décrite dans nos travaux (cf. section **2.1**) est la suivante : en considérant le modèle (1.4) présenté ci-dessus, l'estimateur adaptatif final de $\mathbf{\Theta^0}$ sera donné, pour une constante $s > 0$ fixée, par $\widehat{\mathbf{\Theta}}_s := \widehat{\mathbf{\Theta}}_{\widehat{k}(s),\widehat{\tau}(s)}$ où

$$(\widehat{k}(s), \widehat{\tau}(s)) \in \arg \min_{(k,\tau)\in\mathcal{K}\times\mathcal{T}} \{\|\mathbf{M} - \widehat{\mathbf{\Theta}}_{k,\tau}\|_{\mathcal{F}}^2 + \mathrm{pen}_s(k,\tau)\} \tag{1.6}$$

avec $\mathcal{K} \subset \{1, \dots, d \wedge T\}$, $\mathcal{T} \subset \{1, \dots, T\}$ et $\widehat{\mathbf{\Theta}}_{k,\tau}$ l'estimateur solution de (1.5) avec des paramètres k et τ fixés. L'estimateur $\widehat{\mathbf{\Theta}}_s$ est donc obtenu en minimisant une fonction de coût qui prend en compte la complexité du modèle. La pénalité qui intervient dans ce critère pénalisé permet notamment d'éviter les problèmes de sur-ajustement (*overfitting*) et est définie de la façon suivante :

$$\mathrm{pen}_s(k,\tau) := \mathfrak{c}_{\mathrm{pen}} \|\mathbf{\Sigma}_\varepsilon\|_{\mathrm{op}} k(d + \tau + s) \; ; \quad \forall (k,\tau) \in \mathcal{K} \times \mathcal{T}, \tag{1.7}$$

où $\mathfrak{c}_{\mathrm{pen}} > 0$ est une constante déterministe et $\|.\|_{\mathrm{op}}$ est la norme d'opérateur sur $\mathcal{M}_T(\mathbb{R})$. Cette pénalité, de l'ordre du terme de variance dans la borne de risque de $\widehat{\mathbf{\Theta}}_{k,\tau}$, dépend de plusieurs paramètres :

La constante s : elle est assez facile à choisir en pratique puisqu'on la fixe en fonction du niveau de confiance α intervenant dans la borne de risque obtenue par Alquier et al. dans [24]. Plus précisemment, on prend $s = -\log((1-\alpha)/2)$ avec α fixé à 99%, 95% ou 90%.

La constante $\mathfrak{c}_{\mathrm{pen}}$: elle nécessite d'être calibrée en amont en utilisant des heuristiques classiques développées dans la littérature.

$\|\mathbf{\Sigma}_\varepsilon\|_{\mathrm{op}}$: la norme d'opérateur de la matrice de covariance du bruit, que l'on peut expliciter en fonction des paramètres de la distribution du bruit.

L'un des aspects de nos travaux est alors de proposer une heuristique en deux étapes pour sélectionner à la fois le rang k de la matrice et le paramètre τ lié à la structure temporelle des données. Pour cela, on introduit une constante de pénalité *globale*, que l'on va devoir calibrer avant de procéder à la sélection de k et τ : $\mathfrak{c}_{\mathrm{cal}} = \mathfrak{c}_{\mathrm{pen}} \|\mathbf{\Sigma}_\varepsilon\|_{\mathrm{op}}$. Plus précisemment, en reprenant le modèle (1.6), nous proposons, en se basant sur la stratégie développée par Devijver et al. dans [42], une heuristique en deux étapes pour séléctionner à la fois le rang k et le paramètre τ de cette façon :

(1) Pour chaque rang $k \in \mathcal{K} \subset \{1, \dots, d \wedge T\}$, nous sélectionnons un ensemble de valeurs possibles pour τ en minimisant le critère suivant :

$$\widehat{\tau}(k) \in \arg\min_{\tau\in\mathcal{T}} \{\|\mathbf{M} - \widehat{\mathbf{\Theta}}_{k,\tau}\|_{\mathcal{F}}^2 + \mathfrak{c}_{\mathrm{cal},\tau} k(d + \tau + s)\},$$

avec la constante $\mathfrak{c}_{\mathrm{cal},\tau}$ calibrée en utilisant l'heuristique de pente proposée en 2001 par Birgé et al. dans [40] et dont l'objectif est de déterminer, à l'aide des données, une constante multiplicative optimale devant une pénalité en sélection de modèles.

(2) Puis, pour chaque valeur de $\widehat{\tau}(k)$ calculée à l'étape (1), nous sélectionnons le rang $\widehat{k}$ qui minimise le critère

$$\widehat{k} \in \arg\min_{k \in \mathcal{K}} \{\|\mathbf{M} - \widehat{\mathbf{\Theta}}_{k,\widehat{\tau}(k)}\|_{\mathcal{F}}^2 + \mathfrak{c}_{\text{cal},k} k(d + \widehat{\tau}(k) + s)\},$$

la constante $\mathfrak{c}_{\text{cal},k}$ étant une fois de plus calibrée par heuristique de pente. Notons que cette dernière peut être différente de $\mathfrak{c}_{\text{cal},\tau}$.

(3) Enfin, le $\widehat{\tau}$ sélectionné est

$$\widehat{\tau} = \widehat{\tau}(\widehat{k}).$$

Cette procédure de sélection à été implémentée dans le package R `TrendTM` que nous avons crée et qui est aujourd'hui disponnible sur le CRAN. Au delà de sélectionner conjointement ces deux paramètres, l'intérêt de nos travaux est également de confirmer la borne de risque obtenue dans [24] et de montrer que le prise en compte de la structure temporelle dans la tendance des séries (via la matrice $\mathbf{\Lambda}$) améliore l'estimation. Notre package offre donc la possibilité de prendre en compte différents types de structures temporelles sur les données (`"aucune structure"`, `"périodique"` ou `"smooth"`) et de sélectionner, selon les souhaits de l'utilisateur, soit le rang k de la matrice, soit le paramètre τ lié à la structure des tendances des séries, soit les deux. L'estimateur des moindres carrés $\widehat{\mathbf{T}}_{k,\tau}$ donné par (1.5) est obtenu en utilisant la fonction `softImpute` du package R du même nom développé par Hastie et Mazumder (cf. [50]). Cette fonction renvoie trois matrices $\mathbf{U}_f \in \mathcal{M}_{d,k}(\mathbb{R})$, $\mathbf{V}_f \in \mathcal{M}_{\tau,k}(\mathbb{R})$ et $\mathbf{D} \in \mathcal{M}_{k,k}(\mathbb{R})$ telles que

$$\widehat{\mathbf{T}}_{k,\tau} = \mathbf{U}_f \mathbf{D} \mathbf{V}_f^*.$$

Pour approximer la matrice à estimer, les auteurs ont notamment implémenté deux procédures différentes dans la fonction `softImpute` : la procédure `"svd"` (Singular Value Decomposition ou décomposition en valeurs singulières) et la procédure `"als"` (Alternate Least Square ou méthode des moindres carrés alternés). C'est cette dernière que nous avons retenue comme paramètre par défaut dans notre package `TrendTM` car, malgré son extrême simplicité, elle permet d'obtenir de bons résultats en pratique (cf. [51]). L'idée de base de la méthode ALS est de minimiser l'erreur quadratique entre la matrice d'origine et sa reconstruction partielle en utilisant des moindres carrés alternativement sur chacune des dimensions de la matrice à approximer. En d'autres termes, on cherche à résoudre un problème d'optimisation en alternant entre deux sous-problèmes, l'un portant sur les colonnes de la matrice et l'autre sur ses lignes. Des simulations ont aussi été réalisées pour comparer les deux procédures : toutes les deux fournissent la même précision d'estimation. Le choix est tout de même laissé à l'utilisateur s'il décide d'applique la fonction de notre package avec la procédure `"svd"`. Pour étudier les performances de notre procédure de sélection, nous avons alors réalisé une étude comparative sous les conditions suivantes : lorsque l'on prend en compte la structure de série temporelle dans la tendance des données ou non et en considérant différents niveaux de difficultés, pour des valeurs de σ (écart-type du bruit ε) plus

ou moins grandes. Quel que soit le niveau de difficulté considéré ou le(s) paramètre(s) que l'on cherche à estimer, la prise en compte du caractère temporel des données dans la tendance augmente la précision de l'estimation (cf. section **2.1**). Nous avons également étudié dans nos travaux la robustesse de la méthode proposée lorsque les lignes $\varepsilon_{i,.}$ de la matrice d'erreur, pour tout $i \in \{1, \ldots, d\}$, sont générées par des processus autoregressifs AR(1), induisant ainsi de la dépendance temporelle dans les lignes de la matrice ε.

Une illustration directe du package `TrendTM` a été fournie sur des données réelles de qualité de l'air (cf. section **2.1**). Par ailleurs, le dernier apport de nos travaux a été d'appliquer, de façon moins directe, la méthode proposée à des problèmes statistiques usuels : la classification (*clustering*) de courbes et l'Analyse en Composantes Principales (ACP) de séries temporelles. Ces applications n'avaient pas été developpées dans [24].

Classification de séries temporelles. Réécrire la matrice de tendance $\mathbf{\Theta}^0 = \mathbf{UL}$ avec $\mathbf{L} = \mathbf{V\Lambda} = (\ell_j(t))_{j,t}$ dans le modèle (1.2) permet de capturer les caractéristiques sous-jacentes de la matrice d'origine $\mathbf{M}$. Plus précisément, pour tout $i \in \{1, \ldots, d\}$, la tendance $m_i(.)$ de la i-ème série dans la matrice d'origine peut s'écrire comme une combinaison linéaire de k facteurs (ou séries) latent(e)s :

$$m_i(t) = \sum_{j=1}^{k} \mathbf{U}_{i,j}\ell_j(t) \ ; \forall t \in \{1, \ldots, T\}.$$

Pour classer nos différentes séries, nous avons appliqué la méthode classique de Classification Ascendante Hiérarchique (CAH), qui vise à construire une suite de partitions "emboîtées" des données en un nombre de plus en plus faible de clusters, en se basant sur la distance ou la similarité entre les données pour regrouper ces dernières dans un même cluster. Notre stratégie de classification de courbes a été appliqué sur deux jeux de données réelles :

- le premier contient les données d'electrocardiogramme de 200 patients, répartis en 2 groupes (considérés comme "normal" et "anormal") ;
- le second contient les données de croissance de 93 enfants, répartis également en 2 groupes ("filles" et "garçons").

Une première étude, réalisée en considérant le nombre de clusters fixé à 2, a nouvelle fois montré que la prise en compte de la structure temporelle dans la tendance des séries étudiées (structure "périodique" pour le jeu de données ECG et "*smooth*" pour le jeu de données de croissance d'enfants) améliorait la classification. Nos performances en matière de classification sont les mêmes que celles de la meilleure méthode de classification décrite dans [43] et elles sont également beaucoup plus rapides. Dans une seconde étude, nous avons testé notre package sur la sélection du nombre de séries latentes k et de la période τ.

Analyse en Composantes Principales de séries temporelles (ACP). Cette technique permet de réduire la dimensionnalité de notre ensemble de données en projetant chaque série de la matrice d'origine $\mathbf{M}$ sur un nouvel espace de variables, appelé "espace des composantes principales". Dans notre étude, si $\mathbf{M}$ est une matrice de taille $d \times T$ et de rang k, alors l'information principale sera capturée par les k premiers axes de l'ACP. Plus précisemment, la solution du problème d'ACP est

$$\widehat{\mathbf{\Theta}}_{k,T} \in \arg \min_{\mathbf{U} \in \mathcal{M}_{d,k}, \mathbf{V} \in \mathcal{M}_{k,T}} \|\mathbf{M}^c - \mathbf{UV}\|_{\mathcal{F}}^2$$

avec, pour tout $j \in \{1, \ldots, T\}$, $\mathbf{M}^c_{\cdot,j} = \mathbf{M}_{\cdot,j} - \frac{1}{d}\sum_{i=1}^{d} \mathbf{M}_{i,j}$. L'objectif de nos travaux est de comparer les projections des séries temporelles obtenues lorsque l'ACP est appliquée à :

- $\mathbf{M}$ seulement, c'est-à-dire sans prise en compte de la structure temporelle dans la tendance. Dans ce cas, la matrice des composantes principales est $\widehat{\mathbf{U}}_k = \mathbf{M}^c \widehat{\mathbf{V}}_k^*$. Cette matrice contient, pour tout $i \in \{1, \ldots, d\}$, les coordonnées de la projection de la i-ème série de $\mathbf{M}$ sur les k premiers axes.

- puis à $\underline{\mathbf{M}}^c = \mathbf{M}^c \mathbf{\Lambda}^+$, donc en prenant en compte cette fois la possible structre temporelle des données dans la tendance des séries. Dans ce cas, la matrice des composantes principales est $\underline{\mathbf{M}}^c \widehat{\mathbf{V}}_k^* = \mathbf{M}^c \mathbf{\Lambda} + \widehat{\mathbf{V}}_k^*$.

Nos travaux ont, une fois de plus, révelé une meilleure performance de l'ACP dans le cas où l'on prend en compte la tendance via la matrice $\mathbf{\Lambda}^+$.

1.1.2 Complétion de matrice de séries temporelles de grande dimension

Le modèle de débruitage introduit dans la section précédente nous permet de construire le modèle de complétion sur lequel nous avons travaillé dans le deuxième article présenté dans cette thèse (cf. section **3.2**). Rappelons que la complétion de matrice est une technique qui vise à prédire les valeurs manquantes d'une matrice en utilisant les valeurs connues de cette matrice. Il existe de nombreux modèles de complétion de matrice différents, allant des modèles de régression linéaire simples [44] aux modèles de réseaux de neurones complexes [45, 46]. L'objectif de nos travaux, cette fois-ci, n'est plus d'estimer la matrice $\mathbf{\Theta}^0$ dans le modèle $\mathbf{M} = \mathbf{\Theta}^0 + \varepsilon$ lorsque toutes les entrées de $\mathbf{M}$ sont observées, mais lorsqu'il en manque ou qu'elles sont perturbées par des erreurs d'observation liées à l'instrument de mesure. On dispose donc, pour la matrice $\mathbf{M}$, de $n \in \{1, \ldots, d \times T\}$ entrées bruitées, notées $Y_1, \ldots, Y_n$ et satisfaisant le modèle suivant :

$$Y_i = \text{trace}(\mathbf{X}_i^* \mathbf{M}) + \xi_i \ ; \ i \in \{1, \ldots, n\} \tag{1.8}$$

où $\mathbf{X}_1, \ldots, \mathbf{X}_n$ sont n matrices aléatoires à valeurs dans

$$\mathcal{X} := \{e_{\mathbb{R}^d}(j) e_{\mathbb{R}^T}(t)^* \ ; \ 1 \leqslant j \leqslant d \text{ and } 1 \leqslant t \leqslant T\}$$

et $\xi_1, \ldots, \xi_n$ sont des variables aléatoires centrées, indenpendantes et identiquement distribuées, d'écart-type $\sigma_\xi > 0$ et tels que $\mathbf{X}_i$ et ξ_i sont indépendants pour tout $i \in \{1, \ldots, n\}$. Afin de prendre en compte la structure de série temporelle dans les lignes de $\mathbf{M}$, nous reprenons le modèle de débruitage introduit précedemment, à savoir

$$\left\{ \begin{array}{rcl} \mathbf{M} & = & \mathbf{\Theta}^0 + \varepsilon \\ \mathbf{\Theta}^0 & = & \mathbf{T}^0 \mathbf{\Lambda} = \mathbf{U}^0 \mathbf{V}^0 \mathbf{\Lambda} \end{array} \right. \tag{1.9}$$

avec $\mathbf{\Theta}^0$ une matrice déterministe, ε une matrice aléatoire, toutes les deux de taille $d \times T$, et $\mathbf{U}^0 \in \mathcal{M}_{d,k}(\mathbb{R})$ et $\mathbf{V}^0 \in \mathcal{M}_{k,T}(\mathbb{R})$. Comme dans le modèle de débruitage, la matrice $\mathbf{\Theta}^0$ est supposée de faible rang $k \in \mathbb{N}^*$ ($k \ll d \wedge T$), mettant en avant la forte corrélation qu'il peut y avoir entre les différentes séries temporelles, et la matrice d'erreur ε a des lignes centrées, indépendantes et identiquement distrubées. D'après les équations (1.8) et (1.9), le modèle de complétion final s'écrit

$$Y_i = \mathrm{trace}(\mathbf{X}_i^* \mathbf{\Theta}^0) + \overline{\xi}_i \tag{1.10}$$

avec $\overline{\xi}_i := \mathrm{trace}(\mathbf{X}_i^* \varepsilon) + \xi_i$. Un estimateur naturel de $\mathbf{\Theta}^0$ est alors donné par

$$\left\{ \begin{array}{rcl} \widehat{\mathbf{\Theta}}_{k,\tau} & = & \widehat{\mathbf{T}}_{k,\tau} \mathbf{\Lambda} \\ \widehat{\mathbf{T}}_{k,\tau} & \in & \arg \min\limits_{\mathbf{T} \in \mathcal{S}_{k,\tau}} r_n(\mathbf{T}\mathbf{\Lambda}) \end{array} \right. , \tag{1.11}$$

où

$$r_n(\mathbf{A}) := \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle \mathbf{X}_i, \mathbf{A} \rangle_{\mathcal{F}})^2 \; ; \forall \mathbf{A} \in \mathcal{M}_{d,T}(\mathbb{R})$$

et

$$\mathcal{S}_{k,\tau} \subset \left\{ \mathbf{A} \in \mathcal{M}_{d,\tau}(\mathbb{R}) : \mathrm{rg}(\mathbf{A}) = k \text{ et } \sup_{j,t} |\mathbf{A}_{j,t}| \leqslant \frac{\mathfrak{m}_0}{\mathfrak{m}_{\mathbf{\Lambda}}(\tau)}, \mathfrak{m}_{\mathbf{\Lambda}}(\tau) := \sup_{j,t} |\mathbf{\Lambda}_{j,t}| < \infty \right\}.$$

Les deux mêmes structures de tendances portées par la matrice $\mathbf{\Lambda}$ sont considérées, à savoir le cas où les tendances sont τ-périodiques ou le cas où les tendances sont décomposables dans la base de Fourier.

Le principal apport de nos travaux a été d'améliorer la borne de risque établie dans Koltchinskii et al. en 2011 [17], pour laquelle les termes d'erreur sont supposés i.i.d. et la matrice $\mathbf{\Theta}^0$ est supposée de faible rang k seulement. Ils ont obtenu une borne de risque pour l'estimateur de la matrice de l'ordre de $k(d+T)\log(n)/n$ pour des matrices de taille $d \times T$ et n entrées bruitées. Dans notre contexte, la structure de séries temporelles dans les lignes de $\mathbf{M}$ nous permet d'améliorer la borne de risque de l'estimateur de la matrice, moyennant certaines conditions. Pour obtenir une borne de risque optimale, certaines hypothèses sur la structure du bruit ε et sur les erreurs d'observation ξ_i sont nécessaires :

(1) **Les lignes de ε sont i.i.d et issues d'un processus Φ-mélengeant**. Pour de tels processus, la corrélation entre deux valeurs éloignées dans le temps décroît exponentiellement avec l'écart temporel entre ces deux valeurs. Cette hypothèse est, de plus, nécessaire afin d'utiliser l'inégalité de concentration obtenue par Samson dans [49], qui est en réalité une extension de l'inégalité de Bernstein à des processus Φ-méleangeant, et obtenir la borne de risque optimale. En effet, les auteurs de [23]

ont montré que les bornes de risque obtenues avec d'autres inégalités de concentration classiquement utilisées avec les séries temporelles étaient moins bonnes. [47] fourni notamment une étude détaillée des autres conditions de mélanges que l'on peut retrouver dans la littérature.

(2) **Les coefficients de la matrice ε sont bornés et $\xi_1, \ldots, \xi_n$ sont des variables aléatoires i.i.d. sous-exponentielles.** Cela couvre notamment les variables aléatoires Gaussiennes. Cette hypothèse de bornitude du bruit a notamment été reprise dans plusieurs travaux dans le contexte de complétion de matrices ayant des entrées i.i.d. (cf. [19, 17]) et permet, une nouvelle fois, d'obtenir une borne de risque optimale.

Soit $\alpha \in (0,1)$. Sous les hypothèses précédentes, nous obtenons, avec une probabilité supérieure à $1-\alpha$, la borne de risque suivante

$$\|\widehat{\mathbf{\Theta}}_{k,\tau} - \mathbf{\Theta}^0\|^2_{\mathcal{F},\Pi} \leqslant 3 \min_{\mathbf{T}\in\mathcal{S}_{k,\tau}} \|(\mathbf{T}-\mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} + \mathfrak{c}_1 \left[k(d+\tau)\frac{\log(n)}{n} + \frac{1}{n}\log\left(\frac{4}{\alpha}\right)\right] \quad (1.12)$$

où Π désigne une mesure de probabilités sur $\mathcal{X}$ telle que

$$\mathbb{P}_{\mathbf{X}_i} = \Pi \ ; \ \forall i \in \{1, \ldots, n\},$$

avec

$$\langle \mathbf{A}, \mathbf{B} \rangle_{\mathcal{F},\Pi} := \int_{\mathcal{M}_{d,T}(\mathbb{R})} \langle X, \mathbf{A} \rangle_{\mathcal{F}} \langle X, \mathbf{B} \rangle_{\mathcal{F}} \Pi(dX) \ ; \ \forall \mathbf{A}, \mathbf{B} \in \mathcal{M}_{d,T}(\mathbb{R})$$

et où $\mathfrak{c}_1$ est une constante déterministe, indépendante de n, d, k, τ et α, que nous pouvons expliciter. Le terme de variance dans la borne de risque est donc d'ordre $k(d+\tau)\log(n)/n$. Lorsque les séries étudiées dans la matrice d'origine ne présentent aucune structure temporelle ($\tau = T$ et $\mathbf{\Lambda} = \mathbf{I}_T$) et que la matrice d'erreur ε est nulle, on retrouve la borne de risque obtenue dans le cas i.i.d. (cf. [17]) tandis qu'en exploitant la structure de séries temporelles dans les lignes de $\mathbf{M}$, puisque $\tau \ll T$, on obtient une borne de risque plus fine sur l'estimateur $\widehat{\mathbf{\Theta}}_{k,\tau}$.

Un autre objectif de nos travaux était d'établir une procédure de sélection du rang k de la matrice $\mathbf{\Theta}^0$, paramètre inconnu en pratique. L'estimateur adaptatif de la matrice $\mathbf{\Theta}^0$, dans ce cas, est le suivant : $\widehat{\mathbf{\Theta}}_{\widehat{k}} = \widehat{\mathbf{T}}_{\widehat{k}}\mathbf{\Lambda}$ avec

$$\widehat{k} \in \arg\min_{k\in\mathcal{K}} \{r_n(\widehat{\mathbf{T}}_k\mathbf{\Lambda}) + \mathrm{pen}(k)\} \text{ avec } \mathcal{K} = \{1, \ldots, k^*\} \subset \mathbb{N}^*,$$

et

$$\mathrm{pen}(k) := 16\mathfrak{c}_{\mathrm{pen}}\frac{\log(n)}{n}k(d+\tau) \text{ with } \mathfrak{c}_{\mathrm{pen}} = 2\left(\frac{1}{\mathfrak{c}_1} \wedge \lambda^*\right)^{-1}.$$

L'expression exacte de la constante c_{pen} a été explicitée dans les preuves mais dépend néanmoins de constantes que nous ne connaissons pas en pratique. Nous avons utilisé, une fois de plus, l'heuristique de pente pour calibrer la constante de pénalité. Sur le plan théorique, nous avons établie une borne de risque pour l'estimateur adaptatif $\widehat{\mathbf{\Theta}}_{\widehat{k}}$.

Afin d'illustrer le gain théorique obtenu dans la borne de risque présentée en (1.12), des simulations ont été réalisées sur des données périodiques simulées : nous avons comparé la reconstruction de la matrice obtenue dans le cas où l'on exploite la structure de périodicité dans les tendances des séries avec la procédure standard, c'est-à-dire sans prise en compte de la périodicité. Comme pour le débruitage de séries temporelles, nous observons, avec nos expériences numériques, une nette amélioration de l'estimation lorsque l'on exploite la structure temporelle dans la tendance. Nous illustrons également nos résultats sur des données réelles *Vélib* provenant du système de vélos en libre-service de Paris.

1.2 Estimation dans des modèles d'équations différentielles stochastiques

Les équations différentielles stochastiques permettent depuis de nombreuses années de modéliser un large éventail de systèmes dynamiques aléatoires à temps continu. En effet, contrairement aux équations différentielles ordinaires (EDO), elles ont l'avantage de tenir compte de la présence de fluctuations aléatoires. Ces fluctuations peuvent provenir de diverses sources, telles que des erreurs de mesure ou des facteurs environnementaux externes, et peuvent avoir un impact significatif sur le comportement du système étudié. En incorporant l'aléa dans le modèle, les EDS fournissent une représentation plus réaliste de la dynamique sous-jacente et permettent de faire des prédictions plus précises sur le comportement du système étudié. Elles trouvent, elles aussi, des applications dans de nombreux domaines scientifiques. En physique, par exemple, les EDS sont utilisées pour modéliser le mouvement des particules dans les fluides ou la diffusion d'un gaz (cf. [2]). La notion d'EDS est également centrale en finance, en particulier en ce qui concerne la modélisation des produits dérivés et le comportement des marchés financiers (cf. [1, 3]). Elles ont également été utilisées en biologie, pour modéliser la dynamique des réactions biochimiques dans les cellules, telles que l'expression des gènes et la transduction des signaux (cf. [4]). Considérons $T > 0$, $b, \sigma : [0,T] \times \mathbb{R} \to \mathbb{R}$, $(X_t)_{t\in[0,T]}$ un processus qui dépend du temps et $(W_t)_{t\in[0,T]}$ un mouvement Brownien. L'équation différentielle ordinaire

$$\begin{cases} dX_t = b(t, X_t)dt \\ X_0 \in \mathbb{R} \end{cases}$$

ne modélise que la tendance du phénomène observé. Pour tenir compte de l'aspect aléatoire du phénomène en question, il convient de perturber l'EDO précédente de la façon suivante :

$$\begin{cases} dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t \\ X_0 \in \mathbb{R} \end{cases},$$

que nous pouvons également écrire plus rigoureusement sous la forme

$$\underbrace{X_t = x_0 + \int_0^t b(s, X_s)ds}_{\text{EDO}} + \underbrace{\int_0^t \sigma(s, X_s)dW_s}_{\text{Intégrale d'Itô}}. \tag{1.13}$$

Ce terme aléatoire, qui modélise le bruit dans l'EDS, est dénifi via l'intégrale d'Itô par rapport au mouvement Brownien. La fonction b est appelée fonction de drift et la fonction σ est appelée fonction de diffusion. Pour modéliser les fluctuations aléatoires, notons qu'il existe des processus plus généraux que le mouvement Brownien, tel que les processus de Lévy par exemple, mais nous nous limiterons dans cette thèse à l'étude des équations différentielles stochastiques dirigées par un mouvement Brownien. L'estimation des paramètres dans les EDS est une question statistique légitime et, dans le troisième travail présenté dans cette thèse, nous nous concentrons en particulier sur l'estimation non-paramétrique de la fonction de drift b de l'EDS définie par (1.13). Avant d'introduire nos travaux, faisons un point sur les différents estimateurs qui ont été étudiés dans la littérature. Rappelons tout d'abord que, pour garantir l'existence et l'unicité de la solution de (1.13), b et σ doivent être dans $C^1([0,T]\times\mathbb{R},\mathbb{R})$ et satisfaire les conditions suivantes :

(1) **Condition de croissance linéaire**. Pour tous $t \in [0,T]$ et $x \in \mathbb{R}$, il existe une constante $C > 0$ telle que

$$|b(t,x)| + |\sigma(t,x)| \leqslant C(1 + |x|).$$

Cette condition garantit qu'une solution de l'EDS existe et n'explose pas dans l'intervalle $[0, T]$.

(2) **Condition de Lipschitz**. Pour tous $t \in [0, T]$ et $x, y \in \mathbb{R}$, il existe une constante $K > 0$ telle que

$$|b(t, x) - b(t, y)| + |\sigma(t, x) - \sigma(t, y)| \leqslant K|x - y|.$$

Cette condition garantit l'unicité de la solution de l'EDS.

Notons qu'une condition plus forte que la condition (1) serait d'avoir b et σ de dérivées partielles bornées par la constante K, la condition de Lipschitz serait alors forcément assurée. Traditionnellement (cf. [122, 104, 118]), les estimateurs de la fonction de drift b sont calculés à partir d'une unique solution stationnaire ergodique de l'EDS observée sur un intervalle de temps long $[0, T]$, avec $T \to \infty$. La stationnarité de la solution est assurée par une condition de dissipativité assez restrictive sur la fonction de drift. Pour contourner ce problème et s'affranchir de cette condition sur b, de nouveaux estimateurs ont vu le jour. L'idée, cette fois, est de construire un estimateur de la fonction de drift basé, non plus sur une seule, mais sur N copies indépendantes $X^1, \ldots, X^N$ de la solution et observées sur un intervalle de temps court $[0, T]$ ($T > 0$ fixé) en considérant le plus de copies possibles, soit $N \to \infty$. Ce type d'approche par copies pour l'estimateur de b, à temps continu et à temps discret, a déjà été utilisée de nombreuses fois dans le cadre paramétrique (cf. [113, 114, 110]) et plus récemment dans le cadre non-paramétrique (cf. [102, 111, 103, 112]). Cette approche est particulièrement intéressante dans certains contextes, par exemple, lorsque l'on cherche à modéliser des actifs financiers sujets aux fluctuations aléatoires des prix, dans le but de prédire le comportement futur du système et de prendre des décisions informées en fonction des différents scénarios possibles. On retrouve notamment deux types d'estimateurs lorsqu'il est question d'estimation non-paramétrique :

L'estimateur des moindres carrés en projection. En regression, la méthode par projection généralise la méthode d'estimation des moindres carrés dans le modèle linéaire multidimensionnel. De façon générale, l'estimation non-paramétrique consiste à estimer une fonction sans indication sur sa forme, mais seulement sur sa régularité. L'objectif de la méthode des moindres carrés en projection est d'estimer la fonction de régression, que l'on notera m, par un élement de $S_m = \text{vect}(\varphi_1, \ldots, \varphi_\text{m})$, où $(\varphi_1, \ldots, \varphi_m)$ désigne une famille orthonormée de $\mathbb{L}^2(\mathbb{R})$, minimisant une fonction de contraste de type moindres carrés. Nous ne détaillerons pas plus ce type d'estimateur ici car ce n'est pas l'objet de notre étude (cf. [52], chapitre 4). L'estimateur des moindres carrés en projection a également été utilisé dans le cadres des EDS, lorsque l'on cherche à estimer la fonction de drift en considérant plusieurs copies du processus de diffusion. Comte et Genon-Catalot [52] (respectivement Comte et Marie [53]) exploitent notamment cette méthode en considérant N copies indépendantes (respectivement dépendantes) du processus de diffusion.

L'estimateur de Nadaraya-Watson. Il s'agit, cette fois, d'une méthode à noyau permettant d'estimer la fonction de régression m dans le modèle suivant

$$Y_i = m(X_i) + \varepsilon_i, \quad \forall i \in \{1, \ldots, n\}$$

où $X_1, \ldots, X_n$ sont des variables aléatoires i.i.d. de densité f (on parle alors de *random design*) et $Y_1, \ldots, Y_n$ sont les variables que l'on cherche à expliquer. L'estimateur de Nadaraya-Watson est un estimateur "quotient" : après avoir introduit la fonction $l = mf$,

le principe est d'estimer la fonction de regression $m = \frac{l}{f}$ en proposant, d'une part un estimateur du numérateur l, d'autre part un estimateur du dénominateur f, en prenant garde à ce que ce dernier ne devienne pas trop petit. L'estimateur de Nadaraya-Watson, introduit indépendamment par Nadaraya [54] et Watson [55], est défini comme suit :

$$\hat{m}_h(x) = \frac{\widehat{l}_h(x)}{\widehat{f}_h(x)} = \frac{\sum_{i=1}^n K_h(x - x_i) y_i}{\sum_{i=1}^n K_h(x - x_i)}$$

où le noyau $K : \mathbb{R} \to \mathbb{R}$ est une fonction intégrable telle que $\int_{\mathbb{R}} K(u)du = 1$ et $K_h(.) = (1/h)K(./h)$ et $h > 0$ désigne le paramètre de lissage qui contrôle la largeur de la fenêtre du noyau. En général, le noyau K est choisi de telle sorte à être une fonction symétrique, nous pouvons citer plusieurs options courantes telles que le noyau gaussien, le noyau triangulaire ou encore le noyau d'Epanechnikov (cf. [52], chapitres 3 et 4). La question du choix de la fenêtre h est également une étape cruciale pour obtenir un estimateur performant, nous reviendrons sur cette problématique dans la suite de l'introduction. Dans les modèles d'équations différentielles stochastiques, l'estimateur de Nadaraya-Watson pour la fonction de drift b calculé à partir de copies de la solution n'a été que très peu étudié. Ce type d'estimateur a notamment été utilisé par Della Maestra et Hoffman [111] pour estimer la fonction de drift dans un modèle de particule en interaction. Dans la section **3** de cette thèse, nous présentons l'estimateur de Nadaraya-Watson sur lequel nous nous sommes concentrés pour estimer la fonction de drift b dans notre modèle, en considérant N copies **indépendantes** du processus de diffusion. En effet, nous ne prenons pas en compte les interactions entre les différentes copies dans nos travaux, contrairement à [111]. Notre estimateur est présenté ci-dessous.

Reprenons le modèle (1.13) dans le cas autonome (*i.e.* b et σ ne dépendent pas du temps). Considérons $\mathcal{I} : (x, w) \to \mathcal{I}(x, w)$ l'application d'Itô pour l'équation (1.13), $W^1, \ldots, W^N$, $N \in \mathbb{N}^*$ copies indépendantes de $W = (W_t)_{t\in[0,T]}$ et, pour tout $i \in \{1, \ldots, N\}$, $X^i = \mathcal{I}(x_0, W^i)$. Sous les bonnes conditions sur les fonctions b et σ (cf. section **3**), la distribution de X_t admet une densité de probabilité $p_t(x_0, .)$ pour tout $t \in]t_0, T]$ et cela nous permet d'introduire, pour $t_0 > 0$,

$$f(x) := \frac{1}{T - t_0} \int_{t_0}^T p_t(x_0, x)dt \ ; \ x \in \mathbb{R}.$$

La fonction f est une densité de probabilité, puisqu'en effet,

$$\int_{-\infty}^{\infty} f(x)dx = \frac{1}{T - t_0} \int_{t_0}^T \int_{-\infty}^{\infty} p_t(x_0, x)dxdt = 1.$$

Considérons le noyau $K : \mathbb{R} \to \mathbb{R}$ vérifiant les propriétés suivantes :

(1) K est symétrique, continu et tel que $K \in \mathbb{L}^2(\mathbb{R}, dx)$,

(2) Il existe $\upsilon \in \mathbb{N}^*$ tel que

$$\int_{-\infty}^{\infty} |z^{\upsilon+1} K(z)|dz < \infty \quad \text{et} \quad \int_{-\infty}^{\infty} z^\ell K(z)dz = 0 \ ; \ \forall \ell \in \{1, \ldots, \upsilon\}.$$

L'estimateur de Nadaraya-Watson à temps continu de la fonction de drift b est le suivant :

$$\widehat{b}_{N,h}(x) := \frac{\widehat{bf}_{N,h}(x)}{\widehat{f}_{N,h}(x)} \tag{1.14}$$

avec

$$\widehat{f}_{N,h}(x) := \frac{1}{N(T-t_0)} \sum_{i=1}^{N} \int_{t_0}^{T} K_h(X_t^i - x)dt$$

et

$$\widehat{bf}_{N,h}(x) := \frac{1}{N(T-t_0)} \sum_{i=1}^{N} \int_{t_0}^{T} K_h(X_t^i - x)dX_t^i.$$

Nous proposons également, dans nos travaux, une version discrétisée de l'estimateur introduit en (1.14). Soient $(t_0, t_1, \ldots, t_n)$ la subdivison de l'intervalle $[t_0, T]$ telle que $t_j = t_0 + (T-t_0)j/n$ pour tout $j \in \{1, \ldots, n\}$. Alors, l'estimateur de Nadaraya-Watson de b à temps discret est défini comme suit :

$$\widehat{b}_{n,N,h}(x) := \frac{\widehat{bf}_{n,N,h}(x)}{\widehat{f}_{n,N,h}(x)},$$

où

$$\widehat{f}_{n,N,h}(x) := \frac{1}{nN} \sum_{i=1}^{N} \sum_{j=0}^{n-1} K_h(X_{t_j}^i - x)$$

et

$$\widehat{bf}_{n,N,h}(x) := \frac{1}{N(T-t_0)} \sum_{i=1}^{N} \sum_{j=0}^{n-1} K_h(X_{t_j}^i - x)(X_{t_{j+1}}^i - X_{t_j}^i).$$

Un des premiers objectifs de nos travaux est d'obtenir, à la fois pour la version continue et discrète de l'estimateur de b, des bornes de risque sur les estimateurs du numérateur et du dénominateur, et donc sur l'estimateur de Nadaraya-Watson lui-même. Ces bornes de risque sont obtenues sous des conditions particulières sur l'existence et la régularité de la densité $p_t(x_0, .)$ et donc, sur la densité f. En effet, les principales hypothèses, qui seront présentées plus en détails dans la section **3**, sous lesquelles nous avons travaillé sont les suivantes :

La densité $p_t(x_0, .)$ et sa dérivée doivent vérifier une borne de type Kusuoka-Stroock. Plus précisemment, la densité de la distribution de X_t doit être $\beta \in \mathbb{N}^*$ fois dérivable sur $\mathbb{R}$. Les bornes de Kusuoka-Stroock (cf. [121]) permettent alors d'obtenir un contrôle sous-gaussien de la densité $p_t(x_0, .)$ ainsi que de ses dérivées partielles $\partial_x^\ell p_t(x_0, x)$ et $\partial_t p_t(x_0, x)$.

Ainsi, la densité f doit vérifier une condition de type Nikol'skii. Plus précisemment, f doit être $\ell \in \{0, \ldots, \beta - 1\}$ fois dérivable sur $\mathbb{R}$. La condition de Nikol'skii nous permet d'obtenir un contrôle de la quantité $\int_{-\infty}^{\infty} [f^{(\ell)}(x+\theta) - f^{(\ell)}(x)]^2 dx$.

Soit l'estimateur de Nadaraya-Watson à temps continu

$$\widehat{b}_{N,h,h'}(x) := \frac{\widehat{bf}_{N,h}(x)}{\widehat{f}_{N,h'}(x)} \mathbf{1}_{\widehat{f}_{N,h'}(x) > m/2}$$

avec $f(x) > m > 0$ pour tout $x \in [A, B]$ ($m \in (0, 1]$ et $A, B \in \mathbb{R}$ tels que $A < B$) et $h, h' > 0$ deux fenêtres distinctes. Sous les bonnes hypothèses, nous obtenons la borne de risque suivante pour cet estimateur :

$$\mathbb{E}(\|\widehat{b}_{N,h,h'} - b\|_{f,A,B}^2) \quad \leqslant \quad \frac{\mathfrak{c}_1}{m^2} \left[\|(bf)_h - bf\|_2^2 + \frac{\mathfrak{c}_2(t_0)}{Nh} + 2\|b\|_f^2 \left(\mathfrak{c}_3(t_0) h'^{2\beta} + \frac{\|K\|_2^2}{Nh'} \right) \right]$$

avec $\|\varphi\|_{f,A,B} := \|\varphi \mathbf{1}_{[A,B]}\|_f$ pour tout $\varphi \in \mathbb{L}^2(\mathbb{R}, f(x)dx)$. Plus précisemment, le risque de l'estimateur $\widehat{b}_{N,h,h'}$ est contrôlé par la somme des risques du numérateur $\widehat{bf}_{N,h}$ et du dénominateur $\widehat{f}_{N,h'}$, à une constante multiplicative près. Sous certaines conditions de régularité, nous montrons que le terme de biais $\|(bf)_h - bf\|_2^2$ dans la borne de risque est de l'ordre de $h^{2\gamma} + (h')^{2\beta}$, $\gamma, \beta \in \mathbb{N}^*$, tandis que le terme de variance est de l'ordre de $1/(Nh) + 1/(Nh')$. Si les fenêtres h et h' sont trop petites, l'estimateur de b aura une faible variance car il sera plus sensible aux fluctuations locales des données. En contrepartie, cela peut conduire à un biais beaucoup trop élevé. Le choix des fenêtres h et h' est alors crucial pour trouver le point où les deux sources d'erreur sont équilibrées : c'est ce que l'on appelle le compromis biais-variance. Notons que lorsque celui-ci est atteint (sous certaines conditions), nous obtenons une vitesse de convergence du même ordre que celle obtenue par Della Maestra et Hoffman dans [111]. Nous obtenons également, dans nos travaux, une borne de risque sur l'estimateur de Nadaraya-Watson à temps discret. Enfin, le paragraphe suivant aborde la problématique de sélection de fenêtres à laquelle nous nous sommes intéressés.

Nous proposons effectivement dans nos travaux deux méthodes de sélection de fenêtres pour sélectionner les paramètres h et h'. La première est une extension de la méthode de cross validation leave-one-out, que nous avons illustrée via des expériences numériques. La seconde est une méthode type PCO (Penalized Comparison to Overfitting) pour laquelle nous avons démontré une inégalité d'oracle pour l'estimateur adaptatif obtenu en sélectionnant la fenêtre du numérateur et la fenêtre du dénominateur séparément.

Méthode de cross-validation leave-one-out (looCV). Bien qu'on manque de résultats théoriques avec cette méthode, elle reste néanmoins numériquement intéressante. On considère, dans cette section, l'estimateur discrétisé

$$\widehat{b}_{n,N,h}(x) = \sum_{i=1}^{N} \sum_{j=0}^{n-1} \omega_j^i(x)(X^i_{t^i_{j+1}} - X_{t^i_j})$$

avec

$$\omega_j^i(x) := \frac{K_h(X^i_{t_j} - x)}{\sum_{k=1}^{N} \sum_{\ell=0}^{n-1} K_h(X^k_{t_\ell} - x)(t_{\ell+1} - t_\ell)},$$

et tel que

$$\sum_{i=1}^{N} \sum_{j=0}^{n-1} \omega_j^i(x)(t_{j+1} - t_j) = 1.$$

L'estimateur de b peut ainsi être vu comme une combinaison convexe des $(X^i_{t^i_{j+1}} - X_{t^i_j})$, pondérés par les poids ω_j^i. La méthode looCV consiste à sélectionner la fenêtre h qui minimise le critère suivant :

$$\mathrm{CV}(h) := \sum_{i=1}^{N} \left[\sum_{j=0}^{n-1} \widehat{b}^{-i}_{n,N,h}(X^i_{t_j})^2 (t_{j+1} - t_j) - 2 \sum_{j=0}^{n-1} \widehat{b}^{-i}_{n,N,h}(X^i_{t_j})(X^i_{t_{j+1}} - X^i_{t_j}) \right]$$

avec

$$\widehat{b}^{-i}_{n,N,h}(x) := \sum_{k \in \{1,\dots,N\} \setminus \{i\}} \sum_{j=0}^{n-1} \omega_j^k(x)(X^k_{t_{j+1}} - X^k_{t_j}).$$

Des expériences numériques, dans lesquelles nous avons implémenté la méthode de sélection de fenêtre par looCV, ont été réalisées pour estimer des fonctions de drift dans

différents modèles d'EDS (cf. section **3**).

Extension de la méthode PCO. Contrairement à la looCV, nous avons réussi à établir des ingélités d'oracle pour l'estimateur adaptatif de la fonction de drift. Elle est numériquement moins intéressante que la looCV mais reste néanmoins plus facile à implémenter que la méthode de Goldenschluger et Lepski (cf. [56]), autre méthode de sélection de fenêtre offrant également de bonnes garanties théoriques sur l'estimateur adaptatif. Le principe de la méthode PCO est le suivant : on définit de la même façon un estimateur quotient $\widehat{b}_{N,\widehat{h},\widehat{h}'}(x) = \frac{\widehat{bf}_{N,\widehat{h}}(x)}{\widehat{f}_{N,\widehat{h}'}(x)} \mathbf{1}_{\widehat{f}_{N,\widehat{h}'}(x) > m/2}$ où h (respectivement h') est une fenêtre sélectionnée en minimisant un critère de la forme $\|\widehat{bf}_{N,h} - \widehat{bf}_{N,h_0}\|_{2,\delta}^2 + \mathrm{pen}(h)$ (respectivement $\|\widehat{f}_{N,h} - \widehat{f}_{N,h'_0}\|_2^2 + \mathrm{pen}'(h)$), h_0, h'_0 représentant les fenêtres minimales des collections de fenêtres considérées pour notre étude, d'où le terme de pénalisation par rapport à l'overfitting. Sous les bonnes conditions, nous obtenons la borne de risque suivante pour l'estimateur adaptatif $\widehat{b}_{N,\widehat{h},\widehat{h}'}$:

Si $f(x), \delta(x) > m > 0$ pour tout $x \in [A, B]$ ($m \in (0, 1]$ et $A, B \in \mathbb{R}$ tels que $A < B$), où δ désigne un noyau connu, alors il existe deux constantes déterministes $\mathfrak{c}_1, \mathfrak{c}_2 > 0$, indépendante de N telles que, pour tout $\vartheta \in (0, 1)$,

$$\underbrace{\mathbb{E}(\|\widehat{b}_{N,\widehat{h},\widehat{h}'} - b\|_{f,A,B}^2)}_{\text{risque intégré}} \leqslant \frac{2\mathfrak{c}_1(1 \vee \|\delta\|_\infty)}{m^3} \times \left[(1+\vartheta) \min_{(h,h') \in \mathcal{H}_N \times \mathcal{H}'_N} \{ \underbrace{\mathbb{E}(\|\widehat{bf}_{N,h} - bf\|_2^2) + \mathbb{E}(\|\widehat{f}_{N,h'} - f\|_2^2)}_{\text{risques à } h, h' \text{ fixés}} \} + \underbrace{\frac{\mathfrak{c}_2}{\vartheta} \left(\|(bf)_{h_0} - bf\|_2^2 + \|f_{h'_0} - f\|_2^2 + \frac{1}{N} \right)}_{\text{terme négligeable}} \right].$$

Cette borne de risque signifie que la performance de l'estimateur de b est de l'ordre celle de la somme des deux meilleurs estimateurs $\widehat{bf}_{N,h}$ et $\widehat{f}_{N,h'}$ de la collection

$$\min_{(h,h') \in \mathcal{H}_N \times \mathcal{H}'_N} \{ \mathbb{E}(\|\widehat{bf}_{N,h} - bf\|_2^2) + \mathbb{E}(\|\widehat{f}_{N,h'} - f\|_2^2) \},$$

à un facteur $(1 + \vartheta)$ près et un terme additif négligeable. Les démonstrations seront présentées dans la section **3**.

Dans les sections suivantes, nous allons examiner de manière approfondie les différentes problématiques présentées dans l'introduction. Nous commencerons chaque section en rappelant le modèle étudié. Ce rappel permettra aux lecteurs de se familiariser avec les notations utilisées et d'avoir une compréhension claire de la manière dont chaque problématique est abordée. Nous fournirons également, dans chacune des sections, le détail des démonstrations des théorèmes établis ainsi que les résultats des expériences numériques réalisées.

Chapitre 2

High-dimensional time series estimation

2.1 A R Package for the Trend of High Dimensional Time Series Estimation : `TrendTM`

Let denote by $\mathbf{M}$ the observed $d \times n$ matrix which rows are time series with d and n high, matrix factorization consists in approximating $\mathbf{M}$ by a matrix $\mathbf{\Theta}^0$ of low rank $k \in \mathbb{N}^*$ (i.e. $k \ll d \wedge n$), which can therefore be written as the product $\mathbf{UV}$ of two matrix $\mathbf{U} \in \mathcal{M}_{d,k}(\mathbb{R})$ and $\mathbf{V} \in \mathcal{M}_{k,n}(\mathbb{R})$. Formally, let us consider the model

$$\mathbf{M} = \mathbf{\Theta}^0 + \varepsilon \tag{2.1}$$

The matrix $\mathbf{\Theta}^0$ is usually estimated by using a contrast minimization approach, the most popular being the least squares contrast associated to the Frobenius norm : the best rank-k approximation of $\mathbf{\Theta}^0$ is

$$\widehat{\mathbf{\Theta}} \in \arg \min_{\mathbf{U} \in \mathcal{M}_{d,k}, \mathbf{V} \in \mathcal{M}_{k,n}(\mathbb{R})} \|\mathbf{M} - \mathbf{UV}\|_{\mathcal{F}}^2, \tag{2.2}$$

where $\|.\|_{\mathcal{F}}$ is the Frobenius norm (for a matrix $\mathbf{A}$, $\|\mathbf{A}\|_{\mathcal{F}} := \text{trace}(\mathbf{AA}^*)^{1/2}$). Then, the choice of the rank k can be viewed as a model selection issue. Let recall that in the matrix factorization framework, several approaches have been proposed in the literature (see for instance [129, 134, 86] for criteria based on the estimated eigenvalue study, [32, 131] for the classical BIC criterion or [135] for a cross-validation strategy).

When dealing with temporal series, the matrix $\mathbf{M}$, besides being of low rank, can have a temporal structure as periodicity, smoothness, etc. It is likely that the temporal properties of the data can be exploited to obtain an accurate factorization. Recently, [24] has extended the latter factorization method in order to take into account the time series trends properties. To this aim, they assume that the matrix $\mathbf{\Theta^0}$ is structured as follows :

$$\mathbf{\Theta}^0 = \mathbf{T^0\Lambda}, \tag{2.3}$$

where $\mathbf{T^0}$ is a $d \times \tau$ matrix of low rank k (thus with $\tau > k$) and $\mathbf{\Lambda}$ is a known $\tau \times n$ full rank matrix reflecting the temporal structure of the data. To estimate $\mathbf{\Theta^0}$, they developed a penalized least squares criterion (based on the Frobenius norm) and shown that, on the theoretical side, to take into account trends properties in the definition of the denoising estimator allowed to improve existing risk bounds. The penalization aims

to choose two parameters : the rank k of the matrix and the parameter τ related to the temporal structure. This penalty function depends also on the noise structure and involves an unknown constant.

In practice, this joint model selection issue is not standard. In addition to a penalty constant to be chosen, parameters from the distribution of the noise need to be estimated in advance. In this paper, we propose an automatic way to deal with these two problems. First, the parameters of the noise distribution are combined with the penalty constant to get a penalty function involving a single constant. This avoids having to estimate the parameters beforehand. Then, we propose a two-stage strategy, as in [?, 132], combined with the use of a heuristic for the constant calibration problem. Several heuristics have been considered here, now well-known for the penalty constant calibration in model selection frameworks [37, 40, 38]. We demonstrate the performances of our procedure and compare the considered heuristics in the case of independent Gaussian noises through simulation experiments. The robustness to an autoregressive noise is also studied.
The proposed method has been implemented in the R package `TrendTM`, for "Trend of High-Dimensional Time Series Matrix Estimation", which is available on the CRAN and presented here. Moreover, natural applications to curves clustering and principal component analysis (PCA) are investigated in this paper. A comparison to recent curves clustering methods (reviewed and compared in [43]) is performed on two real datasets (ECG profiles and children growth profiles) and shows that our method is competitive compared to the best existing one in terms of clustering objective but much faster.

The chapter is organized as follows. Section 2.1.1 recalls the estimation procedure proposed in [24]. Section 2.1.2 presents the proposed two-stage heuristic for the joint selection of the rank and the trend parameter whose performances are studied in Section 2.1.3 on simulated data. Section 2.1.4 gives some details and guidelines on the proposed method in the `TrendTM` package and shows an application on real data. In Section 2.1.5, we show that this method can be used for two classical statistical problems which are the time series clustering and PCA. As mentioned above, illustrations on real datasets are given.

2.1.1 Recall of the trend estimation method proposed by [24]

In this section, we present the method they proposed for estimating $\mathbf{\Theta}^0$ in model (2.2) when $\mathbf{\Theta}^0 = \mathbf{T}^0\mathbf{\Lambda}$ (see (2.3)) and when the noise ε has Gaussian i.i.d. rows of covariance matrix $\mathbf{\Sigma}_\varepsilon$. First, the two temporal structures they considered are the following :

- **periodicity :** if the trend of $\mathbf{M}$ is τ-periodic, then $\mathbf{\Lambda} = (\mathbf{I}_\tau \mid \cdots \mid \mathbf{I}_\tau)$ where $\mathbf{I}_\tau$ is the identity matrix in $\mathcal{M}_{\tau,\tau}(\mathbb{R})$,
- **smoothness :** if the form of the trend is $t \in \{1,\dots,n\} \mapsto f(t/n)$ with $f \in \mathbb{L}^2([0,1];\mathbb{R}^d)$, then

$$\mathbf{\Lambda} = \left(\varphi_\ell\left(\frac{t}{n}\right)\right)_{(\ell,t)\in\{1,\dots,\tau\}\times\{1,\dots,n\}},$$

  where τ is odd and $(\varphi_1,\dots,\varphi_\tau)$ is the τ-dimensional trigonometric basis defined by

$$\varphi_\ell(x) := \begin{cases} 1 & \text{if} \quad \ell = 1 \\ \sqrt{2}\cos(2\pi m x) & \text{if} \quad \ell = 2m \\ \sqrt{2}\sin(2\pi m x) & \text{if} \quad \ell = 2m+1 \end{cases}$$

for every $x \in [0,1]$ and $m \in \{1, \dots, (\tau-1)/2\}$.

So, the estimation procedure consists in two steps :

Step 1 : Estimation of $\mathbf{\Theta}^0$ for k and τ being fixed. They define the following auxiliary model

$$\underline{\mathbf{M}} = \mathbf{T^0} + \underline{\varepsilon}, \tag{2.4}$$

where $\underline{\mathbf{M}} := \mathbf{M}\mathbf{\Lambda}^+$, $\underline{\varepsilon} := \varepsilon\mathbf{\Lambda}^+$ and $\mathbf{\Lambda}^+ = \mathbf{\Lambda}^*(\mathbf{\Lambda}\mathbf{\Lambda}^*)^{-1}$ is the Moore-Penrose inverse of $\mathbf{\Lambda}$. This model doesn't embed some trend's property anymore. The least squares estimator of the matrix $\mathbf{T}^0$ is thus classical :

$$\widehat{\mathbf{T}}_{k,\tau} \in \arg\min_{\mathbf{A} \in \mathcal{S}_{k,\tau}} \|\underline{\mathbf{M}} - \mathbf{A}\|_{\mathcal{F}}^2, \tag{2.5}$$

where $\mathcal{S}_{k,\tau} \subset \{\mathbf{UV}\ ;\ \mathbf{U} \in \mathcal{M}_{d,k}(\mathbb{R}) \text{ and } \mathbf{V} \in \mathcal{M}_{k,\tau}(\mathbb{R})\}$. So, a natural estimator of $\mathbf{\Theta}^0$ is given by

$$\widehat{\mathbf{\Theta}}_{k,\tau} := \widehat{\mathbf{T}}_{k,\tau}\mathbf{\Lambda}.$$

Step 2 : Choice of k and τ. For a fixed $s > 0$, the final estimator of $\mathbf{\Theta^0}$ is $\widehat{\mathbf{\Theta}}_s := \widehat{\mathbf{\Theta}}_{\widehat{k}(s),\widehat{\tau}(s)}$ where

$$(\widehat{k}(s), \widehat{\tau}(s)) \in \arg\min_{(k,\tau)\in\mathcal{K}\times\mathcal{T}} \{\|\mathbf{M} - \widehat{\mathbf{\Theta}}_{k,\tau}\|_{\mathcal{F}}^2 + \mathrm{pen}_s(k,\tau)\}$$

with $\mathcal{K} \subset \{1, \dots, d \wedge n\}$, $\mathcal{T} \subset \{1, \dots, n\}$, and

$$\mathrm{pen}_s(k,\tau) := \mathfrak{c}_{\mathrm{pen}}\|\mathbf{\Sigma}_\varepsilon\|_{\mathrm{op}} k(d+\tau+s)\ ;\forall (k,\tau) \in \mathcal{K}\times\mathcal{T}, \tag{2.6}$$

where $\mathfrak{c}_{\mathrm{pen}} > 0$ is a deterministic constant and $\|.\|_{\mathrm{op}}$ is the operator norm on $\mathcal{M}_n(\mathbb{R})$ ($\|\mathbf{A}\|_{\mathrm{op}} := \sup_{\|x\|=1}\|\mathbf{A}x\|$ with $\|.\|$ the Euclidean norm on $\mathbb{R}^n$). They establish an oracle-type inequality on the resulting estimator (see [24] (Theorem 4.1)) : for every $\theta \in (0,1)$, with probability larger than $1 - 2e^{-s}$,

$$\|\widehat{\mathbf{\Theta}}_s - \mathbf{\Theta}^0\|_{\mathcal{F}}^2 \leqslant \min_{(k,\tau)\in\mathcal{K}\times\mathcal{T}} \min_{\mathbf{A}\in\mathcal{S}_{k,\tau}} \left\{ \left(\frac{1+\theta}{1-\theta}\right)^2 \|\mathbf{A}\mathbf{\Lambda} - \mathbf{\Theta}^0\|_{\mathcal{F}}^2 + \frac{4}{\theta(1-\theta)^2}\mathrm{pen}_s(k,\tau) \right\}. \tag{2.7}$$

2.1.2 The proposed two-stage heuristic for the model selection issue in practice

Discussion on the penalty function. The penalty function given by (2.6) depends of some constants s and $\mathfrak{c}_{\mathrm{pen}}$ that must be chosen or calibrated in practice. It also depends on the parameters of the noise distribution through $\|\mathbf{\Sigma}_\varepsilon\|_{\mathrm{op}}$ that must be estimated thus in advance : we explicit just below this norm in two cases that are considered in the simulation study (Section 2.1.3) :

- when the errors are uncorrelated ($\mathrm{cov}(\varepsilon_{1,t}, \varepsilon_{1,t'}) = \sigma^2 \mathbb{1}_{t=t'}$), then

$$\|\mathbf{\Sigma}_\varepsilon\|_{\mathrm{op}} = \sigma^2,$$

- when $(\varepsilon_{1,t})_t$ is a zero-mean stationary AR(1) Gaussian process (defined as the solution of $\varepsilon_{1,t} = \rho\varepsilon_{1,t-1} + \eta_{1,t}$ where $\rho \in (-1,1)$ and $(\eta_{1,t})_t$ is a white noise of standard deviation σ), we can show that

$$\|\mathbf{\Sigma}_\varepsilon\|_{\text{op}} \geqslant \sigma^2(1+\rho) =: f(\rho). \tag{2.8}$$

  Indeed, the covariance matrix of a row noise $(\varepsilon_0, \ldots, \varepsilon_{n-1})$ is

$$\mathbf{\Sigma}_\varepsilon := (\sigma^2\rho^{|i-j|})_{i,j}.$$

  Then, for every $x \in \mathbb{R}^n$ such that $\|x\| = 1$,

$$\begin{aligned} x^*\mathbf{\Sigma}_\varepsilon x &= \sum_{i,j=1}^{n} x_i x_j [\mathbf{\Sigma}_\varepsilon]_{i,j} = \sigma^2 \left(\|x\|^2 + \sum_{i \neq j} x_i x_j \rho^{|i-j|} \right) \\ &= \sigma^2 \left(1 + 2\sum_{i>j} x_i x_j \rho^{i-j} \right). \end{aligned}$$

  Since $\mathbf{\Sigma}_\varepsilon$ is a symmetric matrix,

$$\begin{aligned} \|\mathbf{\Sigma}_\varepsilon\|_{\text{op}} &= \sup_{\|x\|=1} |x^*\mathbf{\Sigma}_\varepsilon x| \geqslant |\mathbf{x}^*\mathbf{\Sigma}_\varepsilon\mathbf{x}| \quad \text{with} \quad \mathbf{x} = \frac{1}{\sqrt{2}}(1,1,0,\ldots,0) \\ &\geqslant \mathbf{x}^*\mathbf{\Sigma}_\varepsilon\mathbf{x} = \sigma^2 \left(1 + 2 \cdot \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} \cdot \rho^{2-1} \right) = \sigma^2(1+\rho). \end{aligned}$$

Two-stage heuristic. We set $\mathfrak{c}_{\text{cal}} = \mathfrak{c}_{\text{pen}}\|\mathbf{\Sigma}_\varepsilon\|_{\text{op}}$, representing a global penalty constant that we propose to calibrate using the data. So, this allows us to avoid the estimation of the noise distribution parameters, which turns out to be a difficult task. The penalty function is thus reduced to

$$\text{pen}(k,\tau) := \mathfrak{c}_{\text{cal}} k(d+\tau+s) \;; \forall (k,\tau) \in \mathcal{K} \times \mathcal{T},$$

and the resulting adaptive estimator is denoted by $\widehat{\mathbf{\Theta}}_s := \widehat{\mathbf{\Theta}}_{\widehat{k},\widehat{\tau}}$.

If the constant s can be easily chosen, this is not the case for the penalty constant $\mathfrak{c}_{\text{cal}}$. Several heuristics have been proposed in the literature for this purpose in model selection frameworks, but for a one-dimensional parameter only (see [37, 40]). First, in practice, we could take $s = -\log((1-\alpha)/2)$ with α fixed to 99%, 95% or 90%. Here we choose to fix $s = 4$. Then, for the selection of both k and τ, we follow the same strategy than in [42], that is a two-stage heuristic. We first recall some heuristics for the selection of one parameter, and then we present the two-stage heuristic for the joint selection of (k,τ).

Up to our knowledge, there exit the three following heuristics dedicated to the constant calibration question in the model selection frameworks of one parameter :

- the one proposed in [37], denoted here ML, that involves a threshold S which is fixed to $S = 0.75$ as suggested by the author, and
- the two proposed in [40] (see the more recent versions [38] and [39]) that are two versions of the well-known slope heuristic : the 'dimension jump' and the 'slope', denoted here BJ and Slope respectively. The both heuristics have been implemented in the R package `capushe` described in [41].

For the joint selection of (k, τ), the two-stage heuristic is the following : first, we choose the best τ for each $k \in \mathcal{K}$ via the criterion

$$\widehat{\tau}(k) \in \arg\min_{\tau \in \mathcal{T}} \{\|\mathbf{M} - \widehat{\mathbf{\Theta}}_{k,\tau}\|_{\mathcal{F}}^2 + \mathfrak{c}_{\text{cal},\tau} k(d + \tau + s)\},$$

where the penalty constant $\mathfrak{c}_{\text{cal},\tau}$ is calibrated using one of the previous heuristics, and then we select the best k among them via the criterion

$$\widehat{k} \in \arg\min_{k \in \mathcal{K}} \{\|\mathbf{M} - \widehat{\mathbf{\Theta}}_{k,\widehat{\tau}(k)}\|_{\mathcal{F}}^2 + \mathfrak{c}_{\text{cal},k} k(d + \widehat{\tau}(k) + s)\},$$

where the penalty constant $\mathfrak{c}_{\text{cal},k}$ is calibrated using the same heuristic to be consistant, and $\widehat{\tau} = \widehat{\tau}(\widehat{k})$.

Note that in practice $\mathcal{K} = \{1, \ldots, k_{\max}\}$ and $\mathcal{T} = \{k+1, \ldots, \tau_{\max}\}$, where $k_{\max}$ is the maximal rank and $\tau_{\max}$ is the maximal value of τ. These two quantities need to be specified.

2.1.3 Simulation study

In this study, we conduct different numerical experiences to both evaluate the performance of the proposed method and compare the three different heuristics :

- Study 1 : we consider the model selection issue for k and τ separatly,
- Study 2 : we illustrate the importance to take into account the trend in the estimation procedure when it exits,
- Study 3 : we consider the model selection issue for both k and τ.

In these three studies, the errors are assumed to be uncorrelated. We perform an additional simulation study :

- Study 4 : we study the robustness of the proposed method to temporal dependency modelled through an AR process when both k and τ are selected.

2.1.3.1 Simulation design and quality criteria

Simulation design. We've simulated datasets with $d = 100$ and $n = 600$ as follows :

(1) we generate a matrix $\underline{\mathbf{T^0}} = \mathbf{UV}$ by simulating $\mathbf{U} \in \mathcal{M}_{d,k}(\mathbb{R})$ and $\mathbf{V} \in \mathcal{M}_{k,\tau}(\mathbb{R})$ for which the entries of $\mathbf{U}$ and $\mathbf{V}$ are assumed to be i.i.d. and follows a centered Gaussian distribution with same standard deviation σ_{uv} fixed to 0.5 ;

(2) two cases are considered according to the presence or not of a trend in the simulated series : if there is no trend, then $\tau = n$ and $\mathbf{\Theta}^0 = \mathbf{T}^0$, and otherwise $\mathbf{\Theta}^0 = \mathbf{T^0\Lambda}$ with the matrix $\mathbf{\Lambda}$ of the smooth case. To distinguish between these two cases in the sequel, we call them $\text{dataset}_{\text{NoTrend}}$ and $\text{dataset}_{\text{Trend}}$ respectively ;

(3) the rows of the error matrix ε are assumed to be i.i.d. and follow a centered Gaussian distribution of variance σ^2 (i.e. $\mathbf{\Sigma}_\varepsilon = \sigma^2 \mathbf{I}_n$) for Studies 1, 2 and 3 ; the rows of the error matrix ε are assumed to be i.i.d. stationary AR(1) Gaussian processes with a white noise of standard deviation $\sigma > 0$ and an autocorrelation parameter $\rho \in (-1, 1)$ for Study 4.

We take $k = 3$ and $\tau = 25$. We consider different values for the residual standard deviation σ in order to have different levels of difficulty for the estimation problem. First, according to the previous considerations, $\text{var}(\mathbf{\Theta^0}_{ij}) = k\sigma_{uv}^4$ for $\text{dataset}_{\text{NoTrend}}$ and $\text{var}(\mathbf{\Theta^0}_{ij}) = \tau k\sigma_{uv}^4$ for $\text{dataset}_{\text{Trend}}$. For Studies 1, 2 and 3, let us consider $s_v \in \{0.1, 0.5, 1.5, 2\}$. In order to have the same estimation difficulty (same ratio between σ and the standard deviation of $\mathbf{\Theta}^0_{ij}$) for the two datasets, we set $\sigma = s_v$ for $\text{dataset}_{\text{NoTrend}}$ and $\sigma = \sqrt{\tau}s_v$ for $\text{dataset}_{\text{Trend}}$. The obtained four cases are judged as 'Easy', 'Medium', 'Difficult' and 'Hard' respectively. Study 4 is the same as Study 3 but with a noise modeled by an autoregressive process. More precisely, we consider two values for the standard deviation of the noise $s_v \in \{0.1, 1.5\}$ and an autocorrelation parameter $\rho \in \{-0.8, -0.3, 0, 0.3, 0.8\}$. For each combination of parameters, we've simulated 200 datasets.

Let us precise that when the trend is not considered in the estimation procedure, the resulting estimator is

$$\widehat{\mathbf{\Theta}}_{k \text{ or } \widehat{k},n} \text{ (if } k \text{ is selected or not)},$$

and when it is considered the resulting estimator is

$$\widehat{\mathbf{\Theta}}_{k \text{ or } \widehat{k},\tau \text{ or } \widehat{\tau}} \text{ (if both } k \text{ and } \tau \text{ are selected or one of them or none)}.$$

Quality criteria. The performance of our procedure is assessed via :

- the estimated k and/or τ ; and
- the squared Frobenius distance between $\mathbf{\Theta^0}$ and its estimate $\widehat{\mathbf{\Theta}}_{\widehat{k},\widehat{\tau}}$.

Moreover, we also consider the Frobenius distance between $\mathbf{M}$ and

- the estimator of $\mathbf{\Theta^0}$ for the true k and/or τ, that is $\widehat{\mathbf{\Theta}}_{k,\tau}$; and
- the trajectorial oracle, that is $\widehat{\mathbf{\Theta}}_{\widetilde{k},\widetilde{\tau}}$ where

$$(\widetilde{k}, \widetilde{\tau}) = \arg \min_{(k,\tau)\in\mathcal{K}\times\mathcal{T}} \|\mathbf{\Theta}^0 - \widehat{\mathbf{\Theta}}_{k,\tau}\|_{\mathcal{F}}^2$$

  when both k and τ are selected, $\widehat{\mathbf{\Theta}}_{k,\widetilde{\tau}}$ where

$$\widetilde{\tau} = \arg \min_{\tau\in\mathcal{T}} \|\mathbf{\Theta}^0 - \widehat{\mathbf{\Theta}}_{k,\tau}\|_{\mathcal{F}}^2$$

  when k is fixed, and $\widehat{\mathbf{\Theta}}_{\widetilde{k},n}$ where

$$\widetilde{k} = \arg \min_{k\in\mathcal{K}} \|\mathbf{\Theta}^0 - \widehat{\mathbf{\Theta}}_{k,n}\|_{\mathcal{F}}^2$$

  when no trend is considered.

2.1.3.2 Study 1 : behavior of the three heuristics for the selection of k or τ

We first study the selection of k for $\text{dataset}_{\text{NoTrend}}$ when no trend is considered in the estimation procedure. We consider two different values of the maximal rank $k_{\max} \in \{15, 35\}$. The results are presented in Figure 2.1. When the noise is small, i.e. the estimation problem is easy (cases 'Easy' and 'Medium'), all the heuristics recover the true rank, and therefore the obtained estimators perform as well as $\widehat{\mathbf{\Theta}}_{k,n}$ (the estimator

of $\mathbf{\Theta^0}$ for k fixed to its true value). When the estimation problem gets more difficult (cases 'Difficult' and 'Hard'), the heuristics tend to underestimate the rank. This underestimation behavior seems to be logical and even desirable in the particular 'Hard' case. Indeed, we observe that in terms of Frobenius norm, the obtained estimators perform better compared to the one with the true rank. Moreover, they have performance close to the oracle. Comparing the three heuristics, the Slope heuristic shows better performances compared to the two other heuristics. This is particularly marked for the 'Medium' case and $k_{\max} = 15$. We can note that the behavior of the three heuristics can be affected by the choice of $k_{\max}$. This problem is well-known for both the BJ and Slope heuristics (see [39] for more explanations in the case univariate series analysis).

Then, we study the selection of τ for dataset$_{\text{Trend}}$ for k fixed to the true value. We fix $\tau_{\max} = 55$. The results are presented in Figure 2.2. Except with BJ that is more unstable, the heuristics retrieve the true value of τ whatever the estimation difficulty with same performance as the oracle.

From this study, we choose the Slope heuristic for the model selection issue for both k and τ in the sequel and in the developed package.

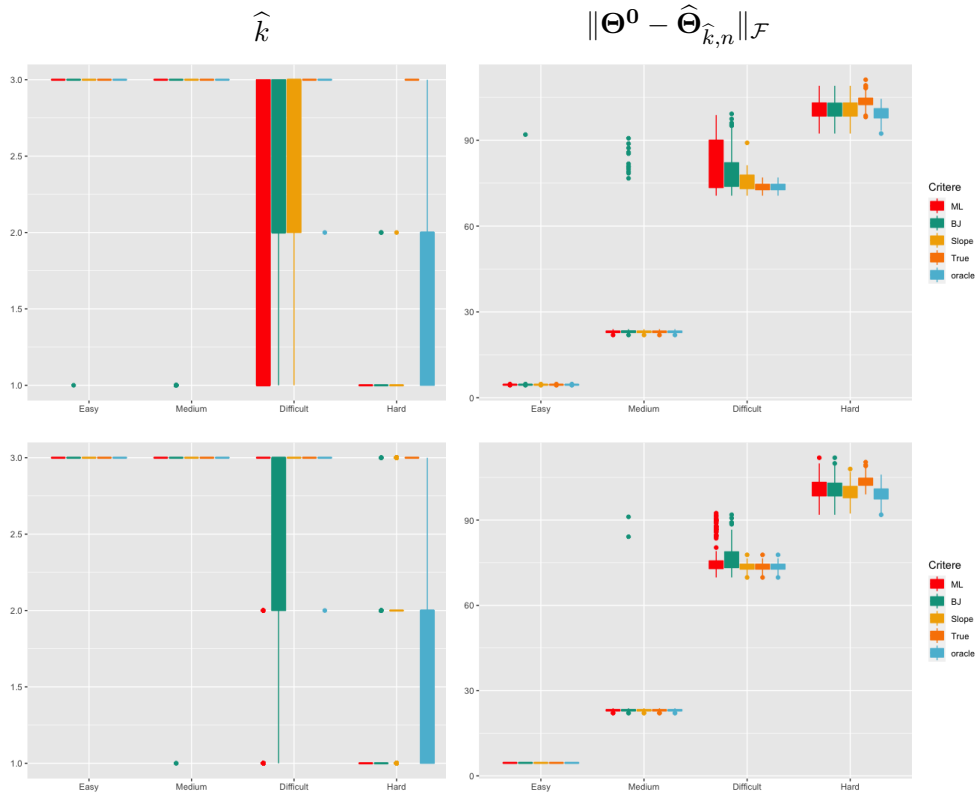

FIGURE 2.1 – Comparison of the three heuristics for the selection of k for dataset$_{\text{NoTrend}}$ (Study 1). Left : estimated number of the rank k and right : boxplot of $\|\mathbf{\Theta^0} - \widehat{\mathbf{\Theta}}_{\widehat{k},n}\|_{\mathcal{F}}$ for two values of $k_{\max} = 15$ (first line) and $k_{\max} = 35$ (second line), and different values of σ. On each graph and for each value of σ, from left to right, we have the result from ML, BJ, Slope ($\widehat{k}$), the true rank (k) and the oracle ($\widetilde{k}$).

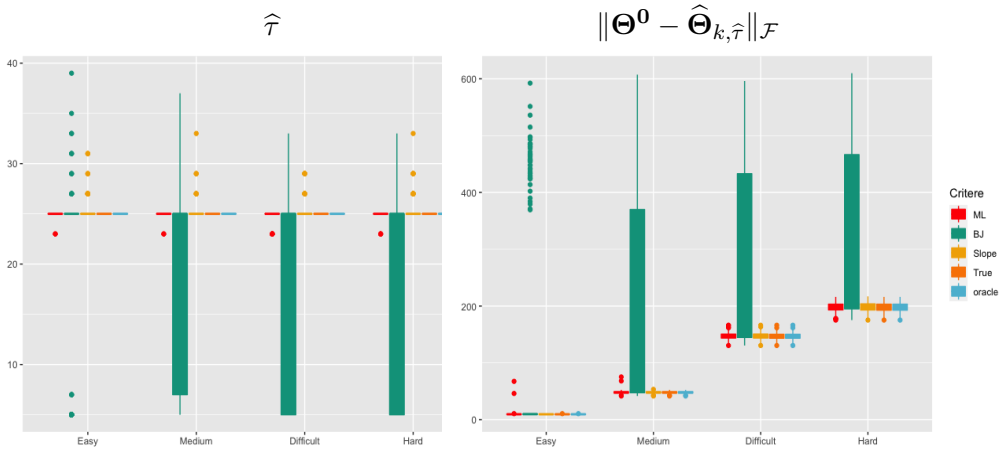

FIGURE 2.2 – Comparison of the three heuristics for the selection of τ for $\text{dataset}_{\text{Trend}}$ when k is fixed to the truth ($k = 3$) for different values of σ (Study 1). Left : estimated τ and right : boxplot of $\|\mathbf{\Theta^0} - \widehat{\mathbf{\Theta}}_{k,\widehat{\tau}}\|_{\mathcal{F}}$. In each graph and each value of σ, from left to right, we have the result from ML, BJ, Slope ($\widehat{\tau}$), the true value (τ) and the oracle ($\widetilde{\tau}$).

2.1.3.3 Study 2 : accounting for the smooth structure in the trend

We compare the performance of the procedure on the $\text{dataset}_{\text{Trend}}$ when the trend is considered ($\tau = \widehat{\tau}$) or not ($\tau = n$) for k fixed to the true value. We choose $\tau_{\max} = 55$. The results are represented in Figure 2.3. Whatever the difficulty of the estimation problem (different values of σ), accounting for the trend increases the precision of the estimation. This is more marked for high values of σ. Note that, similarly as Study 1, the estimation naturally degrades with the increasing of σ.

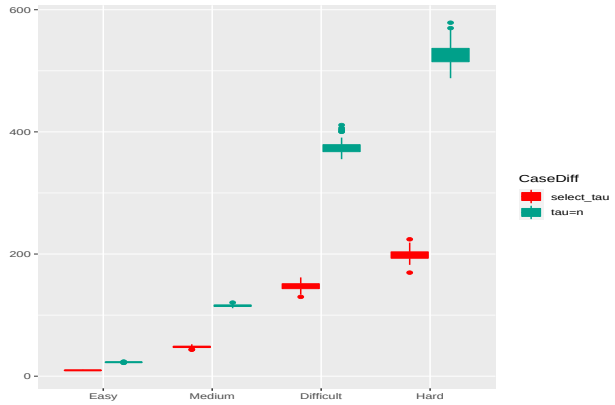

FIGURE 2.3 – Boxplot of $\|\mathbf{\Theta^0} - \widehat{\mathbf{\Theta}}_{k,\tau}\|_{\mathcal{F}}$ with $\tau = \widehat{\tau}$ (`select_tau`) and $\tau = n$ (`tau = n`) for different values of σ (Study 2).

2.1.3.4 Study 3 : selection of k and τ

Figure 2.4 shows that the joint heuristic retrieves the true values of k and τ whatever the difficulty of the estimation problem, except very few times. Thus, the performance of the estimator $\widehat{\mathbf{\Theta}}_{\widehat{k},\widehat{\tau}}$ is comparable to the one of the estimator $\widehat{\mathbf{\Theta}}_{k,\tau}$ and moreover it

has performance close to the oracle (see Figure 2.5). Compared to Study 1 where $\tau = n$, here for difficult estimation problems, k is not underestimated.

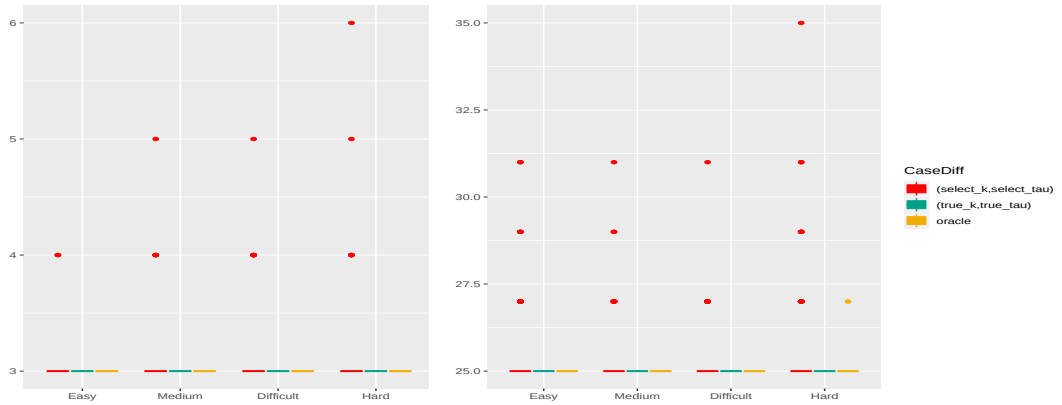

FIGURE 2.4 – Left : estimated k. Right : estimated τ for different values of σ (Study 3).

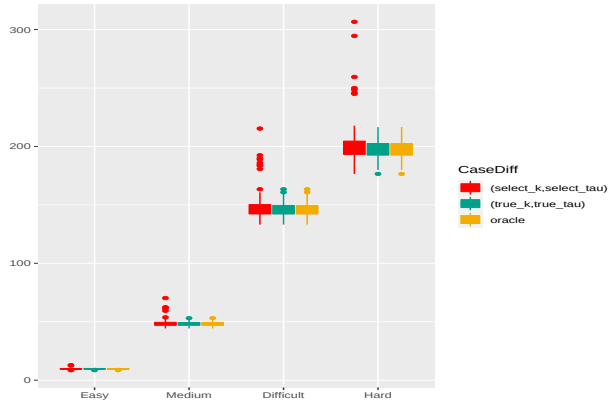

FIGURE 2.5 – Boxplot of $\|\mathbf{\Theta^0} - \widehat{\mathbf{\Theta}}_{k,\tau}\|_{\mathcal{F}}$ with $(k,\tau) = (\widehat{k},\widehat{\tau})$ the selected k and τ, $(k,\tau) = (k,\tau)$ the true values and $(k,\tau) = (\widetilde{k},\widetilde{\tau})$ the oracle for different values of σ (Study 3).

2.1.3.5 Study 4 : robustness to autocorrelated noise

Whatever the dependence and the noise variance, the joint heuristic retrieves the true values of k and τ (see Figures 2.6 and 2.8), except for a large variance ($s_v = 1.5$) and a high positive autocorrelation ($\rho = 0.8$) where it underestimates k and the selection of τ is more variable. For all noise cases, the method leads to estimators that have close performance compared to the oracle (see Figures 2.7 and 2.9) and with better performance than the one with the true values for the excepted case.
Moreover, we can observe that the more the autocorrelation parameter ρ increases (from -1 to 1), the more $\|\mathbf{\Theta^0} - \widehat{\mathbf{\Theta}}_{k,\tau}\|_{\mathcal{F}}$ increases also with a noticeable gap between $\rho = 0.3$ and $\rho = 0.8$ for both values of s_v. First, the estimation is better with high and negative autocorrelation. Then, the observed phenomenon on the norm can be explained. The

variance term in the risk bound of the estimator $\widehat{\mathbf{\Theta}}_{\widehat{k},\widehat{\tau}}$ (i.e. the penalty, see (2.7)) depends on $\|\mathbf{\Sigma}_\varepsilon\|_{\text{op}}$ (see (2.6)). Using (2.8), we can show that this term is lower-bounded by

$$f(\rho)\frac{k(d+\tau)}{dT}$$

up to a multiplicative constant where f is increasing and nonnegative on $[-1,1]$.

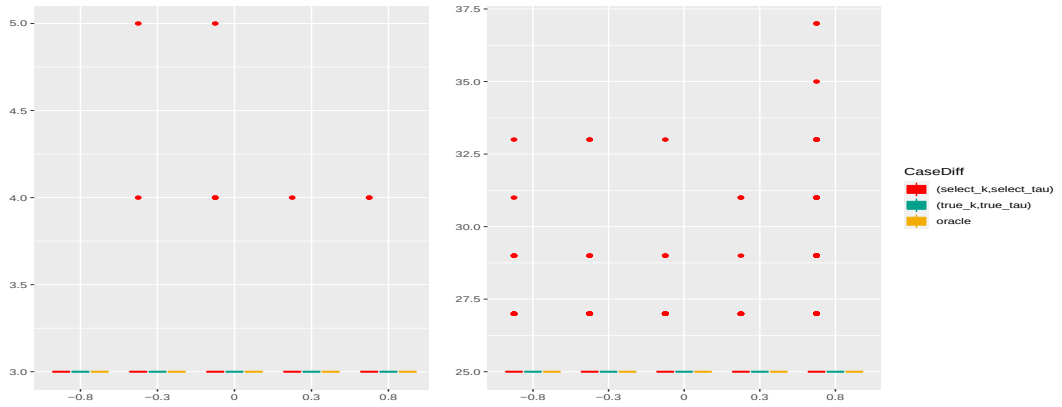

FIGURE 2.6 – Left : estimated k. Right : estimated τ for different values of ρ and for the standard deviation $s_v = 0.1$. (Study 4)

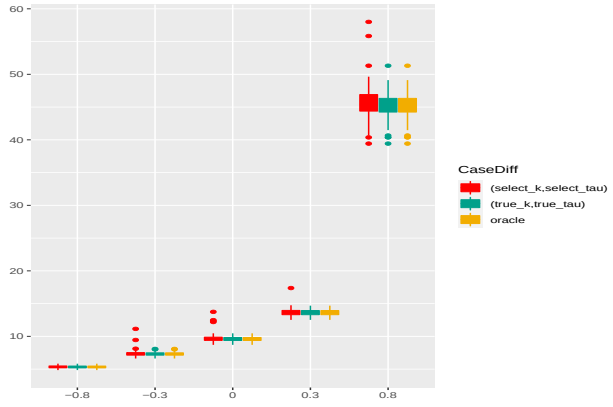

FIGURE 2.7 – Boxplot of $\|\mathbf{\Theta^0} - \widehat{\mathbf{\Theta}}_{k,\tau}\|_{\mathcal{F}}$ with $(k,\tau) = (\widehat{k},\widehat{\tau})$ the selected k and τ, $(k,\tau) = (k,\tau)$ the true values and $(k,\tau) = (\widetilde{k},\widetilde{\tau})$ the oracle for different values of ρ and for the standard deviation $s_v = 0.1$. (Study 4)

2.1.4 Using the `TrendTM` package

2.1.4.1 Comments on the package

The package is organized around the main and unique function `TrendTM`. In this section, we present the arguments used in a call of this function :

```
TrendTM(M, k.select=FALSE, k.max=20, struct.temp="none", tau.select=FALSE,
tau.max=floor(n/2), type.soft="als")
```

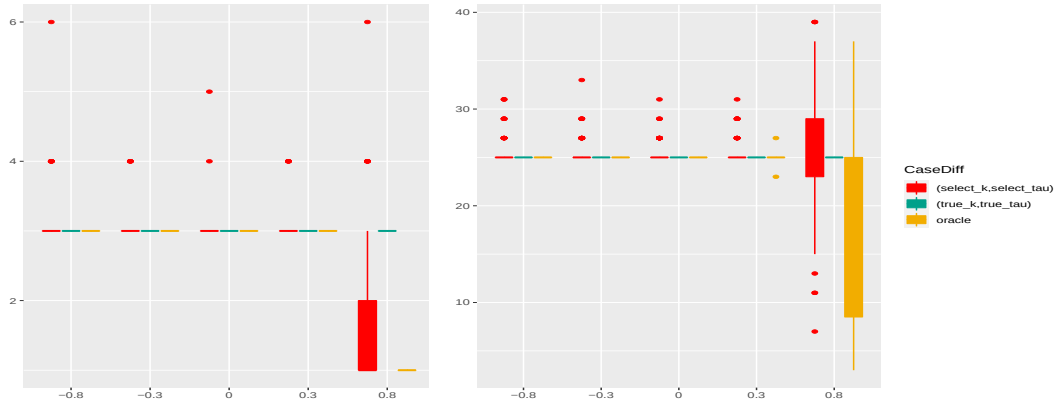

FIGURE 2.8 – Left : estimated k. Right : estimated τ for different values of ρ and for the standard deviation $s_v = 1.5$. (Study 4)

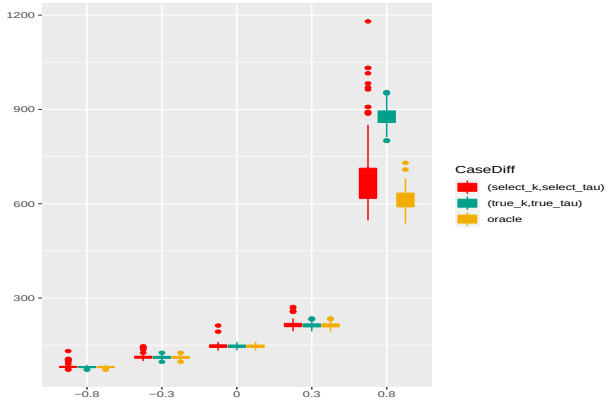

FIGURE 2.9 – Boxplot of $\|\mathbf{\Theta^0} - \widehat{\mathbf{\Theta}}_{k,\tau}\|_{\mathcal{F}}$ with $(k,\tau) = (\widehat{k},\widehat{\tau})$ the selected k and τ, $(k,\tau) = (k,\tau)$ the true values and $(k,\tau) = (\widetilde{k},\widetilde{\tau})$ the oracle for different values of ρ and for the standard deviation $s_v = 1.5$. (Study 4)

This function returns a list containing six elements :

- `k.est`, the estimated k or the true k when no selection is chosen ;
- `tau.est`, the estimated τ or the true τ when no selection is chosen ;
- `Theta.est`, the estimation of $\mathbf{\Theta}^0$ (`Theta.est` = `U.estV.est` if no temporal structure is considered and `Theta.est` = `U.estV.est`$\mathbf{\Lambda}$ if a temporal structure is considered) ;
- `U.est`, the component U of the decomposition of $\widehat{\mathbf{\Theta}}$;
- `V.est`, the component V of the decomposition of $\widehat{\mathbf{\Theta}}$;
- `contrast`, the squared Frobenius norm of `M - Theta.est`. If k and τ are fixed, the contrast is an unique value ; if k is selected and τ is fixed or if τ is selected and k is fixed, the contrast is a vector containing the norms for each visiting values of k or τ respectively ; and if k and τ are selected, the contrast is a matrix with $k_{\max}$ rows and $\tau_{\max}$ columns such that $\text{contrast}_{k,\tau} = \|\mathbf{\Theta^0} - \widehat{\mathbf{\Theta}}_{k,\tau}\|_{\mathcal{F}}^2$.

Model selection. The selection of k or/and τ is requested using the options `k.select` or/and `tau.select` that are booleans. When there is no selection, the option is set to `FALSE` and `k.max` $= k$ or/and `tau.max` $= \tau$. Note that if no trend is considered in the estimation procedure, $\tau = n$, otherwise `tau.max` must be a smaller than n and larger than `k.max` $+ 2$ in order to ensure that the rank of $\mathbf{\Theta^0}$ is k.

Taking into account the trend. Let us give more details about the different arguments of `TrendTM` that need to be specified when accounting for a temporal structure in the estimation procedure.

Two temporal structures are considered : periodic trend and smooth trend. This can be specified using the option `struct.temp`, `struct.temp="periodic"` or `struct.temp="smooth"` respectively. Recall that the selection of τ is only possible when a smooth trend is considered. Thus, when

- `struct.temp="periodic"`, then `tau.select=FALSE` and `tau.max`$= \tau$. In this case, τ must be such that n is a multiple of τ ;
- `struct.temp="smooth"`, then `tau.select` is either `FALSE` or `TRUE`. Whatever this choice, `tau.max` must be an odd number.

When no trend is taken into account, `struct.temp="none"` and `tau.max` $= n$.

Estimation of $\mathbf{\Theta}^0$, k and τ being fixed (Step 1). The least squares estimator $\widehat{\mathbf{\Theta}}_{k,\tau}$ of $\mathbf{\Theta}^0$, given by (2.5), is obtained by using the `softImpute` function from the R package of the same name developed by Hastie and Mazumder for matrix completion [128]. In this package, two algorithms are implemented : 'svd' and 'als'. In a simulation study, we observed that they have both provided the same accuracy of the estimator (results not shown). We decide to use the 'als' algorithm by default but the choice is left free to the user in our package `TrendTM`. In this package, this choice is specified using the option `type.soft`.

The slope heuristic. Let us now focus on the selection problem of the rank k, and write the penalty as $\text{pen}(k) = \mathfrak{c}_{\text{cal},k}\varphi(k)$. The Slope heuristic, proposed by [40], consists in estimating the slope $\widehat{s}$ of the contrast $\|\mathbf{M} - \widehat{\mathbf{\Theta}}_{k,n}\|^2_{\mathcal{F}}$ as a function of $\varphi(k)$ with k 'large enough' and defining $\mathfrak{c}_{\text{cal},k} = -2\widehat{s}$. The implementation of this heuristic requires the choice of the dimensions on which to perform the regression, that can be difficult in practice. To deal with this problem, [41] proposed to make robust regressions for dimensions between k and $k_{\max}$ for $k = 1, 2, \ldots$, resulting in different selected $\widehat{k}$. The choice of the final dimension is the maximal value $\widehat{k}$ such that the length of successive same $\widehat{k}$ is greater than the option point of the function DDSE. In order to avoid some implementation problems as such condition is not reached and no k is selected, we decide to take the value $\widehat{k}$ associated to the maximal length of successive same $\widehat{k}$.

2.1.4.2 Application to pollution dataset

Let us use the package on a real dataset[1]. The dataset contains the amounts of $d = 13$ toxic gases in the air recorded $n = 9357$ times during one year. We do a first

1. available at https ://archive.ics.uci.edu/ml/datasets/Air+quality

step of data imputation using the function `complete` of the package `softImpute` since missing values (coded with -200) exist in this dataset (see [133] for more details on high-dimensional time series completion). Our function `trendTM` is then applied on the completed matrix with a selection of both k and τ. The code is the following :

```
## Used package
library('softImpute')
library('TrendTM')

## Data importation
AirPollution <- read.table('AirQualityUCI.csv',sep=";",header=TRUE)[,-c(1:2)]
AirPollution <- t(as.matrix(AirPollution))

## Imputation
AirPollution[AirPollution==-200] <- NA
AirPollutionSoft=softImpute::softImpute(AirPollution,rank=1,lambda=0)
AirPollutionImp=softImpute::complete(AirPollution,AirPollutionSoft)

## Trend estimation
Trend.AirPollution = TrendTM(AirPollutionImp, k.select=TRUE, k.max=13,
struct.temp="smooth", tau.select=TRUE, tau.max=101)

Trend.AirPollution$k.est
[1] 7
Trend.AirPollution$tau.est
[1] 13
```

The procedure selects $\widehat{k} = 7$ and $\widehat{\tau} = 13$. Figure 2.10 shows the obtained trend estimation for 4 toxic gases among the 13. The denoising process seems to have been well applied to the data.

2.1.5 Method used for usual problems in statistics

This section deals with an application of Model (2.1), first to times series clustering, and then to PCA of times series. These applications were not developed in [24].

2.1.5.1 Application to time series clustering

To assume that $\mathbf{\Theta}^0 = \mathbf{UL}$ with $\mathbf{L} = (\ell_j(t))_{j,t} := \mathbf{V\Lambda}$ means that for any $i \in \{1, \ldots, d\}$, the trend $m_i(.)$ of the i-th row of $\mathbf{M}$ satisfies

$$m_i(t) = \sum_{j=1}^{k} \mathbf{U}_{i,j} \ell_j(t) \ ; \ \forall t \in \{1, \ldots, T\}.$$

In other words, the trend of the i-th time series stored in $\mathbf{M}$ is a linear combination of the trends $\ell_1(.), \ldots, \ell_k(.)$ of k latent time series. Moreover, since $\mathbf{L} = \mathbf{V\Lambda}$, $\ell_1(.), \ldots, \ell_k(.)$ have the same usual time series trend's property, characterized by $\mathbf{\Lambda}$, than $m_1(.), \ldots, m_d(.)$.

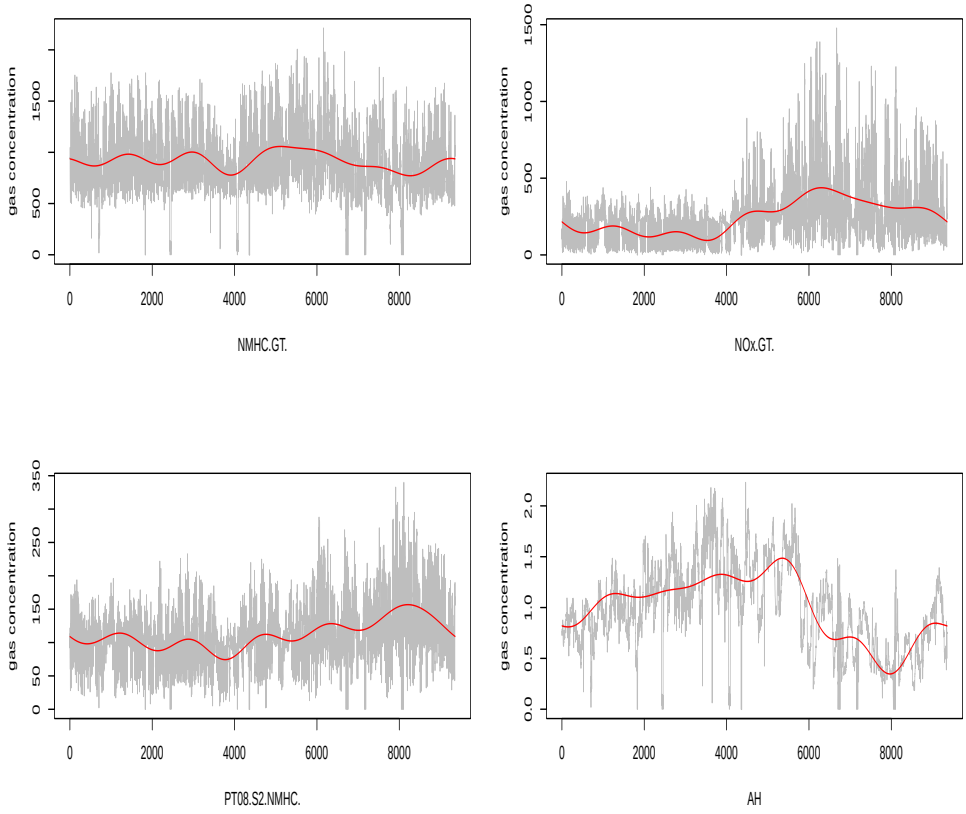

FIGURE 2.10 – Estimation of the trend of 4 toxic gases (red) compared to the initial series (grey).

The identification of the aforementioned latent time series is part of curves or time series clustering framework. To this aim, we propose here to apply the well-known Hierarchical Agglomerative Clustering with the Ward's linkage on the rows of $\widehat{\mathbf{U}}_k$, where $\widehat{\mathbf{T}}_{k,\tau} = \widehat{\mathbf{U}}_k\widehat{\mathbf{V}}_\tau$ or $\widehat{\mathbf{\Theta}}_{k,\tau} = \widehat{\mathbf{U}}_k\widehat{\mathbf{V}}_\tau\mathbf{\Lambda}$.

In [43], the authors applied and compared different curves clustering methods proposed in the literature on real datasets. We propose to apply our clustering strategy on two datasets : the ECG and Growth datasets. The ECG dataset (taken from the UCR Time Series Classification and Clustering website) consists in 200 electrocardiogram from 2 groups of patients sampled at 96 time instants in which 133 are classified as normal and 67 as abnormal, and the Growth dataset (available in the R package fda) contains the heights of 54 girls and 39 boys measured at 31 stages from 1 to 18 years. In both datasets, we have two groups. The time series of the two datasets are plotted in Figure 2.11.

We apply our proposed clustering strategy based on the procedure with and without taking into account for a trend (periodic for ECG and smooth for Growth) with the known number of groups $k = 2$. The Correct Classification Rates (CCR) according to the known partitions are given in Table 2.1 for the ECG dataset and in Table 2.2 for the Growth dataset. We also report the CCR obtained for the best method among the ones tested in [43]. In addition, we indicate the time taken by the different methods on a laptop 1.6 GHz CPU (note that the time for the method HDDC on FPCA scores is not given since this method is not available). For the ECG dataset, accounting for the trend

improves significantly the clustering performance. This is not the case with the Growth dataset which already has a very high CCR without trend. Our clustering performances are the same compared to the best clustering method but it is much more faster. Note that among the compared methods in [43], the best ones for the two datasets are not the same.

The R code for the Growth dataset without accounting the trend is the following

```
library(fda)
data("growth")
Growth <-t(cbind(matrix(growth\$hgtm,31,39),matrix(growth\$hgtf,31,54)))
cls <-c(rep(1,39),rep(0,54))

k.max <- 2
res.Growth <- TrendTM::TrendTM(Growth,k.max=k.max)

Uf.cr <- scale(res.Growth\$U.est,scale=TRUE, center=TRUE)
Uf.ward<- hclust(dist(Uf.cr, method = "euclidean")^2, method = "ward.D")
cluster.Uf <- cutree(Uf.ward, k =k.max)
table(cls,cluster.Uf)
```

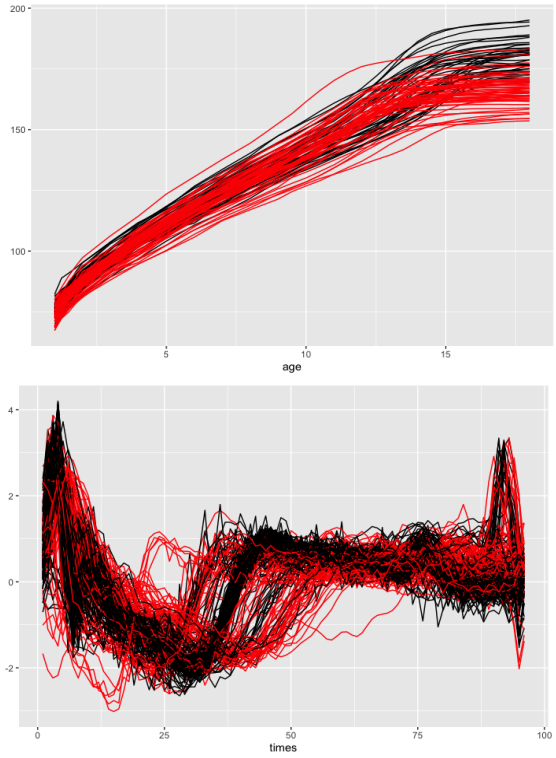

FIGURE 2.11 – Top : Growth dataset (black : boys, red : girls).
Bottom : ECG dataset (black : normal, red : abnormal).

Note that for the Growth dataset without trend, the Slope heuristic selects $\widehat{k} = 6$ clusters and for the ECG dataset with a 32-periodic trend, it selects $\widehat{k} = 7$. In Figure 2.12, the series of the ECG dataset are plotted and colored according to the $\widehat{k} = 7$ obtained clusters with a 32-periodic trend on the left, and the associated obtained trend $m_i(.)$ for $i = 1, \ldots, 6$ are plotted on the right.

	$\text{proc}_{\text{NoTrend}}$	$\text{proc}_{\text{Trend}}$	Best method in [43]
CCR	74.5	83	84 (reported from [43])
Mean times in second (on 30 runs)	0.015	0.008	19.2

TABLE 2.1 – Correct classification rates (CCR) in percentage accounting or not for a trend on the ECG dataset. Mean times in second obtained on 30 runs.

	$\text{proc}_{\text{NoTrend}}$	$\text{proc}_{\text{Trend}}$ smooth with $\widehat{\tau} = 13$	Best method in [43] HDDC on FPCA scores
CCR	97.85	97.85	97.85 (reported from [43])
Mean times in second (on 30 runs)	0.003	0.0028	.

TABLE 2.2 – Correct classification rates (CCR) in percentage accounting or not for a trend on the Growth dataset. Mean times in second obtained on 30 runs.

2.1.5.2 Application to time series PCA

The PCA problem can be rephrased as a k-rank approximation of a matrix $\mathbf{M}$ (information captured by the first k axes). More precisely, if the matrix $\mathbf{M}$ is of dimension $d \times n$, the simple PCA solution of rank k is given by

$$\widehat{\mathbf{\Theta}}_{k,n} = \widehat{\mathbf{U}}_k \widehat{\mathbf{V}}_k \in \arg \min_{\mathbf{U} \in \mathcal{M}_{d,k}, \mathbf{V} \in \mathcal{M}_{k,n}} \|\mathbf{M}^c - \mathbf{U}\mathbf{V}\|_{\mathcal{F}}^2$$

where, if A_j denotes the j-th column of the matrix A, $A_j^c := A_j - \overline{A}_j$ with $\overline{A}_j = d^{-1} \sum_{i=1}^{d} A_{ij}$. Then, the matrix of the principal components is $\mathbf{M}^c \widehat{\mathbf{V}}_k^* = \widehat{\mathbf{U}}_k$. Each line of this $d \times k$ matrix contains the coordinates of the projection of the associated time series on the first k axis. Two projected time series are close if they share globally the same trend's property. However, in high-dimensional space (n high), the euclidean distance used in PCA can lose its meaning and a local trend similarity could be preferred. We thus propose to perform the PCA on the transformed matrix $\underline{\mathbf{M}}^c = \mathbf{M}^c \mathbf{\Lambda}^+$. The solution is $\widehat{\mathbf{\Theta}}_{k,\tau} = \widehat{\mathbf{T}}_{k,\tau} \mathbf{\Lambda} = \widehat{\mathbf{U}}_k \widehat{\mathbf{V}}_\tau \mathbf{\Lambda}$ and the matrix of principal components is $\underline{\mathbf{M}}^c \widehat{\mathbf{V}}_k^* = \mathbf{M}^c \mathbf{\Lambda}^+ \widehat{\mathbf{V}}_k^*$.

The projections of the times series on the principal plan (coordinates given by the two first columns of the matrix of principal components) of $\mathbf{M}^c$ and $\underline{\mathbf{M}}^c$ are plotted in Figure 2.13 on the left and the right respectively for the ECG dataset colored according to the true partition. Recall that for the ECG dataset, we consider that the trend is periodic with period $\tau = 32$.

To illustrate the trend reduction using $\mathbf{\Lambda}^+$ on a period with length τ, the 28th and the 121th time series of $\mathbf{M}^c$ and $\underline{\mathbf{M}}^c$ are plotted in Figure 2.14 on left and right respectively. As expected according to their features in both matrices, they are not close in the PCA of $\mathbf{M}^c$ whereas they are close in the PCA of $\underline{\mathbf{M}}^c$. That also explains the difference when the clustering is performed without trend or with trend (see the previous section). For example, these two time series are clustered in the same group when a trend property

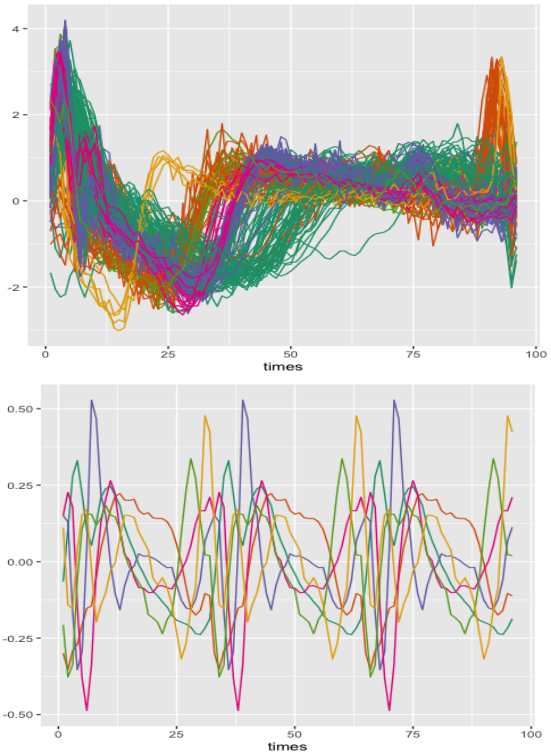

FIGURE 2.12 – Top : ECG dataset colored according to the $\widehat{k} = 7$ obtained clusters with a 32-periodic trend. Bottom : the associated trends $m_i(.)$ for $i = 1, \ldots, 6$.

is taken into account, whereas they are not in the same group when no trend property is considered.

2.1.6 Conclusion

The penalized criterion developed in [24] for high-dimensional time series analysis consists in selecting both the rank k of the matrix and the parameter τ related to the temporal structure. The penalty function involves a constant to be calibrated and depends on the temporal structure through its associated parameters to be estimated. For such minimization contrast estimation context, it is well-known in the literature that despite the selection of both parameters issue that is not standard, the calibration of penalty constant is not an easy task and many heuristics have been proposed to this aim. We proposed in this paper a two-stage strategy based on a popular heuristic : the slope heuristic proposed by [40] and used in many statistical problems. We conducted a large simulation study to compare different heuristics as well as to study the performance of the method. In particular, through these simulations, we show that the joint heuristic performs well : the true values of k and τ are retrieved or underestimated when the estimation problem is more difficult, but with good reasons (the estimation is better than with the true values in this case). Moreover, whatever all the tested cases, the performance of the final estimator is comparable to that of the oracle. The method has been implemented in the R package `TrendTM` yet available on the CRAN and which is detailed in this paper. We also show that this method can be used for classical problems in statistics : the clustering and the PCA. Especially, we show that for times series clustering, the proposed method works as well as the best ones proposed in the literature but is computationally much faster.

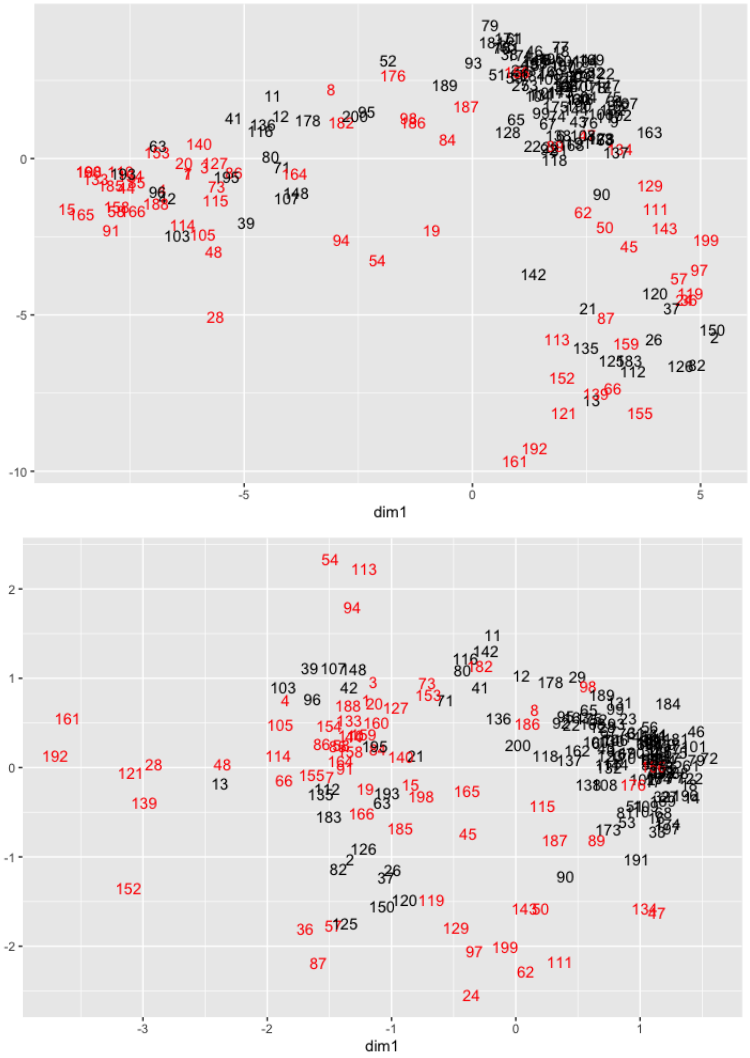

FIGURE 2.13 – PCA on the ECG dataset. Top : on $\mathbf{M}^c$. Bottom : on $\underline{\mathbf{M}}^c$ with a periodic trend ($\tau = 32$).

2.2 Tight Risk Bound for High Dimensional Time Series Completion

As explained in the introduction, low-rank matrix completion methods were studied in depth in the past 10 years. Recall that the first theoretical papers on the topic covered matrix recovery from a few entries observed exactly [67, 68, 80]. The same problem was studied with noisy observations in [65, 66, 81, 77]. The minimax rate of estimation was derived by [17]. Since then, many estimators and many variants of this problem were studied in the statistical literature, see [91, 82, 84, 15, 19, 94, 72, 70, 58, 20, 21] for instance.

High-dimensional time series often have strong correlation, and it is thus natural to assume that the matrix that contains such a series is low-rank (exactly, or approximately). Many econometrics models are designed to generate series with such a structure. For example, the factor model studied in [83, 85, 86, 76, 71, 78] can be interpreted as a high-dimensional autoregressive (AR) process with a low-rank transition matrix. This model (and variants) was used and studied in signal processing [62] and statistics [91, 57].

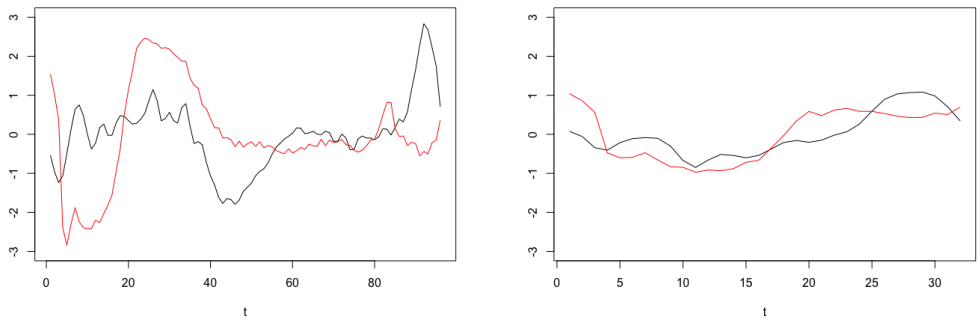

FIGURE 2.14 – The 28th (in black) and the 121th (in red) time series. Left : on $\mathbf{M}^c$. Right : on $\underline{\mathbf{M}}^c$ with a periodic trend ($\tau = 32$).

Other papers focused on a simpler model where the series is represented by a deterministic low-rank trend matrix plus some possibly correlated noise. This model was used by [99] to perform prediction, and studied in [24].
It is thus tempting to use low-rank matrix completion algorithms to recover partially observed high-dimensional time series, and this was indeed done in many applications, as recalled in the introduction. This chapter focuses on low-rank matrix completion for partially observed high-dimensional time series that indeed exhibit a temporal dependence.
The paper is organized as follows. In Section 2.2.1, we recall our model (already defined in the introduction), and the notations used throughout the chapter. In Section 2.2.2, we provide the risk analysis when the rank k is known. We then describe our rank selection procedure in Section 2.2.3 and show that it satisfies a sharp oracle inequality. The numerical experiments are in Section 2.2.4. All the proofs are gathered in Section 2.2.5.

Notations and basic definitions. Throughout the paper, $\mathcal{M}_{d,T}(\mathbb{R})$ is equipped with the Fröbénius scalar product

$$\langle ., . \rangle_{\mathcal{F}} : (\mathbf{A}, \mathbf{B}) \in \mathcal{M}_{d,T}(\mathbb{R})^2 \longmapsto \mathrm{trace}(\mathbf{A}^* \mathbf{B}) = \sum_{j,t} \mathbf{A}_{j,t} \mathbf{B}_{j,t}$$

or with the spectral norm

$$\|.\|_{\mathrm{op}} : \mathbf{A} \in \mathcal{M}_{d,T}(\mathbb{R}) \longmapsto \sup_{\|x\|=1} \|\mathbf{A}x\| = \sigma_1(\mathbf{A}).$$

Let us finally remind the definition of the ϕ-mixing condition on stochastic processes. Given two σ-algebras $\mathcal{A}$ and $\mathcal{B}$, we define the ϕ-mixing coefficient between $\mathcal{A}$ and $\mathcal{B}$ by

$$\phi(\mathcal{A}, \mathcal{B}) := \sup \left\{ |\mathbb{P}(B) - \mathbb{P}(B|A)| \ ; \ (A, B) \in \mathcal{A} \times \mathcal{B}, \ \mathbb{P}(A) \neq 0 \right\}.$$

When $\mathcal{A}$ and $\mathcal{B}$ are independent, $\phi(\mathcal{A}, \mathcal{B}) = 0$, more generally, this coefficient measure how dependent $\mathcal{A}$ and $\mathcal{B}$ are. Given a process $(Z_t)_{t \in \mathbb{N}}$, we define its ϕ-mixing coefficients by

$$\phi_Z(i) := \sup \left\{ \phi(A, B) \ ; \ t \in \mathbb{Z}, \ A \in \sigma(X_h, h \leqslant t), \ B \in \sigma(X_\ell, \ell \geqslant t + i) \right\}.$$

Some properties and examples of applications of ϕ-mixing coefficients can be found in [73].

2.2.1 Setting of the problem and notations

Consider $d, T \in \mathbb{N}^*$ and a $d \times T$ random matrix $\mathbf{M}$. Assume that the rows $\mathbf{M}_{1,.}, \dots, \mathbf{M}_{d,.}$ are time series and that $Y_1, \dots, Y_n$ are $n \in \{1, \dots, d \times T\}$ noisy entries of the matrix $\mathbf{M}$:

$$Y_i = \text{trace}(\mathbf{X}_i^* \mathbf{M}) + \xi_i \ ; \ i \in \{1, \dots, n\}, \tag{2.9}$$

where $\mathbf{X}_1, \dots, \mathbf{X}_n$ are i.i.d random matrices distributed on

$$\mathcal{X} := \{e_{\mathbb{R}^d}(j) e_{\mathbb{R}^T}(t)^* \ ; \ 1 \leqslant j \leqslant d \text{ and } 1 \leqslant t \leqslant T\},$$

and $\xi_1, \dots, \xi_n$ are i.i.d. centered random variables, with standard deviation $\sigma_\xi > 0$, such that $\mathbf{X}_i$ and ξ_i are independent for every $i \in \{1, \dots, n\}$. Note that, as $\mathbf{X}_1, \dots, \mathbf{X}_n$ are independent, we do not exclude multiple observations of the same entry. That is, our model of matrix completion is the one studied in [17, 91, 19] rather than the model in [65, 66] where this is not possible.
Let us now describe the time series structure of each $\mathbf{M}_{1,.}, \dots, \mathbf{M}_{d,.}$. We assume that each series $\mathbf{M}_{j,.}$ can be decomposed as a deterministic component $\mathbf{\Theta}^0_{j,.}$ plus some random noise $\varepsilon_{j,.}$. The noise can exhibit some temporal dependence : $\varepsilon_{j,t}$ will not be independent from $\varepsilon_{j,t'}$ in general. Moreover, as discussed in [24], $\mathbf{\Theta}^0_{j,.}$ can have some more structure : $\mathbf{\Theta}^0_{j,.} = \mathbf{T}^0_{j,.} \mathbf{\Lambda}$ for some known matrix $\mathbf{\Lambda}$. Examples of such structures (smoothness or periodicity) are discussed below. This gives

$$\left\{ \begin{array}{rcl} \mathbf{M} & = & \mathbf{\Theta}^0 + \varepsilon \\ \mathbf{\Theta}^0 & = & \mathbf{T}^0 \mathbf{\Lambda} \end{array} \right. , \tag{2.10}$$

where ε is a $d \times T$ random matrix having i.i.d. and centered rows, $\mathbf{\Lambda} \in \mathcal{M}_{\tau,T}(\mathbb{C})$ $(\tau \leqslant T)$ is known and $\mathbf{T}^0$ is an unknown element of $\mathcal{M}_{d,\tau}(\mathbb{R})$ such that

$$\sup_{j,t} |\mathbf{T}^0_{j,t}| \leqslant \frac{\mathfrak{m}_0}{\mathfrak{m}_{\mathbf{\Lambda}}(\tau)} \text{ with } \mathfrak{m}_0 > 0$$
$$\text{and } 1 \vee \sup_{\mathbf{T} \in \mathcal{M}_{d,\tau}(\mathbb{R})} \left\{ \frac{\sup_{j,t} |(\mathbf{T}\mathbf{\Lambda})_{j,t}|}{\sup_{j,\ell} |\mathbf{T}_{j,\ell}|} \right\} \leqslant \mathfrak{m}_{\mathbf{\Lambda}}(\tau) < \infty. \tag{2.11}$$

Note that this leads to

$$\sup_{j,t} |\mathbf{\Theta}^0_{j,t}| \leqslant \sup_{j,\ell} |\mathbf{T}^0_{j,\ell}| \cdot \frac{\sup_{j,t} |(\mathbf{T}^0 \mathbf{\Lambda})_{j,t}|}{\sup_{j,\ell} |\mathbf{T}^0_{j,\ell}|} \leqslant \mathfrak{m}_0$$

and

$$\mathfrak{m}_{\mathbf{\Lambda}} := \sup_{j,t} |\mathbf{\Lambda}_{j,t}| < \infty.$$

We now make the additional assumption that the deterministic component is low-rank, reflecting the strong correlation between the different series. Precisely, we assume that $\mathbf{T}^0$ is of rank $k \in \{1, \dots, d \wedge T\}$: $\mathbf{T}^0 = \mathbf{U}^0 \mathbf{V}^0$ with $\mathbf{U}^0 \in \mathcal{M}_{d,k}(\mathbb{R})$ and $\mathbf{V}^0 \in \mathcal{M}_{k,\tau}(\mathbb{R})$. The rows of the matrix $\mathbf{V}^0$ may be understood as latent factors. By Equations (2.9) and (2.10), for any $i \in \{1, \dots, n\}$,

$$Y_i = \text{trace}(\mathbf{X}_i^* \mathbf{\Theta}^0) + \overline{\xi}_i \tag{2.12}$$

with $\overline{\xi}_i := \text{trace}(\mathbf{X}_i^* \varepsilon) + \xi_i$. It is reasonable to assume that $\mathbf{X}_i$ and ξ_i, which are random terms related to the observation instrument, are independent to ε, which is the stochastic component of the observed process. Then, since ξ_i is a centered random variable and ε is a centered random matrix,

$$\mathbb{E}(\overline{\xi}_i) = \mathbb{E}(\langle \mathbf{X}_i, \varepsilon \rangle_{\mathcal{F}}) + \mathbb{E}(\xi_i) = \sum_{j=1}^{d} \sum_{t=1}^{T} \mathbb{E}((\mathbf{X}_i)_{j,t}) \mathbb{E}(\varepsilon_{j,t}) = 0.$$

This legitimates to consider the following least-square estimator of the matrix $\mathbf{\Theta}^0$:

$$\left\{ \begin{array}{rcl} \widehat{\mathbf{\Theta}}_{k,\tau} & = & \widehat{\mathbf{T}}_{k,\tau}\mathbf{\Lambda} \\ \widehat{\mathbf{T}}_{k,\tau} & \in & \arg\min_{\mathbf{T}\in\mathcal{S}_{k,\tau}} r_n(\mathbf{T}\mathbf{\Lambda}) \end{array} \right. , \tag{2.13}$$

where $\mathcal{S}_{k,\tau}$ is a subset of

$$\mathcal{M}_{d,k,\tau} := \left\{ \mathbf{UV} \; ; \; (\mathbf{U},\mathbf{V}) \in \mathcal{M}_{d,k}(\mathbb{R}) \times \mathcal{M}_{k,\tau}(\mathbb{R}) \text{ s.t. } \sup_{j,\ell} |\mathbf{U}_{j,\ell}| \leqslant \sqrt{\frac{\mathfrak{m}_0}{k\mathfrak{m}_{\mathbf{\Lambda}}(\tau)}} \text{ and } \sup_{\ell,t} |\mathbf{V}_{\ell,t}| \leqslant \sqrt{\frac{\mathfrak{m}_0}{k\mathfrak{m}_{\mathbf{\Lambda}}(\tau)}} \right\},$$

and

$$r_n(\mathbf{A}) := \frac{1}{n}\sum_{i=1}^{n} (Y_i - \langle \mathbf{X}_i, \mathbf{A}\rangle_{\mathcal{F}})^2 \; ; \; \forall \mathbf{A} \in \mathcal{M}_{d,T}(\mathbb{R}).$$

Remark 2.2.1. — *In many cases, we will simply take $\mathcal{S}_{k,\tau} = \mathcal{M}_{d,k,\tau}$. However, in many applications, it is natural to impose stronger constraints on the estimators. For example, in nonnegative matrix factorization, we would have*

$$\mathcal{S}_{k,\tau} = \{\mathbf{UV} \; ; \; (\mathbf{U},\mathbf{V}) \in \mathcal{M}_{d,k,\tau} \; s.t. \; \forall j,\ell,t, \; \mathbf{U}_{j,\ell} \geqslant 0 \; and \; \mathbf{V}_{\ell,t} \geqslant 0\}$$

(see e.g. [89]). So for now, we only assume that $\mathcal{S}_{k,\tau} \subset \mathcal{M}_{d,k,\tau}$. Later, we will specify some sets $\mathcal{S}_{k,\tau}$.

Let us conclude this section with two examples of matrices $\mathbf{\Lambda}$ corresponding to usual time series structures. On the one hand, if the trend of the multivalued time series $\mathbf{M}$ is τ-periodic, with $T \in \tau\mathbb{N}^*$, one can take $\mathbf{\Lambda} = (\mathbf{I}_\tau|\cdots|\mathbf{I}_\tau)$, and then $\mathfrak{m}_{\mathbf{\Lambda}} = 1$ and $\mathfrak{m}_{\mathbf{\Lambda}}(\tau) := 1$ works. So, in this case, note that the usual matrix completion model of [17] is part of our framework by taking $T = \tau$. On the other hand, assume that for any $j \in \{1,\ldots,d\}$, the trend of $\mathbf{M}_{j,\cdot}$ is a sample on $\{0, 1/T, 2/T, \ldots, 1\}$ of a function $f_j : [0,1] \to \mathbb{R}$ belonging to a Hilbert space $\mathcal{H}$. In this case, if $(\mathbf{e}_n)_{n\in\mathbb{Z}}$ is a Hilbert basis of $\mathcal{H}$, one can take $\mathbf{\Lambda} = (\mathbf{e}_n(t/T))_{|n|\leqslant N, 1\leqslant t\leqslant T}$. For instance, if $f_j \in \mathbb{L}^2([0,1];\mathbb{R})$, a natural choice is the Fourier basis $\mathbf{e}_n(t) = e^{2i\pi nt/T}$, and then $\mathfrak{m}_{\mathbf{\Lambda}} = 1$ and

$$\frac{\sup_{j,t}|(\mathbf{T}\mathbf{\Lambda})_{j,t}|}{\sup_{j,\ell}|\mathbf{T}_{j,\ell}|} \leqslant \frac{1}{\sup_{j,\ell}|\mathbf{T}_{j,\ell}|} \cdot \sup_{j,t} \sum_{\ell=1}^{\tau} |\mathbf{T}_{j,\ell} e^{2i\pi nt/T}| \leqslant \tau =: \mathfrak{m}_{\mathbf{\Lambda}}(\tau).$$

Here, the usual matrix completion model of [17] is not part of our framework because T is possibly huge and to take $\tau = T$ implies that the coefficients of the matrix $\mathbf{T}^0$ are all unrealistically small by Condition (2.11). However, wathever the time series structure taken into account via $\mathbf{\Lambda}$, our model is designed for small values of τ. Else, the model of [17] is appropriate. So, when $\mathbf{\Lambda}$ is the previous *Fourier matrix*, and in general when $\mathfrak{m}_{\mathbf{\Lambda}}(\tau)$ is a non constant increasing function of τ, we assume that $\tau \in [\![1, \tau_0]\!]$ with $\tau_0 \ll T$.

2.2.2 Risk bound on $\widehat{\mathbf{T}}_{k,\tau}$

2.2.2.1 Upper bound

First of all, since $\mathbf{X}_1,\ldots,\mathbf{X}_n$ are i.i.d $\mathcal{X}$-valued random matrices, there exists a probability measure Π on $\mathcal{X}$ such that

$$\mathbb{P}_{\mathbf{X}_i} = \Pi \; ; \; \forall i \in \{1,\ldots,n\}.$$

In addition to the two norms on $\mathcal{M}_{d,T}(\mathbb{R})$ introduced above, let us consider the scalar product $\langle .,. \rangle_{\mathcal{F},\Pi}$ defined on $\mathcal{M}_{d,T}(\mathbb{R})$ by

$$\langle \mathbf{A}, \mathbf{B} \rangle_{\mathcal{F},\Pi} := \int_{\mathcal{M}_{d,T}(\mathbb{R})} \langle X, \mathbf{A} \rangle_{\mathcal{F}} \langle X, \mathbf{B} \rangle_{\mathcal{F}} \Pi(dX) \ ; \forall \mathbf{A}, \mathbf{B} \in \mathcal{M}_{d,T}(\mathbb{R}).$$

Remarks :

(1) For any deterministic $d \times T$ matrices $\mathbf{A}$ and $\mathbf{B}$,

$$\langle \mathbf{A}, \mathbf{B} \rangle_{\mathcal{F},\Pi} = \mathbb{E}(\langle \mathbf{A}, \mathbf{B} \rangle_n)$$

where $\langle .,. \rangle_n$ is the empirical scalar product on $\mathcal{M}_{d,T}(\mathbb{R})$ defined by

$$\langle \mathbf{A}, \mathbf{B} \rangle_n := \frac{1}{n} \sum_{i=1}^{n} \langle \mathbf{X}_i, \mathbf{A} \rangle_{\mathcal{F}} \langle \mathbf{X}_i, \mathbf{B} \rangle_{\mathcal{F}}.$$

However, note that this relationship between $\langle .,. \rangle_{\mathcal{F},\Pi}$ and $\langle .,. \rangle_n$ doesn't hold anymore when $\mathbf{A}$ and $\mathbf{B}$ are random matrices.

(2) If the sampling distribution Π is uniform, then $\|.\|^2_{\mathcal{F},\Pi} = (dT)^{-1} \|.\|^2_{\mathcal{F}}$.

Notation. For every $i \in \{1, \dots, n\}$, let χ_i be the couple of *coordinates* of the nonzero element of $\mathbf{X}_i$, which is a $\mathcal{E}$-valued random variable with $\mathcal{E} = \{1, \dots, d\} \times \{1, \dots, T\}$.

In the sequel, ε, $\xi_1, \dots, \xi_n$ and $\mathbf{X}_1, \dots, \mathbf{X}_n$ fulfill the following additional conditions.

Assumption 2.2.2. — *The rows of ε are independent and identically distributed. There is a process $(\varepsilon_t)_{t \in \mathbb{Z}}$ such that each $\varepsilon_{j,\cdot}$ has the same distribution than $(\varepsilon_1, \dots, \varepsilon_T)$, and such that*

$$\Phi_\varepsilon := 1 + \sum_{i=1}^{n} \phi_\varepsilon(i)^{1/2} < \infty.$$

Assumption 2.2.3. — *There exists a deterministic constant $\mathfrak{m}_\varepsilon > 0$ such that*

$$\sup_{j,t} |\varepsilon_{j,t}| \leqslant \mathfrak{m}_\varepsilon.$$

Moreover, there exist two deterministic constants $\mathfrak{c}_\xi, \mathfrak{v}_\xi > 0$ such that

$$\sup_{i \in \{1,\dots,n\}} \mathbb{E}(\xi_i^2) \leqslant \mathfrak{v}_\xi$$

and, for every $q \geqslant 3$,

$$\sup_{i \in \{1,\dots,n\}} \mathbb{E}(|\xi_i|^q) \leqslant \frac{\mathfrak{v}_\xi \mathfrak{c}_\xi^{q-2} q!}{2}.$$

This assumption means that the $\varepsilon_{j,t}$'s are bounded, and that the ξ_i's are sub-exponential random variables. Sub-exponential variables include bounded and Gaussian variables as special cases. Note that this is the assumption made on the noise for the matrix completion in the i.i.d. framework in the papers mentioned above [19, 17]. The boundedness of the $\varepsilon_{j,t}$'s can be seen as quite restrictive. However, we are not aware of any way to avoid this assumption in this setting. Indeed, it allows to apply Samson's concentration inequality for ϕ-mixing processes (see Samson [49]). In [23], the authors prove sharp sparsity inequalities under a similar assumption, using Samson's inequality. They also show that the other concentration inequalities known for time series lead to slow rates of convergence.

Assumption 2.2.4. — *There is a constant* $\mathfrak{c}_\Pi > 0$ *such that*

$$\Pi(\{e_{\mathbb{R}^d}(j)e_{\mathbb{R}^T}(t)^*\}) \leqslant \frac{\mathfrak{c}_\Pi}{dT} \; ; \forall (j,t) \in \mathcal{E}.$$

Note that when the sampling distribution Π is uniform, Assumption 2.2.4 is trivially satisfied with $\mathfrak{c}_\Pi = 1$.

Theorem 2.2.5. — *Let* $\alpha \in (0,1)$. *Under Assumptions 2.2.2, 2.2.3 and 2.2.4, if* $n \geqslant \max(d,\tau)$, *then*

$$\|\widehat{\mathbf{\Theta}}_{k,\tau} - \mathbf{\Theta}^0\|^2_{\mathcal{F},\Pi} \leqslant 3 \min_{\mathbf{T}\in\mathcal{S}_{k,\tau}} \|(\mathbf{T}-\mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} + \mathfrak{c}_{2.2.5}\left[k(d+\tau)\frac{\log(n)}{n} + \frac{1}{n}\log\left(\frac{4}{\alpha}\right)\right]$$

with probability larger than $1-\alpha$, *where* $\mathfrak{c}_{2.2.5}$ *is a constant depending only on* $\mathfrak{m}_0$, $\mathfrak{v}_\xi$, $\mathfrak{c}_\xi$, $\mathfrak{m}_\varepsilon$, $\mathfrak{m}_{\mathbf{\Lambda}}$, Φ_ε *and* $\mathfrak{c}_\Pi$.

Actually, from the proof of the theorem, we know $\mathfrak{c}_{2.2.5}$ explicitly. Indeed,

$$\mathfrak{c}_{2.2.5} = 72\mathfrak{m}_0\mathfrak{m}_{\mathbf{\Lambda}}\mathfrak{c}_\xi + 5\mathfrak{c}_{2.2.11,1} + 9\mathfrak{m}_0\mathfrak{c}_{2.2.11,2}$$

where $\mathfrak{c}_{2.2.11,1}$ and $\mathfrak{c}_{2.2.11,2}$ are constants (explicitly given in Theorem 2.2.11 in Section 2.2.5) depending themselves only on $\mathfrak{m}_0$, $\mathfrak{v}_\xi$, $\mathfrak{c}_\xi$, $\mathfrak{m}_\varepsilon$, $\mathfrak{m}_{\mathbf{\Lambda}}$, Φ_ε and $\mathfrak{c}_\Pi$.

Remarks :

(1) Note that another classic way to formulate the risk bound in Theorem 2.2.5 is that for every $s > 0$, with probability larger than $1 - e^{-s}$,

$$\|\widehat{\mathbf{\Theta}}_{k,\tau} - \mathbf{\Theta}^0\|^2_{\mathcal{F},\Pi} \leqslant 3 \min_{\mathbf{T}\in\mathcal{S}_{k,\tau}} \|(\mathbf{T}-\mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} + \bar{\mathfrak{c}}_{2.2.5}\left[k(d+\tau)\frac{\log(n)}{n} + \frac{s}{n}\right].$$

(2) The ϕ-mixing assumption (Assumption 2.2.2) is known to be restrictive, we refer the reader to [73] where it is compared to other mixing conditions. Some examples are provided in Examples 7, 8 and 9 in [23], including stationary AR processes with a noise that has a density with respect to the Lebesgue measure on a compact interval. Interestingly, [23] also discusses weaker notions of dependence. Under these conditions, we could here apply the inequalities used in [23], but it is important to note that this would prevent us from taking λ of the order of n in the proof of Proposition 2.2.8. In other words, this would deteriorate the rates of convergence. A complete study of all the possible dependence conditions on ε goes beyond the scope of this paper.

2.2.2.2 Lower bound

In the case where $T = \tau$ and $\mathbf{\Lambda} = \mathbf{I}_T$, the model in [17] is included in our model, and corresponds to the case where the temporally dependent noise ε is null : $\varepsilon = 0$. This means that the lower bound provided by Theorem 5 in [17] holds in our setting. That is, when the ξ_i's are $\mathcal{N}(0,1)$ and the $\mathbf{X}_i$'s are uniform (so $\mathfrak{c}_\Pi = 1$), there are absolute constants $\mathfrak{c}_{\inf}, \beta > 0$ such that for any $k \leqslant \frac{n}{d\vee T}$,

$$\inf_{\widehat{\mathbf{A}}} \sup_{\mathbf{\Theta}^0\in\mathcal{M}_{d,k,T}} \mathbb{P}\left(\|\widehat{\mathbf{A}} - \mathbf{\Theta}^0\|^2_{\mathcal{F},\Pi} \geqslant \mathfrak{c}_{\inf}\frac{k(d+T)}{n}\right) \geqslant 1-\beta.$$

In other words, the bound in Theorem 2.2.5 is tight, maybe up to the $\log(n)$ term (there is also a log term in the upper bounds of [17]). We now extend this result to the case $\tau \leqslant T$, in the special case where the deterministic component of the series is τ-periodic : $\mathbf{\Lambda} = (\mathbf{I}_\tau | \dots | \mathbf{I}_\tau)$.

Theorem 2.2.6. — *Assume the ξ_i's are $\mathcal{N}(0,1)$, the $\mathbf{X}_i$'s are uniform (so $\mathfrak{c}_\Pi = 1$) and the temporally dependent noise $\varepsilon = 0$. There are absolute constants $\mathfrak{c}_{\inf}, \beta > 0$ such that for any $\tau \in \{1, \dots, T\}$, in the case $\mathbf{\Lambda} = (\mathbf{I}_\tau | \dots | \mathbf{I}_\tau)$, for any $k \leqslant (256\mathfrak{m}_0^2 n/(d \vee \tau))^{1/3}$,*

$$\inf_{\widehat{\mathbf{A}}} \sup_{\mathbf{\Theta}^0 \in \mathcal{M}_{d,k,\tau}} \mathbb{P}\left(\|\widehat{\mathbf{A}} - \mathbf{\Theta}^0 \mathbf{\Lambda}\|_{\mathcal{F},\Pi}^2 \geqslant \mathfrak{c}_{\inf} \frac{k(d+\tau)}{n} \right) \geqslant 1 - \beta.$$

For $\tau = T$, we recover Theorem 5 in [17], but our result also guarantees that the bound in Theorem 2.2.5 is tight (up to log terms) even when $\tau < T$.

2.2.3 Model selection

The purpose of this section is to provide a selection method of the parameter k. First, for the sake of readability, $\mathcal{S}_{k,\tau}$ and $\widehat{\mathbf{T}}_{k,\tau}$ are respectively denoted by $\mathcal{S}_k$ and $\widehat{\mathbf{T}}_k$ in the sequel. The adaptive estimator studied here is $\widehat{\mathbf{\Theta}} := \widehat{\mathbf{T}}\mathbf{\Lambda}$, where $\widehat{\mathbf{T}} := \widehat{\mathbf{T}}_{\widehat{k}}$,

$$\widehat{k} \in \arg\min_{k \in \mathcal{K}} \{ r_n(\widehat{\mathbf{T}}_k \mathbf{\Lambda}) + \mathrm{pen}(k) \} \text{ with } \mathcal{K} = \{1, \dots, k^*\} \subset \mathbb{N}^*,$$

and

$$\mathrm{pen}(k) := 16 \mathfrak{c}_{\mathrm{pen}} \frac{\log(n)}{n} k(d+\tau) \text{ with } \mathfrak{c}_{\mathrm{pen}} = 2\left(\frac{1}{\mathfrak{c}_{2.2.8}} \wedge \lambda^* \right)^{-1}.$$

Note that the value of the constant $\mathfrak{c}_{\mathrm{pen}}$ could be deduced from the proofs. It would however depend on quantities that are unknown in practice, such as $\mathfrak{c}_\Pi$ or Φ_ε. Moreover, the value of $\mathfrak{c}_{\mathrm{pen}}$ provided by the proofs would probably be too large for practical purposes. In practice, we recommend to use the slope heuristics to estimate this constant. The slope heuristic is defined as follows : for each $C > 0$, let us define

$$k(C) \in \arg\min_{k \in \mathcal{K}} \{ r_n(\widehat{\mathbf{T}}_k \mathbf{\Lambda}) + C \cdot k \}.$$

Then, let us define $\widetilde{C}$ as the location of the largest jump of the function

$$C \longmapsto r_n(\widehat{\mathbf{T}}_{k(C)} \mathbf{\Lambda})$$

and choose the rank $\widetilde{k} = k(2\widetilde{C})$. This popular procedure leads to good practical results in most situations. Its theoretical properties are available only in limited situations (see [60]), though, so we will focus our theoretical result to $\widehat{k}$.

Theorem 2.2.7. — *Under Assumptions 2.2.2, 2.2.3 and 2.2.4, if $n \geqslant \max(d, \tau)$, then*

$$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^0\|_{\mathcal{F},\Pi}^2 \leqslant 4 \min_{k \in \mathcal{K}} \left\{ 3 \min_{\mathbf{T} \in \mathcal{S}_k} \|(\mathbf{T} - \mathbf{T}^0)\mathbf{\Lambda}\|_{\mathcal{F},\Pi}^2 + \mathfrak{c}_{2.2.7,1} k(d+\tau) \frac{\log(n)}{n} \right\} + \frac{\mathfrak{c}_{2.2.7,1}}{n} \log\left(\frac{4k^*}{\alpha} \right) + \frac{\mathfrak{c}_{2.2.7,2}}{n}$$

with probability larger than $1 - \alpha$, where

$$\mathfrak{c}_{2.2.7,1} = 4\mathfrak{c}_{2.2.5} + 16\mathfrak{c}_{\mathrm{pen}} + 72\mathfrak{m}_0\mathfrak{c}_\xi \text{ and } \mathfrak{c}_{2.2.7,2} = 9\mathfrak{c}_{2.2.11,2}\mathfrak{m}_0.$$

2.2.4 Numerical experiments

This section describes an experimental study of the estimator of the matrix $\mathbf{T}^0$ introduced at Section 2.2.1. In particular, we compare on simulated periodic data the completion procedure using the periodicity information, to the standard procedure, and we observe a clear improvement. We also illustrate our results on real data from Paris sharing bike system.
In the case where no particular temporal structure is used, that is, $\mathbf{\Lambda} = \mathbf{I}_T$, standard packages such as softImpute [128] could be used. However, this is not necessarily the case for a general Λ, thus we implemented a standard alternate least square (ALS) procedure. That is, we iterate $\mathbf{U} := \arg\min_U \mathbf{r_n}(U\mathbf{V}\mathbf{\Lambda})$ and $\mathbf{V} := \arg\min_V \mathbf{r_n}(\mathbf{U}V\mathbf{\Lambda})$ until convergence. Each step is a linear regression and has an explicit solution. Despite its extreme simplicity, this type of alternate optimization is known to lead to very good results in practice [87], and such a method is actually used by softImpute [128].
The code of all the experiments can be found on the third author webpage
https ://amelierosier8.wixsite.com/website

2.2.4.1 Experiments on simulated datas

The experiments in this subsection are done on datas simulated the following way :

(1) We generate a matrix $\mathbf{T}^0 = \mathbf{U}^0\mathbf{V}^0$ with $\mathbf{U}^0 \in \mathcal{M}_{d,k}(\mathbb{R})$ and $\mathbf{V}^0 \in \mathcal{M}_{k,\tau}(\mathbb{R})$. Each entries of $\mathbf{U}^0$ and $\mathbf{V}^0$ are generated independently by simulating i.i.d. $\mathcal{N}(0,1)$ random variables.

(2) We multiply $\mathbf{T}^0$ by a known matrix $\mathbf{\Lambda} \in \mathcal{M}_{\tau,T}(\mathbb{R})$. This matrix depends on the time series structure assumed on $\mathbf{M}$. Here, we consider the periodic case : $T = p\tau$, $p \in \mathbb{N}^*$ and $\mathbf{\Lambda} = (\mathbf{I}_\tau | \dots | \mathbf{I}_\tau)$.

(3) The matrix $\mathbf{M}$ is then obtained by adding a matrix ε such that $\varepsilon_{1,.}, \dots, \varepsilon_{d,.}$ are generated independently by simulating i.i.d. AR(1) processes with compactly supported error in order to meet the *ϕ-mixing* condition. We multiply ε by the coefficient σ_ε which value will vary according to the experiments. The goal is to evaluate the impact of adding more noise in the estimation.
Only 30% of the entries of $\mathbf{M}$, taken randomly, are observed. These entries are then corrupted by i.i.d. observation errors $\xi_1, \dots, \xi_n \rightsquigarrow \mathcal{N}(0, 0.01^2)$. To meet Assumption 2.2.3, we also consider uniform errors $\xi_1, \dots, \xi_n \rightsquigarrow \mathcal{U}([-a, a])$, where $a = \sqrt{3}/100 \approx 0.017$ to keep the same variance than previously. The first experiments will show that the estimation remains quite good even if the ξ_i's are not bounded.

Given the observed entries, our goal is to complete the missing values of the matrix and check if they correspond to the simulated data in two different cases :

(1) Our first model doesn't take into account the time series structure in the matrix $\mathbf{M}$. Thus, we simply apply our fonction `als` to the dataframe containing the values of the noisy entries in addition to their position in the matrix $\mathbf{M}$ (number of the line j and number of the column t with $1 \leqslant j \leqslant d$, $1 \leqslant t \leqslant T$). The output of the function gives directly an estimator of the matrix $\mathbf{\Theta}^0$.

(2) Our second model takes into account the time series structure in $\mathbf{M}$ and more precisely the periodicity of the time series datas. In order to have an estimator of the matrix $\mathbf{\Theta}^0$, some transformation are required on the data : the fonction `als`

is now applied to the dataframe in which all the observed entries at the position (j,t) $(1 \leqslant j \leqslant d,\ 1 \leqslant t \leqslant T)$ are now moved to the position $(j, t[\mathrm{mod}]\tau)$. The output of this function needs to be remultiplied by $\mathbf{\Lambda}$ to have an estimator of $\mathbf{\Theta}^0$.

We will evaluate the MSE of the estimator with respect to several parameters and show that there is a gain to take into account the time series structure in the model. As expected, the more $\mathbf{\Theta}^0$ is perturbed, either with ε or $\xi_1, \ldots, \xi_n$, the more difficult it is to reconstruct the matrix. In the same way, increasing the value of the rank k will lead to a worse estimation. Finally, we study the effect of replacing the uniform error in each AR(1) by a Gaussian one.

The first experiments are done with $d = 1000$, $T = 100$ and $\tau = 25$ to be in concordance with the experiments on real data (see subsection 5.3). Here are the MSE obtained for both models, 3 values of the rank k and for two kinds of observation errors $\xi_1, \ldots, \xi_n$: Gaussian $\mathcal{N}(0, 0.01^2)$ v.s. uniform $\mathcal{U}([-0.017, 0.017])$. The errors in the AR(1) processes generating the rows of ε remain uniform $\mathcal{U}([-1, 1])$.

MSE	$\xi_i \rightsquigarrow \mathcal{N}(0, 0.01^2)$	$\xi_i \rightsquigarrow \mathcal{U}([-0.017, 0.017])$
Model w/o time series struct.	0.00012	0.00014
Model with time series struct.	0.00009	0.00010

TABLE 2.3 – MSE for both models, $k = 2$.

MSE	$\xi_i \rightsquigarrow \mathcal{N}(0, 0.01^2)$	$\xi_i \rightsquigarrow \mathcal{U}([-0.017, 0.017])$
Model w/o time series struct.	0.00018	0.00022
Model with time series struct.	0.00012	0.00013

TABLE 2.4 – MSE for both models, $k = 5$.

MSE	$\xi_i \rightsquigarrow \mathcal{N}(0, 0.01^2)$	$\xi_i \rightsquigarrow \mathcal{U}([-0.017, 0.017])$
Model w/o time series struct.	0.00026	0.00045
Model with time series struct.	0.00013	0.00017

TABLE 2.5 – MSE for both models, $k = 8$.

Thus, both of the rank k and the nature of the error considered for the ξ_i's seem to play a key role on the reduction of the MSE. Regarding the rank k (d, T and τ being fixed) being fixed), our numerical results are consistent with respect to the theoretical rate of convergence of order $O(k(d+\tau)\log(n)/n)$ obtained at Theorem 2.2.5 when we consider the time series structure of the data (see Tables 2.3, 2.4 and 2.5). Indeed, for both models, the MSE is increasing when the value of the rank k is higher but this increase is always more significant in the model without time series structure, which is also consistent with the rate of convergence of order $O(k(d+T)\log(n)/n)$ obtained in this case. Note that when we look at one model at a time, for each tested values of k, whatever the distribution of the errors $\xi_1, \ldots, \xi_n$ (Gaussian or uniform), the MSE

remains of same order with a slight improvement when we considered Gaussian errors. This justifies to take $\xi_1, \ldots, \xi_n \rightsquigarrow \mathcal{N}(0, 0.01^2)$ in the following experiments.

This study can be summarized in the following experiment which shows the evolution of the MSE with respect to the rank k ($k = 1, \ldots, 10$) for both models. Once again, we take $d = 1000$, $T = 100$, $\tau = 25$ but the ξ_i's remain i.i.d. $\mathcal{N}(0, 0.01^2)$ random variables, and $\varepsilon_{1,.}, \ldots, \varepsilon_{d,.}$ are i.i.d. AR(1) processes with Gaussian errors.

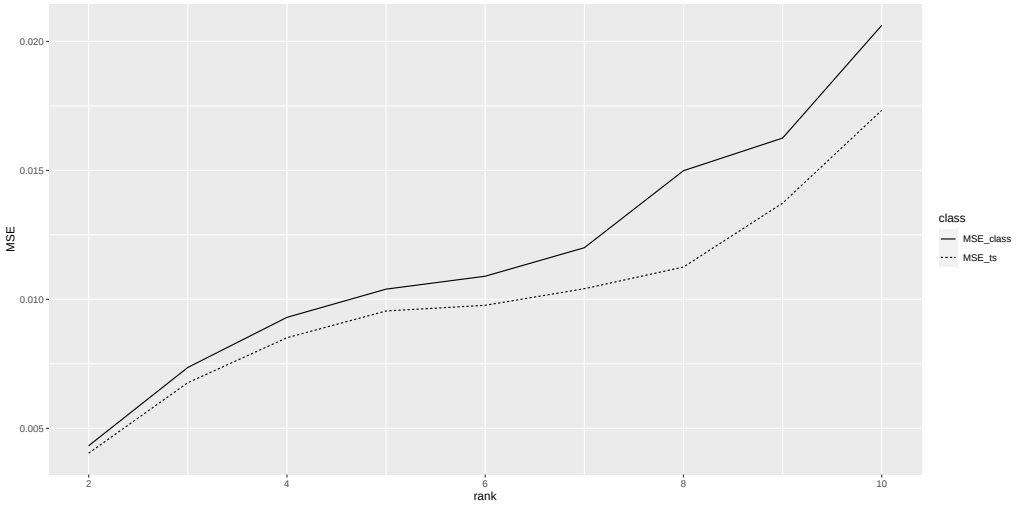

FIGURE 2.15 – Models (time series (dotted line) v.s. classic (solid line)) MSEs with respect to the rank k.

As expected (see Figure 2.15), the MSE is much better with the model taking into account the time series structure. The MSE in both cases degrades when the value of the rank is increasing, the maximum being reached for $k = 10$ with the value 0.0173 for the time series model compared to 0.0206 in the classic case, which still remains very low.

As we said, the estimation seems to be more precise with Gaussian errors in ε, and the more $\mathbf{\Theta}^0$ is perturbed via ε or $\xi_1, \ldots, \xi_n$, the more the completion process is complicated and the MSE degrades. So, we now evaluate the consequence on the MSE of changing the value of σ_ε. For both models (with or without taking into account the time series structure), the following figure shows the evolution of the MSE with respect to σ_ε when the errors in ε are $\mathcal{N}(0, 1/3)$ random variables and all the other parameters remain the same than previously, we are still considering 30% of observed entries.
Once again, as expected (see Figure 2.16), the MSE with time series model is smaller than the one with the classic model for each values of σ_ε. The fact that the MSE increases with respect to σ_ε with both models illustrates that *more noise* always complicates the completion process. In our experiments, the values of σ_ε range from 0.02 to 2. We can notice that, the more we add noise with σ_ε, the more significant the gap between the MSE of both models is. With σ_ε equal to 2, the MSE reaches the value 0.2241 for the time series model and 0.3040 for the classic one. Our method has increasing difficulty in reconstructing the matrix when we add too much noise to the model. See also Table 2.6.
Let us do the same experiment but with uniform $\mathcal{U}([-1, 1])$ errors in the AR(1) processes generating the rows of ε.

The curves shape on Figure 2.17 is pretty much the same as in the previous graph : the

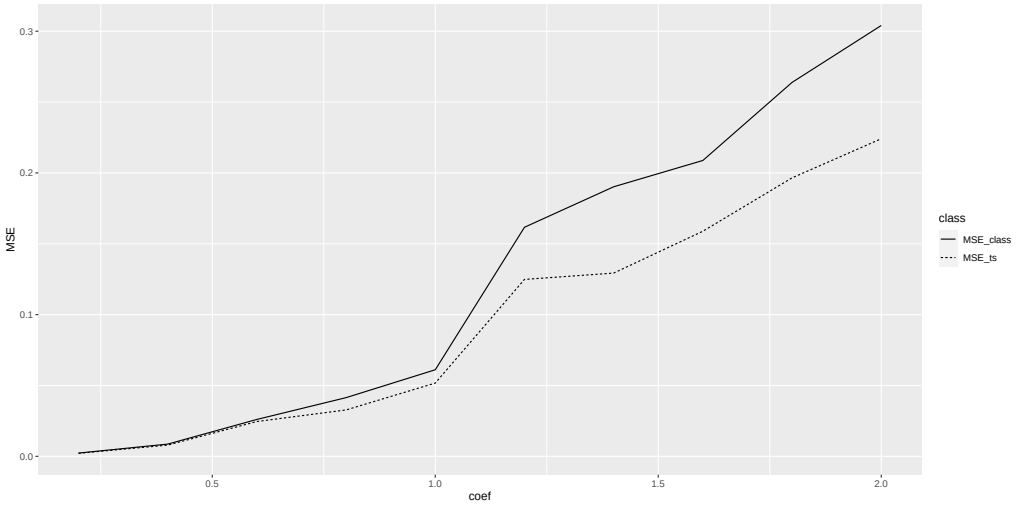

FIGURE 2.16 – Models (time series (dotted line) v.s. classic (solid line)) MSEs with respect to σ_ε, Gaussian errors.

	Min. MSE	**Max. MSE**
Model w/o time series struct.	0.0023	0.3040
Model with time series struct.	0.0021	0.2241

TABLE 2.6 – Min. and max. values reached by the MSE with Gaussian errors in ε.

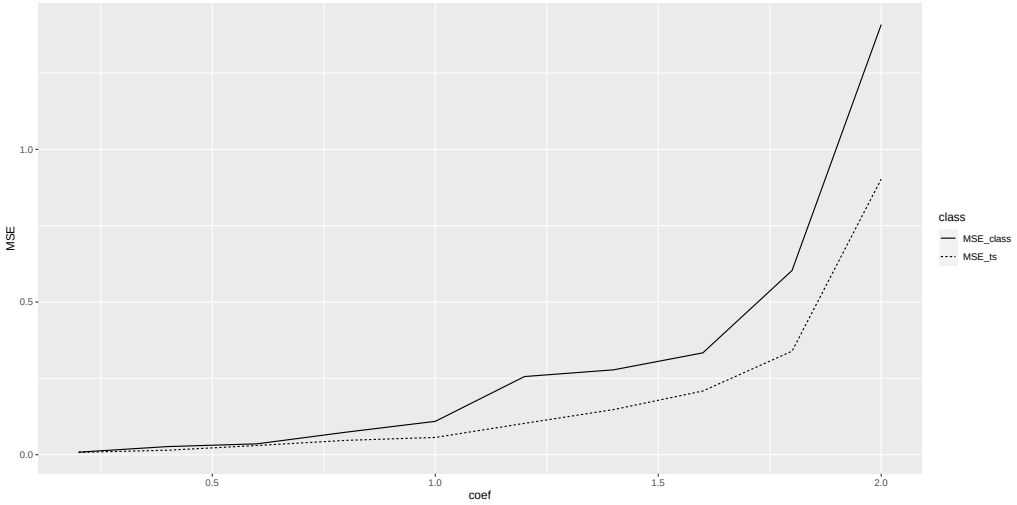

FIGURE 2.17 – Models (time series (dotted line) v.s. classic (solid line)) MSEs with respect to σ_ε, uniform errors.

MSE for the model taking into account the time series structure is still smaller than for the classic model and this difference between the two models is even greater when we increase the value of σ_ε. However, this time, the MSE for both models reaches higher values, leading to a huge misestimation when $\sigma_\varepsilon = 2$ (see Table 2.7).

	Min. MSE	Max. MSE
Model w/o time series struct.	0.0082	1.4088
Model with time series struct.	0.0076	0.9027

TABLE 2.7 – Min. and max. values reached by the MSE with uniform errors in ε.

Finally, as mentioned, the previous numerical experiments were done by assuming that k is known, which is mostly uncommon in practice. So, our purpose in the last part of this section is to implement the model selection method introduced at Section 2.2.3. Let us recall the criterion to minimize :

$$\left\{ \begin{array}{rcl} \text{crit}(k) & = & r_n(\widehat{\mathbf{T}}_k\mathbf{\Lambda}) + \text{pen}(k) \\ \text{pen}(k) & = & \mathfrak{c}_{\text{cal}} k(d+\tau)\log(n)/n \end{array} \right. ; k \in \{1,\ldots,20\}.$$

In the sequel, $\xi_1,\ldots,\xi_n \rightsquigarrow \mathcal{N}(0,0.5)$, $\varepsilon_{1,.},\ldots,\varepsilon_{d,.}$ are i.i.d. AR(1) processes with $\mathcal{N}(0,1/3)$ errors, and $\sigma_\varepsilon = 0.2$. Percentage of observed entries is still 30%. The penalty term in crit(.) depends on the constant $\mathfrak{c}_{\text{cal}} > 0$ which is calibrated here by using the slope heuristic presented at Section 2.2.3.
On 20 independent experiments, Table 2.8 gives the mean MSE obtained for the estimator computed with the true rank $k = 5$ and the associated adaptative estimator computed with $\widehat{k}$ selected by minimizing the criterion studied in Section 2.2.3. Table 2.9

Mean MSE for $\widehat{\mathbf{T}}_k\mathbf{\Lambda}$	0.10712
Mean MSE for $\widehat{\mathbf{T}}_{\widehat{k}}\mathbf{\Lambda}$	0.17601

TABLE 2.8 – Mean MSE over 20 simulations for $\widehat{\mathbf{T}}_k\mathbf{\Lambda}$ and $\widehat{\mathbf{T}}_{\widehat{k}}\mathbf{\Lambda}$.

gives the frequence of the different values of k selected. Our method select the true k 8 times over 20.

k selected	4	5	6	7	8	9
Frequence	0.05	0.4	0.1	0.15	0.2	0.1

TABLE 2.9 – Frequence of k-values selected

2.2.4.2 Experiments on real datas

Modern transportation data are often high-dimensional and have strong patterns including periodicity. For this reason, matrix factorization methods are very popular in this field [69, 95]. The data used in this section comes from the `funFEM` package (the real time data are available at *https ://developer.jcdecaux.com/*). We used the *Velib* data set which contains data from the bike sharing system of Paris. These data provide the occupancy (number of available bikes/number of bike docks) of 1189 bike stations over one week. The data were collected every hour during the following period : Sunday 1st

Sept. - Sunday 7th Sept., 2014. We removed the time points collected during the week-end (50 time points in total) insofar as the week-end occupancy of the bike stations differs from the week. Loading profiles of 6 different stations (week-end excluded) are represented on Figure 2.18.

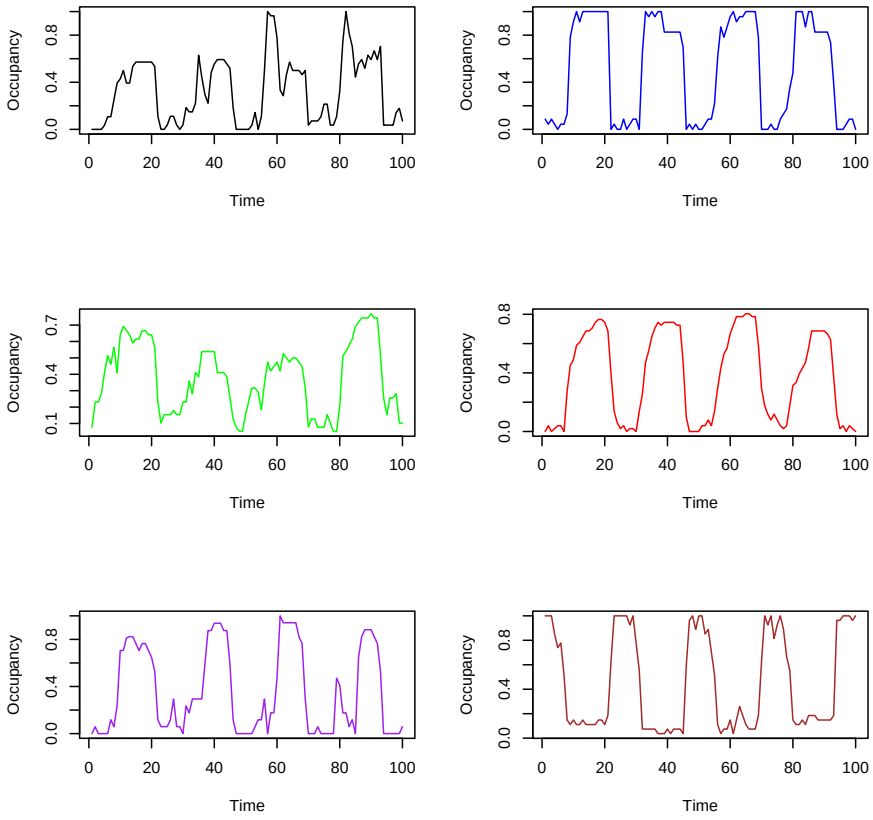

FIGURE 2.18 – Occupancy of six *Velib* stations over one week (week-end excluded).

We clearly notice the daily periodic behaviour of our time series. Thus, the experiments of this section are done with the real time data in the matrix $\mathbf{M}$ of dimensions $d = 1189$, $\tau = 25$ (which corresponds to one day) and $T = 125$ (five days, from Monday to Thursday). Once again, we evaluate the MSE of the estimator with and without taking into account the time series structure, that is the periodicity in this case. Different percentages of the entries observed are tested. As for the simulated data, for the model without considering the temporal structure of our series, we apply directly our function `als` on the dataframe containing the observed entries with their position in the matrix, without any additional transformation on the data. The output gives directly an estimator of $\mathbf{M}$. As regards the model considering the periodic behaviour of the *Velib* time series in $\mathbf{M}$, the ALS optimization procedure is applied on the dataframe which has received the same transformation than the one explained at point (2) in the previous section. Once again, the output needs to be multiplied by $\mathbf{\Lambda}$ to have an estimator of $\mathbf{M}$ at the end. The MSEs obtained for both models are gathered in Table 2.10. We study how the MSEs vary according to the percentage of observed entries.

Of course, the real data is not *exactly* periodic (as can bee seen in some of the series in Figure 2.18. This means that the bias term of in Theorem 2.2.5, is larger

	15%	30%
Model w/o time series struct.	0.0609	0.0315
Model with time series struct.	0.0436	0.0381

TABLE 2.10 – MSE according to the number of observed entries (%).

for the method imposing periodicity than for the standard method : $\min_{\mathbf{T}\in\mathcal{S}_{k,\tau}} \|(\mathbf{T}-\mathbf{T}^0)\mathbf{\Lambda}\|_{\mathcal{F},\Pi} \geq \min_{\mathbf{T}\in\mathcal{S}_{k,T}} \|\mathbf{T}-\mathbf{T}^0\|_{\mathcal{F},\Pi}$. On the other hand, the variance term of the method using periodicity is much smaller : $k(d+\tau)/n \leq k(d+T)/n$. Thus, it is expected that when the sample size n is small, using periodicity can improve on the standard method, but that this is not the case for larger values of n. This is perfectly illustrated by our experiments : Table 2.10 show that when we observe 15% of the original data, exploiting periodicity improves on the reconstruction of the data by the standard method by more than 25%. On the other hand, when we the sample size doubles, the standard method already performs slightly better.

2.2.5 Proofs

This section is organized as follows. We first state an exponential inequality that will serve as a basis for all the proofs. From this inequality, we prove Theorem 2.2.11, a prototype of Theorem 2.2.5 that holds when the set $\mathcal{S}_{k,\tau}$ is finite or infinite but compact by using ϵ-nets ($\epsilon > 0$). In the proof of Theorem 2.2.5, we provide an explicit risk-bound by using the ϵ-net $\mathcal{S}^{\epsilon}_{k,\tau}$ of $\mathcal{S}_{k,\tau}$ constructed in Candès and Plan [66], Lemma 3.1.

2.2.5.1 Exponential inequality

This sections deals with the proof of the following exponential inequality, the cornerstone of the paper, which is derived from the usual Bernstein inequality and its extension to ϕ-mixing processes due to Samson [49].

Proposition 2.2.8. — *Let $\mathbf{T} \in S_{k,\tau}$. Under Assumptions 2.2.2, 2.2.3 and 2.2.4,*

$$\mathbb{E}\left[\exp\left(\frac{\lambda}{4}\left(\left(1+\mathfrak{c}_{2.2.8}\frac{\lambda}{n}\right)(R(\mathbf{T}^0\mathbf{\Lambda})-R(\mathbf{T}\mathbf{\Lambda}))+r_n(\mathbf{T}\mathbf{\Lambda})-r_n(\mathbf{T}^0\mathbf{\Lambda})\right)\right)\right] \leqslant 1 \quad (2.14)$$

and

$$\mathbb{E}\left[\exp\left(\frac{\lambda}{4}\left(\left(1-\mathfrak{c}_{2.2.8}\frac{\lambda}{n}\right)(R(\mathbf{T}\mathbf{\Lambda})-R(\mathbf{T}^0\mathbf{\Lambda}))+r_n(\mathbf{T}^0\mathbf{\Lambda})-r_n(\mathbf{T}\mathbf{\Lambda})\right)\right)\right] \leqslant 1 \quad (2.15)$$

for every $\mathbf{T} \in \mathcal{S}_{k,\tau}$ and $\lambda \in (0, n\lambda^*)$, where

$$R(\mathbf{A}) := \mathbb{E}(|Y_1 - \langle \mathbf{X}_1, \mathbf{A}\rangle_{\mathcal{F}}|^2) \ ; \ \forall \mathbf{A} \in \mathcal{M}_{d,T}(\mathbb{R}),$$

$\mathfrak{c}_{2.2.8} = 4\max\{4\mathfrak{m}_0^2, 4\mathfrak{v}_\xi, 4\mathfrak{m}_\varepsilon^2, 2\mathfrak{m}_\varepsilon^2\Phi_\varepsilon^2\mathfrak{c}_\Pi\}$ *and* $\lambda^* = (16\mathfrak{m}_0\max\{\mathfrak{m}_0, \mathfrak{m}_\varepsilon, \mathfrak{c}_\xi\})^{-1}$.

Proof of Proposition 2.2.8. The proof relies on Bernstein's inequality as stated in [64], that we remind in the following lemma.

Lemma 2.2.9. — *Let $T_1, \ldots, T_n$ be some independent and real-valued random variables. Assume that there are $v > 0$ and $c > 0$ such that*

$$\sum_{i=1}^{n} \mathbb{E}(T_i^2) \leqslant v$$

and, for any $q \geqslant 3$,

$$\sum_{i=1}^{n} \mathbb{E}(T_i^q) \leqslant \frac{v c^{q-2} q!}{2}.$$

Then, for every $\lambda \in (0, 1/c)$,

$$\mathbb{E}\left[\exp\left[\lambda \sum_{i=1}^{n} (T_i - \mathbb{E}(T_i))\right]\right] \leqslant \exp\left(\frac{v\lambda^2}{2(1-c\lambda)}\right).$$

We will also use a variant of this inequality for time series due to Samson, stated in the proof of Theorem 3 in [49].

Lemma 2.2.10. — *Consider* $m \in \mathbb{N}^*$, $M > 0$, *a stationary sequence of* $\mathbb{R}^m$*-valued random variables* $Z = (Z_t)_{t\in\mathbb{Z}}$, *and*

$$\Phi_Z := 1 + \sum_{t=1}^{T} \phi_Z(t)^{1/2},$$

where $\phi_Z(t)$, $t \in \mathbb{Z}$, *are the* ϕ*-mixing coefficients of* Z. *For every smooth and convex function* $f : [0, M]^T \to \mathbb{R}$ *such that* $\|\nabla f\| \leqslant L$ *a.e, for any* $\lambda > 0$,

$$\mathbb{E}(\exp(\lambda(f(Z_1, \ldots, Z_T) - \mathbb{E}[f(Z_1, \ldots, Z_T)]))) \leqslant \exp\left(\frac{\lambda^2 L^2 \Phi_Z^2 M^2}{2}\right).$$

Let $\mathbf{T} \in \mathcal{S}_{k,\tau}$ be arbitrarily chosen. Consider the deterministic map $\mathbf{X} : \mathcal{E} \to \mathcal{M}_{d,T}(\mathbb{R})$ such that

$$\mathbf{X}_i = \mathbf{X}(\chi_i) \ ; \ \forall i \in \{1, \ldots, n\},$$

$\Xi_i := (\overline{\xi}_i, \chi_i)$ for any $i \in \{1, \ldots, n\}$, and $h : \mathbb{R} \times \mathcal{E} \to \mathbb{R}$ the map defined by

$$h(x, y) := \frac{1}{n}(2x\langle \mathbf{X}(y), (\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\rangle_{\mathcal{F}} + \langle \mathbf{X}(y), (\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\rangle_{\mathcal{F}}^2) \ ; \ \forall (x, y) \in \mathbb{R} \times \mathcal{E}.$$

Note that

$$\begin{aligned} h(\Xi_i) &= \frac{1}{n}(2\overline{\xi}_i\langle \mathbf{X}_i, (\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\rangle_{\mathcal{F}} + \langle \mathbf{X}_i, (\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\rangle_{\mathcal{F}}^2) \\ &= \frac{1}{n}((\overline{\xi}_i + \langle \mathbf{X}_i, (\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\rangle_{\mathcal{F}})^2 - \overline{\xi}_i^2) \\ &= \frac{1}{n}((Y_i - \langle \mathbf{X}_i, \mathbf{T}\mathbf{\Lambda}\rangle_{\mathcal{F}})^2 - (Y_i - \langle \mathbf{X}_i, \mathbf{T}^0\mathbf{\Lambda}\rangle_{\mathcal{F}})^2) \end{aligned}$$

and

$$\sum_{i=1}^{n} (h(\Xi_i) - \mathbb{E}(h(\Xi_i))) = r_n(\mathbf{T}\mathbf{\Lambda}) - r_n(\mathbf{T}^0\mathbf{\Lambda}) + R(\mathbf{T}^0\mathbf{\Lambda}) - R(\mathbf{T}\mathbf{\Lambda}).$$

Now, replacing $\overline{\xi}_i$ by its expression in terms of $\mathbf{X}_i$, ξ_i and ε,

$$\begin{aligned} \sum_{i=1}^{n} (h(\Xi_i) - \mathbb{E}(h(\Xi_i))) &= \sum_{i=1}^{n} \left(\frac{2}{n}\xi_i\langle \mathbf{X}_i, (\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\rangle_{\mathcal{F}}\right) \\ &\quad + \sum_{i=1}^{n} \left(\frac{2}{n}\langle \mathbf{X}_i, \varepsilon\rangle_{\mathcal{F}}\langle \mathbf{X}_i, (\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\rangle_{\mathcal{F}}\right) \\ &\quad + \sum_{i=1}^{n} \left(\frac{1}{n}\langle \mathbf{X}_i, (\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\rangle_{\mathcal{F}}^2 - \mathbb{E}(h(\Xi_i))\right) \end{aligned}$$

$$=: \sum_{i=1}^{n} A_i + \sum_{i=1}^{n} B_i + \sum_{i=1}^{n} (C_i - \mathbb{E}(h(\Xi_i))).$$

In order to conclude, by using Lemmas 2.2.9 and 2.2.10, let us provide suitable bounds for the exponentiel moments of each terms of the previous decomposition :

- **Bounds for the A_i's and the C_i's.** First, note that since $\mathbf{X}_1$, ξ_1 and ε are independent,

$$\begin{aligned} R(\mathbf{T}\mathbf{\Lambda}) - R(\mathbf{T}^0\mathbf{\Lambda}) &= \mathbb{E}((Y_1 - \langle \mathbf{X}_1, \mathbf{T}\mathbf{\Lambda}\rangle_{\mathcal{F}})^2 - (Y_1 - \langle \mathbf{X}_1, \mathbf{T}^0\mathbf{\Lambda}\rangle_{\mathcal{F}})^2) \\ &= 2\mathbb{E}(\overline{\xi}_1 \langle \mathbf{X}_1, (\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\rangle_{\mathcal{F}}) + \mathbb{E}(\langle \mathbf{X}_1, (\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\rangle_{\mathcal{F}}^2) \\ &= 2\langle \mathbb{E}(\langle \mathbf{X}_1, (\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\rangle_{\mathcal{F}} \mathbf{X}_1), \mathbb{E}(\varepsilon)\rangle_{\mathcal{F}} \\ &\quad + 2\mathbb{E}(\xi_1)\mathbb{E}(\langle \mathbf{X}_1, (\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\rangle_{\mathcal{F}}) + \|(\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\|_{\mathcal{F},\Pi}^2 \\ &= \|(\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\|_{\mathcal{F},\Pi}^2. \end{aligned} \tag{2.16}$$

  On the one hand,

$$\mathbb{E}(A_i^2) \leqslant \frac{4}{n^2}\mathbb{E}(\xi_i^2)\mathbb{E}(\langle \mathbf{X}_i, (\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\rangle_{\mathcal{F}}^2) \leqslant \frac{4}{n^2}\mathfrak{v}_\xi (R(\mathbf{T}^0\mathbf{\Lambda}) - R(\mathbf{T}\mathbf{\Lambda}))$$

  thanks to Equality (2.16). Moreover,

$$\begin{aligned} \mathbb{E}(|A_i|^q) &\leqslant \frac{2^q}{n^q}\mathbb{E}(|\xi_i|^q)\mathbb{E}(\langle \mathbf{X}_i, (\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\rangle_{\mathcal{F}}^q) \\ &\leqslant \left(\frac{4\mathfrak{c}_\xi \mathfrak{m}_0}{n}\right)^{q-2} \frac{q!}{2} \cdot \frac{4\mathfrak{v}_\xi}{n^2}(R(\mathbf{T}^0\mathbf{\Lambda}) - R(\mathbf{T}\mathbf{\Lambda})). \end{aligned}$$

  So, we can use Lemma 2.2.9 with

$$v = \frac{4}{n}\mathfrak{v}_\xi (R(\mathbf{T}^0\mathbf{\Lambda}) - R(\mathbf{T}\mathbf{\Lambda})) \text{ and } c = \frac{4\mathfrak{c}_\xi \mathfrak{m}_0}{n}$$

  to obtain :

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} A_i\right)\right] \leqslant \exp\left[\frac{2\mathfrak{v}_\xi (R(\mathbf{T}^0\mathbf{\Lambda}) - R(\mathbf{T}\mathbf{\Lambda}))\lambda^2}{n - 4\mathfrak{c}_\xi \mathfrak{m}_0 \lambda}\right]$$

  for any $\lambda \in (0, n/(4\mathfrak{c}_\xi \mathfrak{m}_0))$. On the other hand, $|C_i| \leqslant 4\mathfrak{m}_0^2/n$ and

$$\begin{aligned} \mathbb{E}(C_i^2) &= \frac{1}{n^2}\mathbb{E}(\langle \mathbf{X}_i, (\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\rangle_{\mathcal{F}}^4) \\ &\leqslant \frac{4\mathfrak{m}_0^2}{n^2}\|(\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\|_{\mathcal{F},\Pi}^2 = \frac{4}{n^2}\mathfrak{m}_0^2 (R(\mathbf{T}^0\mathbf{\Lambda}) - R(\mathbf{T}\mathbf{\Lambda})) \end{aligned} \tag{2.17}$$

  thanks to Equality (2.16). So, we can use Lemma 2.2.9 with

$$v = \frac{4}{n}\mathfrak{m}_0^2 (R(\mathbf{T}^0\mathbf{\Lambda}) - R(\mathbf{T}\mathbf{\Lambda})) \text{ and } c = \frac{4\mathfrak{m}_0^2}{n}$$

  to obtain :

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} (C_i - \mathbb{E}(h(\Xi_i)))\right)\right] \leqslant \exp\left[\frac{2\mathfrak{m}_0^2 (R(\mathbf{T}^0\mathbf{\Lambda}) - R(\mathbf{T}\mathbf{\Lambda}))\lambda^2}{n - 4\mathfrak{m}_0^2 \lambda}\right]$$

  for any $\lambda \in (0, n/(4\mathfrak{m}_0^2))$.

- **Bounds for the B_i's.** First, write

$$\sum_{i=1}^{n} B_i = \sum_{i=1}^{n} (B_i - \mathbb{E}(B_i|\varepsilon)) + \sum_{i=1}^{n} \mathbb{E}(B_i|\varepsilon) =: \sum_{i=1}^{n} D_i + \sum_{i=1}^{n} E_i,$$

and note that

$$\begin{aligned} \mathbb{E}(B_i|\varepsilon) &= \frac{2}{n}\mathbb{E}(\langle \mathbf{X}_i, \varepsilon\rangle_{\mathcal{F}} \langle \mathbf{X}_i, (\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\rangle_{\mathcal{F}}|\varepsilon) \\ &= \frac{2}{n}\sum_{j,t} \mathbb{E}(\mathbf{1}_{\chi_i=(j,t)}[(\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}]_{\chi_i})\varepsilon_{j,t} \\ &= \frac{2}{n}\sum_{j,t} p_{j,t}[(\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}]_{j,t}\varepsilon_{j,t} \end{aligned} \tag{2.18}$$

and

$$\begin{aligned} \|(\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} &= \mathbb{E}(\langle \mathbf{X}_i, (\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\rangle^2_{\mathcal{F}}) \\ &= \mathbb{E}([(\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}]^2_{\chi_i}) \\ &= \sum_{j,t} p_{j,t}[(\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}]^2_{j,t}, \end{aligned} \tag{2.19}$$

where

$$p_{j,t} := \mathbb{P}(\chi_1 = (j,t)) = \Pi(\{e_{\mathbb{R}^d}(j) e_{\mathbb{R}^T}(t)^*\})$$

for every $(j,t) \in \mathcal{E}$. On the one hand, given ε, the D_i's are i.i.d, $|D_i| \leqslant 8\mathfrak{m}_\varepsilon \mathfrak{m}_0/n$ and

$$\begin{aligned} \mathbb{E}(B_i^2|\varepsilon) &= \frac{4}{n^2}\mathbb{E}(\langle \mathbf{X}_i, \varepsilon\rangle^2_{\mathcal{F}} \langle \mathbf{X}_i, (\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\rangle^2_{\mathcal{F}}|\varepsilon) \\ &\leqslant \frac{4}{n^2}\mathfrak{m}_\varepsilon^2 \mathbb{E}(\langle \mathbf{X}_i, (\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\rangle^2_{\mathcal{F}}|\varepsilon) \\ &= \frac{4}{n^2}\mathfrak{m}_\varepsilon^2 \mathbb{E}(\langle \mathbf{X}_i, (\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}\rangle^2_{\mathcal{F}}) = \frac{4}{n^2}\mathfrak{m}_\varepsilon^2 (R(\mathbf{T}^0\mathbf{\Lambda}) - R(\mathbf{T}\mathbf{\Lambda})) \end{aligned}$$

thanks to Equality (2.16). So, *conditionnally on* ε, we can apply Lemma 2.2.9 with

$$v = \frac{4}{n}\mathfrak{m}_\varepsilon^2 (R(\mathbf{T}^0\mathbf{\Lambda}) - R(\mathbf{T}\mathbf{\Lambda})) \text{ and } c = \frac{8\mathfrak{m}_\varepsilon \mathfrak{m}_0}{n}$$

to obtain :

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} D_i\right)\middle|\varepsilon\right] \leqslant \exp\left[\frac{2\mathfrak{m}_\varepsilon^2 (R(\mathbf{T}^0\mathbf{\Lambda}) - R(\mathbf{T}\mathbf{\Lambda}))\lambda^2}{n - 8\mathfrak{m}_\varepsilon \mathfrak{m}_0 \lambda}\right]$$

for any $\lambda \in (0, n/(8\mathfrak{m}_\varepsilon \mathfrak{m}_0))$. Taking the expectation of both sides gives :

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} D_i\right)\right] \leqslant \exp\left[\frac{2\mathfrak{m}_\varepsilon^2 (R(\mathbf{T}^0\mathbf{\Lambda}) - R(\mathbf{T}\mathbf{\Lambda}))\lambda^2}{n - 8\mathfrak{m}_\varepsilon \mathfrak{m}_0 \lambda}\right].$$

On the other hand, let us focus on the E_i's. Thanks to Equality (2.18) and since the rows of ε are independent,

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} E_i\right)\right] = \mathbb{E}\left[\exp\left[2\lambda \sum_{j,t} p_{j,t}[(\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}]_{j,t}\varepsilon_{j,t}\right]\right]$$

$$= \prod_{j=1}^{d} \mathbb{E}\left[\exp\left(2\lambda \sum_{t=1}^{T} p_{j,t}[(\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}]_{j,t}\varepsilon_{j,t}\right)\right].$$

Now, for any $j \in \{1, \dots, d\}$, let us apply Lemma 2.2.10 to $(\varepsilon_{j,1}, \dots, \varepsilon_{j,T})$, which is a sample of a ϕ-mixing sequence, and to the function $f_j : [0, \mathfrak{m}_\varepsilon]^T \to \mathbb{R}$ defined by

$$f_j(u_1, \dots, u_T) := 2\sum_{t=1}^{T} p_{j,t}[(\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}]_{j,t}u_t \ ; \forall u \in [0, \mathfrak{m}_\varepsilon]^T.$$

Since

$$\|\nabla f_j(u_1, \dots, u_T)\|^2 = 4\sum_{t=1}^{T} p_{j,t}^2[(\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}]_{j,t}^2 \ ; \forall u \in [0, \mathfrak{m}_\varepsilon]^T,$$

by Lemma 2.2.10 :

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(2\lambda \sum_{t=1}^{T} p_{j,t}[(\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}]_{j,t}\varepsilon_{j,t}\right)\right] \\
&= \mathbb{E}(\exp(\lambda(f_j(\varepsilon_{j,1}, \dots, \varepsilon_{j,T}) - \mathbb{E}[f_j(\varepsilon_{j,1}, \dots, \varepsilon_{j,T})])))\\
&\leqslant \exp\left(2\mathfrak{m}_\varepsilon^2\lambda^2\Phi_\varepsilon^2 \sum_{t=1}^{T} p_{j,t}^2[(\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}]_{j,t}^2\right).
\end{aligned}$$

Thus, for any $\lambda > 0$, by Equalities (2.16) and (2.19) together with $n \leqslant dT$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} E_i\right)\right] &= \prod_{j=1}^{d} \mathbb{E}\left[\exp\left(2\lambda \sum_{t=1}^{T} p_{j,t}[(\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}]_{j,t}\varepsilon_{j,t}\right)\right] \\
&\leqslant \prod_{j=1}^{d} \exp\left(2\mathfrak{m}_\varepsilon^2\lambda^2\Phi_\varepsilon^2 \sum_{t=1}^{T} p_{j,t}^2[(\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}]_{j,t}^2\right) \\
&\leqslant \exp\left[\frac{2\mathfrak{m}_\varepsilon^2\lambda^2\Phi_\varepsilon^2\mathfrak{c}_\Pi}{dT} \sum_{j,t} p_{j,t}[(\mathbf{T}^0 - \mathbf{T})\mathbf{\Lambda}]_{j,t}^2\right] \\
&\leqslant \exp\left[\frac{2\mathfrak{m}_\varepsilon^2\lambda^2\Phi_\varepsilon^2\mathfrak{c}_\Pi}{n}(R(\mathbf{T}^0\mathbf{\Lambda}) - R(\mathbf{T}\mathbf{\Lambda}))\right].
\end{aligned}$$

Therefore, these bounds together with Jensen's inequality give :

$$\begin{aligned}
&\mathbb{E}\exp\left(\frac{\lambda}{4}[r_n(\mathbf{T}\mathbf{\Lambda}) - r_n(\mathbf{T}^0\mathbf{\Lambda}) + R(\mathbf{T}^0\mathbf{\Lambda}) - R(\mathbf{T}\mathbf{\Lambda})]\right) \\
&= \mathbb{E}\left[\exp\left(\frac{\lambda}{4}\sum_{i=1}^{n}(h(\Xi_i) - \mathbb{E}(h(\Xi_i)))\right)\right] \\
&= \mathbb{E}\left[\exp\left(\frac{\lambda}{4}\sum_{i=1}^{n} A_i + \frac{\lambda}{4}\sum_{i=1}^{n}(C_i - \mathbb{E}(h(\Xi_i))) + \frac{\lambda}{4}\sum_{i=1}^{n} D_i + \frac{\lambda}{4}\sum_{i=1}^{n} E_i\right)\right] \\
&\leqslant \frac{1}{4}\left[\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n} A_i\right)\right] + \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n}(C_i - \mathbb{E}(h(\Xi_i)))\right)\right]\right. \\
&\quad \left. + \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n} D_i\right)\right] + \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n} E_i\right)\right]\right] \\
&\leqslant \exp\left[\frac{2\mathfrak{v}_\xi}{1 - 4\mathfrak{c}_\xi\mathfrak{m}_0\lambda/n} \cdot \frac{\lambda^2}{n}(R(\mathbf{T}^0\mathbf{\Lambda}) - R(\mathbf{T}\mathbf{\Lambda}))\right]
\end{aligned}$$

$$
\begin{aligned}
&+\exp\left[\frac{2\mathfrak{m}_0^2}{1-4\mathfrak{m}_0^2\lambda/n}\cdot\frac{\lambda^2}{n}(R(\mathbf{T}^0\mathbf{\Lambda})-R(\mathbf{T}\mathbf{\Lambda}))\right]\\
&+\exp\left[\frac{2\mathfrak{m}_\varepsilon^2}{1-8\mathfrak{m}_\varepsilon\mathfrak{m}_0\lambda/n}\cdot\frac{\lambda^2}{n}(R(\mathbf{T}^0\mathbf{\Lambda})-R(\mathbf{T}\mathbf{\Lambda}))\right]\\
&+\exp\left[2\mathfrak{m}_\varepsilon^2\Phi_\varepsilon^2\mathfrak{c}_\Pi\frac{\lambda^2}{n}(R(\mathbf{T}^0\mathbf{\Lambda})-R(\mathbf{T}\mathbf{\Lambda}))\right]\\
&\leqslant\exp\left[\mathfrak{c}_\lambda\frac{\lambda^2}{n}(R(\mathbf{T}^0\mathbf{\Lambda})-R(\mathbf{T}\mathbf{\Lambda}))\right]
\end{aligned}
$$

with

$$
\mathfrak{c}_\lambda=\max\left\{\frac{2\mathfrak{v}_\xi}{1-4\mathfrak{c}_\xi\mathfrak{m}_0\lambda/n},\frac{2\mathfrak{m}_0^2}{1-4\mathfrak{m}_0^2\lambda/n},\frac{2\mathfrak{m}_\varepsilon^2}{1-8\mathfrak{m}_\varepsilon\mathfrak{m}_0\lambda/n},2\mathfrak{m}_\varepsilon^2\Phi_\varepsilon^2\mathfrak{c}_\Pi\right\}
$$

and

$$
0<\lambda<n\min\left\{\frac{1}{4\mathfrak{c}_\xi\mathfrak{m}_0},\frac{1}{4\mathfrak{m}_0^2},\frac{1}{8\mathfrak{m}_\varepsilon\mathfrak{m}_0}\right\}.
$$

In particular, for

$$
\lambda<\frac{n}{16\mathfrak{m}_0\max\{\mathfrak{m}_0,\mathfrak{m}_\varepsilon,\mathfrak{c}_\xi\}},
$$

we have

$$
\mathfrak{c}_\lambda\leqslant\max\{4\mathfrak{m}_0^2,4\mathfrak{v}_\xi,4\mathfrak{m}_\varepsilon^2,2\mathfrak{m}_\varepsilon^2\Phi_\varepsilon^2\mathfrak{c}_\Pi\}.
$$

This ends the proof of the first inequality.

2.2.5.2 A preliminary non-explicit risk bound

We now provide a simpler version of Theorem 2.2.5, that holds in the case where $\mathcal{S}_{k,\tau}$ is finite : (1) in the following theorem. When this is not the case, we provide a similar bound using a general ϵ-net, that is (2) in the theorem.

Theorem 2.2.11. — *Consider $\alpha\in]0,1[$.*

(1) Under Assumptions 2.2.2, 2.2.3 and 2.2.4, if $|\mathcal{S}_{k,\tau}|<\infty$, then

$$
\|\widehat{\mathbf{\Theta}}_{k,\tau}-\mathbf{\Theta}^0\|_{\mathcal{F},\Pi}^2\leqslant 3\min_{\mathbf{T}\in\mathcal{S}_{k,\tau}}\|(\mathbf{T}-\mathbf{T}^0)\mathbf{\Lambda}\|_{\mathcal{F},\Pi}^2+\frac{\mathfrak{c}_{2.2.11,1}}{n}\log\left(\frac{2}{\alpha}|\mathcal{S}_{k,\tau}|\right)
$$

with probability larger than $1-\alpha$, where $\mathfrak{c}_{2.2.11,1}=32(\mathfrak{c}_{2.2.8}^{-1}\wedge\lambda^*)^{-1}$.

(2) Under Assumptions 2.2.2, 2.2.3 and 2.2.4, for every $\epsilon>0$, there exists a finite subset $\mathcal{S}_{k,\tau}^\epsilon$ of $\mathcal{S}_{k,\tau}$ such that

$$
\begin{aligned}
\|\widehat{\mathbf{\Theta}}_{k,\tau}-\mathbf{\Theta}^0\|_{\mathcal{F},\Pi}^2\leqslant 3\min_{\mathbf{T}\in\mathcal{S}_{k,\tau}}\|(\mathbf{T}-\mathbf{T}^0)\mathbf{\Lambda}\|_{\mathcal{F},\Pi}^2+\frac{\mathfrak{c}_{2.2.11,1}}{n}\log\left(\frac{2}{\alpha}|\mathcal{S}_{k,\tau}^\epsilon|\right)\\
+\left[\mathfrak{c}_{2.2.11,2}+8\mathfrak{m}_{\mathbf{\Lambda}}\mathfrak{c}_\xi\log\left(\frac{1}{\alpha}\right)\right]\epsilon \quad (2.20)
\end{aligned}
$$

with probability larger than $1-\alpha$, where $\mathfrak{c}_{2.2.11,2}=4\mathfrak{m}_{\mathbf{\Lambda}}(\mathfrak{v}_\xi^{1/2}+\mathfrak{v}_\xi/(2\mathfrak{c}_\xi)+\mathfrak{m}_\varepsilon+3\mathfrak{m}_0)$.

Proof of Theorem 2.2.11.

(1) Assume that $|\mathcal{S}_{k,\tau}| < \infty$. For any $x > 0$, $\lambda \in (0, n\lambda^*)$ and $\mathcal{S} \subset \mathcal{M}_{d,\tau}(\mathbb{R})$, consider the events

$$\Omega^-_{x,\lambda,\mathcal{S}}(\mathbf{T}) := \left\{ \left(1 - \mathfrak{c}_{2.2.8}\frac{\lambda}{n}\right) \|(\mathbf{T} - \mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} - (r_n(\mathbf{T}\mathbf{\Lambda}) - r_n(\mathbf{T}^0\mathbf{\Lambda})) > 4x \right\}$$

for $\mathbf{T} \in \mathcal{S}$ and

$$\Omega^-_{x,\lambda,\mathcal{S}} := \bigcup_{\mathbf{T}\in\mathcal{S}} \Omega^-_{x,\lambda,\mathcal{S}}(\mathbf{T}).$$

By Markov's inequality together with Proposition 2.2.8, Inequality (2.15),

$$\begin{aligned}\mathbb{P}(\Omega^-_{x,\lambda,\mathcal{S}_{k,\tau}}) \leqslant \sum_{\mathbf{T}\in\mathcal{S}_{k,\tau}} \mathbb{P}\Bigg\{ \exp\Bigg[\frac{\lambda}{4}\left(1 - \frac{\mathfrak{c}_{2.2.8}\lambda}{n}\right) \\ \times \Big(R(\mathbf{T}\mathbf{\Lambda}) - R(\mathbf{T}^0\mathbf{\Lambda})) - (r_n(\mathbf{T}\mathbf{\Lambda}) - r_n(\mathbf{T}^0\mathbf{\Lambda})\Big)\Bigg] > e^{\lambda x}\Bigg\} \\ \leqslant |\mathcal{S}_{k,\tau}|e^{-\lambda x}.\end{aligned}$$

In the same way, with

$$\Omega^+_{x,\lambda,\mathcal{S}}(\mathbf{T}) := \left\{ -\left(1 + \mathfrak{c}_{2.2.8}\frac{\lambda}{n}\right) \|(\mathbf{T} - \mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} + r_n(\mathbf{T}\mathbf{\Lambda}) - r_n(\mathbf{T}^0\mathbf{\Lambda}) > 4x \right\},$$

$\mathbf{T} \in \mathcal{S}$ and

$$\Omega^+_{x,\lambda,\mathcal{S}} := \bigcup_{\mathbf{T}\in\mathcal{S}} \Omega^+_{x,\lambda,\mathcal{S}}(\mathbf{T}),$$

by Markov's inequality together with Proposition 2.2.8, Inequality (2.14), $\mathbb{P}(\Omega^+_{x,\lambda,\mathcal{S}_{k,\tau}}) \leqslant |\mathcal{S}_{k,\tau}|e^{-\lambda x}$. Then,

$$\mathbb{P}(\Omega_{x,\lambda,\mathcal{S}_{k,\tau}}) \geqslant 1 - 2|\mathcal{S}_{k,\tau}|e^{-\lambda x}$$

with

$$\Omega_{x,\lambda,\mathcal{S}} := (\Omega^-_{x,\lambda,\mathcal{S}})^c \cap (\Omega^+_{x,\lambda,\mathcal{S}})^c \subset \Omega^-_{x,\lambda,\mathcal{S}}(\widehat{\mathbf{T}}_{k,\tau})^c \cap \Omega^+_{x,\lambda,\mathcal{S}}(\widehat{\mathbf{T}}_{k,\tau})^c =: \Omega_{x,\lambda,\mathcal{S}_{k,\tau}}(\widehat{\mathbf{T}}_{k,\tau}).$$

Moreover, on the event $\Omega_{x,\lambda,\mathcal{S}_{k,\tau}}$, by the definition of $\widehat{\mathbf{T}}_{k,\tau}$,

$$\begin{aligned}\|\widehat{\mathbf{\Theta}}_{k,\tau} - \mathbf{\Theta}^0\|^2_{\mathcal{F},\Pi} &\leqslant \left(1 - \mathfrak{c}_{2.2.8}\frac{\lambda}{n}\right)^{-1} (r_n(\widehat{\mathbf{T}}_{k,\tau}\mathbf{\Lambda}) - r_n(\mathbf{T}^0\mathbf{\Lambda}) + 4x) \\ &= \left(1 - \mathfrak{c}_{2.2.8}\frac{\lambda}{n}\right)^{-1} \left(\min_{\mathbf{T}\in\mathcal{S}_{k,\tau}} \{r_n(\mathbf{T}\mathbf{\Lambda}) - r_n(\mathbf{T}^0\mathbf{\Lambda})\} + 4x\right) \\ &\leqslant \frac{1 + \mathfrak{c}_{2.2.8}\lambda n^{-1}}{1 - \mathfrak{c}_{2.2.8}\lambda n^{-1}} \min_{\mathbf{T}\in\mathcal{S}_{k,\tau}} \|(\mathbf{T} - \mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} + \frac{8x}{1 - \mathfrak{c}_{2.2.8}\lambda n^{-1}}.\end{aligned}$$

So, for any $\alpha \in]0,1[$, with probability larger than $1 - \alpha$,

$$\|\widehat{\mathbf{\Theta}}_{k,\tau} - \mathbf{\Theta}^0\|^2_{\mathcal{F},\Pi} \leqslant \frac{1 + \mathfrak{c}_{2.2.8}\lambda n^{-1}}{1 - \mathfrak{c}_{2.2.8}\lambda n^{-1}} \min_{\mathbf{T}\in\mathcal{S}_{k,\tau}} \|(\mathbf{T} - \mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} + \frac{8\lambda^{-1}\log(2\alpha^{-1}|\mathcal{S}_{k,\tau}|)}{1 - \mathfrak{c}_{2.2.8}\lambda n^{-1}}.$$

Now, let us take

$$\lambda = \frac{n}{2}\left(\frac{1}{\mathfrak{c}_{2.2.8}} \wedge \lambda^*\right) \in (0, n\lambda^*) \text{ and } x = \frac{1}{\lambda}\log\left(\frac{2}{\alpha}|\mathcal{S}_{k,\tau}|\right).$$

In particular, $\mathfrak{c}_{2.2.8}\lambda n^{-1} \leqslant 1/2$, and then

$$\frac{1+\mathfrak{c}_{2.2.8}\lambda n^{-1}}{1-\mathfrak{c}_{2.2.8}\lambda n^{-1}} \leqslant 3 \text{ and } \frac{8\lambda^{-1}}{1-\mathfrak{c}_{2.2.8}\lambda n^{-1}} \leqslant 32\left(\frac{1}{\mathfrak{c}_{2.2.8}} \wedge \lambda^*\right)^{-1}\frac{1}{n}.$$

Therefore, with probability larger than $1-\alpha$,

$$\|\widehat{\mathbf{\Theta}}_{k,\tau} - \mathbf{\Theta}^0\|^2_{\mathcal{F},\Pi} \leqslant 3\min_{\mathbf{T}\in\mathcal{S}_{k,\tau}} \|(\mathbf{T}-\mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} + 32\left(\frac{1}{\mathfrak{c}_{2.2.8}} \wedge \lambda^*\right)^{-1}\frac{1}{n}\log\left(\frac{2}{\alpha}|\mathcal{S}_{k,\tau}|\right).$$

(2) Now, assume that $|\mathcal{S}_{k,\tau}| = \infty$. Since $\dim(\mathcal{M}_{d,\tau}(\mathbb{R})) < \infty$ and $\mathcal{S}_{k,\tau}$ is a bounded subset of $\mathcal{M}_{d,\tau}(\mathbb{R})$ (equipped with $\mathbf{T} \mapsto \sup_{j,t}|\mathbf{T}_{j,t}|$), $\mathcal{S}_{k,\tau}$ is compact in $(\mathcal{M}_{d,\tau}(\mathbb{R}), \|.\|_{\mathcal{F}})$. Then, for any $\epsilon > 0$, there exists a finite subset $\mathcal{S}^\epsilon_{k,\tau}$ of $\mathcal{S}_{k,\tau}$ such that

$$\forall \mathbf{T} \in \mathcal{S}_{k,\tau}, \exists \mathbf{T}^\epsilon \in \mathcal{S}^\epsilon_{k,\tau} : \|\mathbf{T}-\mathbf{T}^\epsilon\|_{\mathcal{F}} \leqslant \epsilon. \tag{2.21}$$

On the one hand, for any $\mathbf{T} \in \mathcal{S}_{k,\tau}$ and $\mathbf{T}^\epsilon \in \mathcal{S}^\epsilon_{k,\tau}$ satisfying (2.21), since $\langle \mathbf{X}_i, (\mathbf{T}-\mathbf{T}^\epsilon)\mathbf{\Lambda}\rangle_{\mathcal{F}} = \langle \mathbf{X}_i\mathbf{\Lambda}^*, \mathbf{T}-\mathbf{T}^\epsilon\rangle_{\mathcal{F}}$ for every $i \in \{1,\dots,n\}$,

$$\begin{aligned}
|r_n(\mathbf{T}\mathbf{\Lambda}) - r_n(\mathbf{T}^\epsilon\mathbf{\Lambda})| &\leqslant \frac{1}{n}\sum_{i=1}^n |\langle \mathbf{X}_i, (\mathbf{T}-\mathbf{T}^\epsilon)\mathbf{\Lambda}\rangle_{\mathcal{F}}(2Y_i - \langle \mathbf{X}_i, (\mathbf{T}+\mathbf{T}^\epsilon)\mathbf{\Lambda}\rangle_{\mathcal{F}})| \\
&\leqslant \frac{\epsilon}{n}\sum_{i=1}^n \|\mathbf{X}_i\mathbf{\Lambda}^*\|_{\mathcal{F}}\left(2|Y_i| + \sup_{j,t}\left|\sum_{\ell=1}^\tau (\mathbf{T}+\mathbf{T}^\epsilon)_{j,\ell}\mathbf{\Lambda}_{\ell,t}\right|\right) \\
&\leqslant \epsilon\mathfrak{m}_{\mathbf{\Lambda}}\left(\frac{2}{n}\sum_{i=1}^n |Y_i| + 2\mathfrak{m}_0\right) \leqslant \mathfrak{c}_1(\xi_1,\dots,\xi_n)\epsilon
\end{aligned} \tag{2.22}$$

with

$$\mathfrak{c}_1(\xi_1,\dots,\xi_n) := 2\mathfrak{m}_{\mathbf{\Lambda}}\left(\frac{1}{n}\sum_{i=1}^n |\xi_i| + \mathfrak{m}_\varepsilon + 2\mathfrak{m}_0\right),$$

and thanks to Equality (2.16),

$$\begin{aligned}
|R(\mathbf{T}\mathbf{\Lambda}) - R(\mathbf{T}^\epsilon\mathbf{\Lambda})| &= |R(\mathbf{T}\mathbf{\Lambda}) - R(\mathbf{T}^0\mathbf{\Lambda}) - (R(\mathbf{T}^\epsilon\mathbf{\Lambda}) - R(\mathbf{T}^0\mathbf{\Lambda}))| \\
&= |\|(\mathbf{T}-\mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} - \|(\mathbf{T}^\epsilon-\mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi}| \\
&\leqslant \mathbb{E}(|\langle \mathbf{X}_i, (\mathbf{T}-\mathbf{T}^\epsilon)\mathbf{\Lambda}\rangle_{\mathcal{F}}\langle \mathbf{X}_i, (\mathbf{T}+\mathbf{T}^\epsilon-2\mathbf{T}^0)\mathbf{\Lambda}\rangle_{\mathcal{F}}|) \\
&\leqslant \mathfrak{c}_2\epsilon
\end{aligned} \tag{2.23}$$

with $\mathfrak{c}_2 = 4\mathfrak{m}_0\mathfrak{m}_{\mathbf{\Lambda}}$. On the other hand, consider

$$\widehat{\mathbf{T}}^\epsilon_{k,\tau} = \arg\min_{\mathbf{T}\in\mathcal{S}^\epsilon_{k,\tau}} \|\mathbf{T}-\widehat{\mathbf{T}}_{k,\tau}\|_{\mathcal{F}}. \tag{2.24}$$

On the event $\Omega_{x,\lambda,\mathcal{S}^\epsilon_{k,\tau}}$ with $x > 0$ and $\lambda \in (0, n\lambda^*)$, by the definitions of $\widehat{\mathbf{T}}^\epsilon_{k,\tau}$ and $\widehat{\mathbf{T}}_{k,\tau}$, and thanks to Inequalities (2.22) and (2.23),

$$\begin{aligned}
\|\widehat{\mathbf{\Theta}}_{k,\tau} - \mathbf{\Theta}^0\|^2_{\mathcal{F},\Pi} &\leqslant \|(\widehat{\mathbf{T}}^\epsilon_{k,\tau}-\mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} + \mathfrak{c}_2\epsilon \\
&\leqslant \left(1-\mathfrak{c}_{2.2.8}\frac{\lambda}{n}\right)^{-1}(r_n(\widehat{\mathbf{T}}^\epsilon_{k,\tau}\mathbf{\Lambda}) - r_n(\mathbf{T}^0\mathbf{\Lambda}) + 4x) + \mathfrak{c}_2\epsilon
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \left(1-\mathfrak{c}_{2.2.8}\frac{\lambda}{n}\right)^{-1}\left[r_n(\widehat{\mathbf{T}}_{k,\tau}\mathbf{\Lambda})-r_n(\mathbf{T}^0\mathbf{\Lambda})\right.\\
&\qquad\left.+\mathfrak{c}_1(\xi_1,\ldots,\xi_n)\epsilon+4x\right]+\mathfrak{c}_2\epsilon\\
&= \left(1-\mathfrak{c}_{2.2.8}\frac{\lambda}{n}\right)^{-1}\left[\min_{\mathbf{T}\in\mathcal{S}_{k,\tau}}\{r_n(\mathbf{T}\mathbf{\Lambda})-r_n(\mathbf{T}^0\mathbf{\Lambda})\}\right.\\
&\qquad\left.+\mathfrak{c}_1(\xi_1,\ldots,\xi_n)\epsilon+4x\right]+\mathfrak{c}_2\epsilon\\
&\leqslant \frac{1+\mathfrak{c}_{2.2.8}\lambda n^{-1}}{1-\mathfrak{c}_{2.2.8}\lambda n^{-1}}\min_{\mathbf{T}\in\mathcal{S}_{k,\tau}}\|(\mathbf{T}-\mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi}+\frac{8x}{1-\mathfrak{c}_{2.2.8}\lambda n^{-1}}\\
&\quad+\left[\frac{\mathfrak{c}_1(\xi_1,\ldots,\xi_n)}{1-\mathfrak{c}_{2.2.8}\lambda n^{-1}}+\mathfrak{c}_2\right]\epsilon.
\end{aligned}
$$

So, by taking

$$
\lambda=\frac{n}{2}\left(\frac{1}{\mathfrak{c}_{2.2.8}}\wedge\lambda^*\right)\text{ and }x=\frac{1}{\lambda}\log\left(\frac{2}{\alpha}|\mathcal{S}^\epsilon_{k,\tau}|\right),
$$

as in the proof of Theorem 2.2.11.(1), with probability larger than $1-\alpha$,

$$
\begin{aligned}
\|\widehat{\mathbf{\Theta}}_{k,\tau}-\mathbf{\Theta}^0\|^2_{\mathcal{F},\Pi} &\leqslant 3\min_{\mathbf{T}\in\mathcal{S}_{k,\tau}}\|(\mathbf{T}-\mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi}\\
&\quad+32\left(\frac{1}{\mathfrak{c}_{2.2.8}}\wedge\lambda^*\right)^{-1}\frac{1}{n}\log\left(\frac{2}{\alpha}|\mathcal{S}^\epsilon_{k,\tau}|\right)\\
&\quad+\left[4\mathfrak{m}_{\mathbf{\Lambda}}\left(\frac{1}{n}\sum_{i=1}^n|\xi_i|+\mathfrak{m}_\varepsilon+2\mathfrak{m}_0\right)+\mathfrak{c}_2\right]\epsilon. \qquad (2.25)
\end{aligned}
$$

Thanks to Markov's inequality together with Lemma 2.2.9, for $\lambda_0=1/(2n\mathfrak{c}_\xi)$,

$$
\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^n|\xi_i|>\sum_{i=1}^n\mathbb{E}(|\xi_i|)+s\right) &\leqslant \exp\left[\frac{n\mathfrak{v}_\xi\lambda_0^2}{2(1-n\mathfrak{c}_\xi\lambda_0)}-\lambda_0 s\right]\\
&= \exp\left(\frac{\mathfrak{v}_\xi}{4n\mathfrak{c}_\xi^2}-\frac{s}{2n\mathfrak{c}_\xi}\right)=\alpha
\end{aligned}
$$

with

$$
s=\frac{\mathfrak{v}_\xi}{2\mathfrak{c}_\xi}+2n\mathfrak{c}_\xi\log\left(\frac{1}{\alpha}\right).
$$

Then, since $\mathbb{E}(|\xi_i|)\leqslant\mathbb{E}(\xi_i^2)^{1/2}\leqslant\mathfrak{v}_\xi^{1/2}$ for every $i\in\{1,\ldots,n\}$,

$$
\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n|\xi_i|>\mathfrak{v}_\xi^{1/2}+\frac{\mathfrak{v}_\xi}{2n\mathfrak{c}_\xi}+2\mathfrak{c}_\xi\log\left(\frac{1}{\alpha}\right)\right]\leqslant\alpha. \qquad (2.26)
$$

Finally, note that if $\mathbb{P}(U>V+c)\leqslant\alpha$ and $\mathbb{P}(V>v)\leqslant\alpha$ with $c,v\in\mathbb{R}_+$ and (U,V) a $\mathbb{R}^2$-valued random variable, then

$$
\begin{aligned}
\mathbb{P}(U>v+c) &= \mathbb{P}(U>v+c,V>v)+\mathbb{P}(U>v+c,V\leqslant v)\\
&\leqslant \mathbb{P}(V>v)+\mathbb{P}(U>V+c,V\leqslant v)\leqslant 2\alpha. \qquad (2.27)
\end{aligned}
$$

Therefore, by (2.25) and (2.26), with probability larger than $1-2\alpha$,

$$
\|\widehat{\mathbf{\Theta}}_{k,\tau}\quad-\quad\mathbf{\Theta}^0\|^2_{\mathcal{F},\Pi}
$$

$$
\begin{aligned}
\leqslant \; & 3 \min_{\mathbf{T} \in \mathcal{S}_{k,\tau}} \|(\mathbf{T} - \mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} \\
& + 32 \left(\frac{1}{\mathfrak{c}_{2.2.8}} \wedge \lambda^* \right)^{-1} \frac{1}{n} \log \left(\frac{2}{\alpha} |\mathcal{S}^\epsilon_{k,\tau}| \right) \\
& + \left[4\mathfrak{m}_{\mathbf{\Lambda}} \left(2\mathfrak{c}_\xi \log \left(\frac{1}{\alpha} \right) + \mathfrak{v}_\xi^{1/2} + \frac{\mathfrak{v}_\xi}{2\mathfrak{c}_\xi} + \mathfrak{m}_\varepsilon + 2\mathfrak{m}_0 \right) + \mathfrak{c}_2 \right] \epsilon.
\end{aligned}
$$

2.2.5.3 Proof of Theorem 2.2.5

The proof is dissected in two steps :

Step 1. Consider

$$
\mathcal{M}_{d,\tau,k}(\mathbb{R}) := \{\mathbf{T} \in \mathcal{M}_{d,\tau}(\mathbb{R}) : \mathrm{rank}(\mathbf{T}) = k\}.
$$

For every $\mathbf{T} \in \mathcal{M}_{d,\tau,k}(\mathbb{R})$ and $\rho > 0$, let us denote the closed ball (resp. the sphere) of center $\mathbf{T}$ and of radius ρ of $\mathcal{M}_{d,\tau,k}(\mathbb{R})$ by $\mathbb{B}_k(\mathbf{T}, \rho)$ (resp. $\mathbb{S}_k(\mathbf{T}, \rho)$). For any $\epsilon > 0$, thanks to Candès and Plan [66], Lemma 3.1, there exists an ϵ-net $\mathbb{S}^\epsilon_k(0,1)$ covering $\mathbb{S}_k(0,1)$ and such that

$$
|\mathbb{S}^\epsilon_k(0,1)| \leqslant \left(\frac{9}{\epsilon} \right)^{k(d+\tau+1)}.
$$

Then, for every $\rho > 0$, there exists an ϵ-net $\mathbb{S}^\epsilon_k(0,\rho)$ covering $\mathbb{S}_k(0,\rho)$ and such that

$$
|\mathbb{S}^\epsilon_k(0,\rho)| \leqslant \left(\frac{9\rho}{\epsilon} \right)^{k(d+\tau+1)}.
$$

Moreover, for any $\rho^* > 0$,

$$
\mathbb{B}_k(0,\rho^*) = \bigcup_{\rho \in [0,\rho^*]} \mathbb{S}_k(0,\rho).
$$

So,

$$
\mathbb{B}^\epsilon_k(0,\rho^*) := \bigcup_{j=0}^{[\rho^*/\epsilon]+1} \mathbb{S}^\epsilon_k(0, j\epsilon)
$$

is an ϵ-net covering $\mathbb{B}_k(0,\rho^*)$ and such that

$$
|\mathbb{B}^\epsilon_k(0,\rho^*)| \leqslant \sum_{j=0}^{[\rho^*/\epsilon]+1} |\mathbb{S}^\epsilon_k(0,j\epsilon)| \leqslant \left(\left[\frac{\rho^*}{\epsilon} \right] + 2 \right) \left(\frac{9\rho^*}{\epsilon} \right)^{k(d+\tau+1)}.
$$

If in addition $\rho^* \geqslant \epsilon$, then

$$
|\mathbb{B}^\epsilon_k(0,\rho^*)| \leqslant \frac{3\rho^*}{\epsilon} \left(\frac{9\rho^*}{\epsilon} \right)^{k(d+\tau+1)} \leqslant \left(\frac{9\rho^*}{\epsilon} \right)^{2k(d+\tau)}.
$$

Step 2. For any $\mathbf{T} \in \mathcal{S}_{k,\tau}$,

$$
\sup_{j,t} |\mathbf{T}_{j,t}| \leqslant \frac{\mathfrak{m}_0}{\mathfrak{m}_{\mathbf{\Lambda}}(\tau)}.
$$

Then,

$$
\|\mathbf{T}\|_{\mathcal{F}} = \left(\sum_{j=1}^{d} \sum_{t=1}^{\tau} \mathbf{T}^2_{j,t} \right)^{1/2} \leqslant \rho^*_{d,\tau} := \mathfrak{m}_0 \frac{d^{1/2}\tau^{1/2}}{\mathfrak{m}_{\mathbf{\Lambda}}(\tau)}.
$$

So, $\mathcal{S}_{k,\tau} \subset \mathbb{B}_k(0, \rho^*_{d,\tau})$, and by the first step of the proof, there exists an ϵ-net $\mathcal{S}^\epsilon_{k,\tau}$ covering $\mathcal{S}_{k,\tau}$ and such that

$$|\mathcal{S}^\epsilon_{k,\tau}| \leqslant \left(\frac{9\rho^*_{d,\tau}}{\epsilon}\right)^{2k(d+\tau)} = \left(9\mathfrak{m}_0 \frac{d^{1/2}\tau^{1/2}}{\mathfrak{m}_{\mathbf{\Lambda}}(\tau)\epsilon}\right)^{2k(d+\tau)}.$$

By taking $\epsilon = 9\mathfrak{m}_0 d^{1/2}\tau^{1/2}\mathfrak{m}_{\mathbf{\Lambda}}(\tau)^{-1}n^{-2}$, thanks to Theorem 2.2.11.(2), with probability larger than $1-\alpha$,

$$\begin{aligned}
\|\widehat{\mathbf{\Theta}}_{k,\tau} - \mathbf{\Theta}^0\|^2_{\mathcal{F},\Pi} &\leqslant 3 \min_{\mathbf{T}\in\mathcal{S}_{k,\tau}} \|(\mathbf{T}-\mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} \\
&\quad + \frac{\mathfrak{c}_{2.2.11,1}}{n}\left[\log\left(\frac{2}{\alpha}\right) + 2k(d+\tau)\log\left(9\mathfrak{m}_0\frac{d^{1/2}\tau^{1/2}}{\mathfrak{m}_{\mathbf{\Lambda}}(\tau)\epsilon}\right)\right] \\
&\quad + \left[\mathfrak{c}_{2.2.11,2} + 8\mathfrak{m}_{\mathbf{\Lambda}}\mathfrak{c}_\xi \log\left(\frac{1}{\alpha}\right)\right]\epsilon \\
&= 3 \min_{\mathbf{T}\in\mathcal{S}_{k,\tau}} \|(\mathbf{T}-\mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} \\
&\quad + \frac{\mathfrak{c}_{2.2.11,1}}{n}\left[\log\left(\frac{2}{\alpha}\right) + 4k(d+\tau)\log(n)\right] \\
&\quad + 9\mathfrak{m}_0\frac{d^{1/2}\tau^{1/2}}{\mathfrak{m}_{\mathbf{\Lambda}}(\tau)n^2}\left[\mathfrak{c}_{2.2.11,2} + 8\mathfrak{m}_{\mathbf{\Lambda}}\mathfrak{c}_\xi\log\left(\frac{1}{\alpha}\right)\right].
\end{aligned}$$

Therefore, since $n \geqslant \max(d,\tau)$ and $\mathfrak{m}_{\mathbf{\Lambda}}(\tau) \geqslant 1$, with probability larger than $1-2\alpha$,

$$\begin{aligned}
\|\widehat{\mathbf{\Theta}}_{k,\tau} - \mathbf{\Theta}^0\|^2_{\mathcal{F},\Pi} &\leqslant 3 \min_{\mathbf{T}\in\mathcal{S}_{k,\tau}} \|(\mathbf{T}-\mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} \\
&\quad + (4\mathfrak{c}_{2.2.11,1} + 9\mathfrak{m}_0\mathfrak{c}_{2.2.11,2})k(d+\tau)\frac{\log(n)}{n} \\
&\quad + \frac{\mathfrak{c}_{2.2.11,1} + 72\mathfrak{m}_0\mathfrak{m}_{\mathbf{\Lambda}}\mathfrak{c}_\xi}{n}\log\left(\frac{2}{\alpha}\right).
\end{aligned}$$

Let us replace α by $\alpha/2$ to end the proof.

2.2.5.4 Proof of Theorem 2.2.6

Put $\overline{k} = 2^{\lfloor \log_2(k)\rfloor}$, and note that $k/2 \leqslant \overline{k} \leqslant k$. Fix $a > 0$ and define the set of matrices

$$\mathcal{A} = \left\{\mathbf{A} = (\mathbf{A}_{i,j})_{1\leqslant i\leqslant d\vee\tau, 1\leqslant j\leqslant\overline{k}} : \mathbf{A}_{i,j} \in \{0,a\}\right\}.$$

By Varshamov-Gilbert bound, there is a finite subset $\mathcal{B} \subset \mathcal{A}$ with $\mathrm{card}(\mathcal{B}) \geqslant 2^{\frac{\overline{k}(d\vee\tau)}{8}} + 1$, $0 \in \mathcal{B}$, and each pair $\mathbf{A} \neq \mathbf{A}'$ in $\mathcal{B}$ differ by at least $\overline{k}(d\vee\tau)$ coordinates. This implies

$$\|\mathbf{A}-\mathbf{A}'\|^2_{\mathcal{F}} \geqslant \frac{\overline{k}(d\vee\tau)}{8}a^2 \geqslant \frac{k(d\vee\tau)}{16}a^2.$$

For any $\mathbf{A}$, define by block $\overline{\mathbf{A}} = (\mathbf{A}|\mathbf{0})$ of dimension $(d\vee\tau)\times k$ (so the $\mathbf{0}$ has $k-\overline{k}$ columns). We then define $\widetilde{\mathbf{A}}$ of dimension $d\times\tau$. The construction differs depending on d and τ :

- If $d \geqslant \tau$,
$$\widetilde{\mathbf{A}} = (\mathbf{A}|\ldots|\mathbf{A}|\mathbf{0}).$$
- If $d < \tau$,
$$\widetilde{\mathbf{A}} = (\mathbf{A}|\ldots|\mathbf{A}|\mathbf{0})^*.$$

Note that this is clearly inspired by the construction in the proof of Theorem 5 in [17], however, here, we have to take care that, for a small enough, each $\widetilde{\mathbf{A}} \in \mathcal{A}$ is also in $\mathcal{M}_{d,k,\tau}$. In order to do so, we introduce the vectors in $\mathbb{R}^{\overline{k}}$:

$$\begin{aligned} v[1] &= \sqrt{\frac{1}{k}}(\underbrace{1 \quad \dots \quad 1}_{\overline{k}})^*, \\ v[2] &= \sqrt{\frac{1}{k}}(\underbrace{1 \quad \dots \quad 1}_{\overline{k}/2} \mid \underbrace{-1 \quad \dots \quad -1}_{\overline{k}/2})^*, \\ &\vdots \\ v[\overline{k}] &= \sqrt{\frac{1}{k}}(\underbrace{1 \quad -1 \quad \dots \quad 1 \quad -1}_{\overline{k}})^*. \end{aligned}$$

Now, remark that for $\mathbf{A} \in \mathcal{A}$ we have

$$\mathbf{A} = \underbrace{\sqrt{ak}\begin{pmatrix} v[1]^* \\ \hline \vdots \\ \hline v[\overline{k}]^* \end{pmatrix}}_{\mathbf{B}} \underbrace{\left(\sum_{i=1}^n v[i]\mathbf{1}_{\mathbf{A}_{i,1}\neq 0} \middle| \dots \middle| \sum_{i=1}^n v[i]\mathbf{1}_{\mathbf{A}_{i,k}\neq 0} \right)\sqrt{\frac{a}{k}}}_{\mathbf{C}}$$

and under this decomposition, it is clear that the entries of $\mathbf{B}$ and $\mathbf{C}$ are in $[0, \sqrt{a}]$. Playing with blocks, this gives trivially to a decomposition $\widetilde{\mathbf{A}} = \mathbf{UV}$ where $\mathbf{U}$ is $d \times k$, $\mathbf{V}$ is $k \times \tau$ and the entries of $\mathbf{U}$ and $\mathbf{V}$ are also in $[0, \sqrt{a}]$. In other words, $\widetilde{\mathbf{A}} \in \mathcal{M}_{d,k,\tau}$ holds as soon as $a \leqslant \mathfrak{m}_0/k$. Now, let $\mathbb{P}_{\mathbf{A}}$ be the data-generating distribution when $\mathbf{\Theta}^0 = \widetilde{\mathbf{A}}$ for $\mathbf{A} \in \mathcal{B}$, and KL be the Kullback-Leibler divergence. We have

$$\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_{\mathbf{A}}) = \frac{n}{2}\|\widetilde{\mathbf{A}}\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} \leqslant \frac{n}{2}a^2.$$

Thus, we look for a such that the condition

$$\frac{n}{2}a^2 \leqslant \alpha \log(\mathrm{card}(\mathcal{B}) - 1) = \frac{\alpha\overline{k}(d \vee \tau)}{8}$$

is satisfied for a given $0 < \alpha < 1/8$. Fix $\alpha = 1/16$. As $k \leqslant \overline{k}/2$, it's easy to check that

$$a = \frac{1}{8}\sqrt{\frac{k(d \vee \tau)}{2n}}$$

satisfies the condition. Also, remember that $\widetilde{\mathbf{A}} \in \mathcal{M}_{d,k,\tau}$ if $a \leqslant \mathfrak{m}_0/k$, which adds the condition

$$k \leqslant \left(\frac{256\mathfrak{m}_0^2 n}{d \vee \tau}\right)^{1/3}.$$

Theorem 2.5 in [97] then tells us that the rate is given by the minimal distance, for $\mathbf{A} \neq \mathbf{A}'$ in $\mathcal{B}$:

$$\begin{aligned} \|\widetilde{\mathbf{A}}\mathbf{\Lambda} - \widetilde{\mathbf{A}}'\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} = \frac{1}{d\tau}\|\widetilde{\mathbf{A}} - \widetilde{\mathbf{A}}'\|^2_{\mathcal{F}} &= \frac{1}{d\tau}\left\lfloor \frac{d \wedge \tau}{k} \right\rfloor \|\widetilde{\mathbf{A}} - \widetilde{\mathbf{A}}'\|^2_{\mathcal{F}} \\ &\geqslant \frac{1}{d\tau}\left\lfloor \frac{d \wedge \tau}{k} \right\rfloor \frac{k(d \vee \tau)}{16}a^2 \\ &\geqslant \frac{a^2}{32} = \frac{k(d \vee \tau)}{4096n}. \end{aligned}$$

2.2.5.5 Proof of Theorem 2.2.7

For any $k \in \mathcal{K}$, let $\mathcal{S}_k^\epsilon := \mathcal{S}_{k,\tau}^\epsilon$ be the ϵ-net introduced in the proof of Theorem 2.2.5, and recall that for $\epsilon = 9\mathfrak{m}_0 d^{1/2}\tau^{1/2}\mathfrak{m}_{\mathbf{\Lambda}}(\tau)^{-1}n^{-2}$,

$$|\mathcal{S}_k^\epsilon| \leqslant \left(9\mathfrak{m}_0 \frac{d^{1/2}\tau^{1/2}}{\mathfrak{m}_{\mathbf{\Lambda}}(\tau)\epsilon}\right)^{2k(d+\tau)} = n^{4k(d+\tau)}.$$

Then, for $\alpha \in (0,1)$ and $x_{k,\epsilon} := \lambda^{-1}\log(2\alpha^{-1}|\mathcal{K}|\cdot|\mathcal{S}_k^\epsilon|)$ with $\lambda = n\mathfrak{c}_{\text{pen}}^{-1} \in (0, n\lambda^*)$,

$$\begin{aligned}
4x_{k,\epsilon} - \text{pen}(k) &= \frac{4\mathfrak{c}_{\text{pen}}}{n}\log\left(\frac{2}{\alpha}|\mathcal{K}|\cdot|\mathcal{S}_k^\epsilon|\right) - 16\mathfrak{c}_{\text{pen}}\frac{\log(n)}{n}k(d+\tau) \\
&\leqslant \frac{4\mathfrak{c}_{\text{pen}}}{n}\left[4k(d+\tau)\log(n) + \log\left(\frac{2}{\alpha}|\mathcal{K}|\right)\right] \\
&\quad -16\mathfrak{c}_{\text{pen}}\frac{\log(n)}{n}k(d+\tau) \\
&\leqslant \frac{4\mathfrak{c}_{\text{pen}}}{n}\log\left(\frac{2}{\alpha}|\mathcal{K}|\right) =: \mathfrak{m}_n. \qquad (2.28)
\end{aligned}$$

Now, consider the event $\Omega_{\lambda,\epsilon} := (\Omega_{\lambda,\epsilon}^-)^c \cap (\Omega_{\lambda,\epsilon}^+)^c$ with

$$\Omega_{\lambda,\epsilon}^- := \bigcup_{k\in\mathcal{K}}\bigcup_{\mathbf{T}\in\mathcal{S}_k^\epsilon}\Omega_{x_{k,\epsilon},\lambda,\mathcal{S}_k^\epsilon}^-(\mathbf{T}) \text{ and } \Omega_{\lambda,\epsilon}^+ := \bigcup_{k\in\mathcal{K}}\bigcup_{\mathbf{T}\in\mathcal{S}_k^\epsilon}\Omega_{x_{k,\epsilon},\lambda,\mathcal{S}_k^\epsilon}^+(\mathbf{T}).$$

So,

$$\begin{aligned}
\mathbb{P}(\Omega_{\lambda,\epsilon}^c) &\leqslant \sum_{k\in\mathcal{K}}\sum_{\mathbf{T}\in\mathcal{S}_k^\epsilon}\left[\mathbb{P}(\Omega_{x_{k,\epsilon},\lambda,\mathcal{S}_k^\epsilon}^-(\mathbf{T})) + \mathbb{P}(\Omega_{x_{k,\epsilon},\lambda,\mathcal{S}_k^\epsilon}^+(\mathbf{T}))\right] \\
&\leqslant 2\sum_{k\in\mathcal{K}}|\mathcal{S}_k^\epsilon|e^{-\lambda x_{k,\epsilon}} = \alpha \qquad (2.29)
\end{aligned}$$

and $\Omega_{x_{\widehat{k},\epsilon},\lambda,\mathcal{S}_{\widehat{k}}^\epsilon}(\widehat{\mathbf{T}}_{\widehat{k}}^\epsilon) \subset \Omega_{\lambda,\epsilon}$, where $\widehat{\mathbf{T}}_k^\epsilon$ is a solution of the minimization problem (2.24) for every $k \in \mathcal{K}$.

On the event $\Omega_{\lambda,\epsilon}$, by the definition of $\widehat{k}$, and thanks to Inequalities (2.22), (2.23) and (2.24),

$$\begin{aligned}
\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^0\|_{\mathcal{F},\Pi}^2 &\leqslant \|(\widehat{\mathbf{T}}_{\widehat{k}}^\epsilon - \mathbf{T}^0)\mathbf{\Lambda}\|_{\mathcal{F},\Pi}^2 + \mathfrak{c}_2\epsilon \\
&\leqslant \left(1 - \mathfrak{c}_{2.2.8}\frac{\lambda}{n}\right)^{-1}(r_n(\widehat{\mathbf{T}}_{\widehat{k}}^\epsilon\mathbf{\Lambda}) - r_n(\mathbf{T}^0\mathbf{\Lambda}) + 4x_{\widehat{k},\epsilon}) + \mathfrak{c}_2\epsilon \\
&\leqslant \left(1 - \mathfrak{c}_{2.2.8}\frac{\lambda}{n}\right)^{-1}\Big(r_n(\widehat{\mathbf{T}}_{\widehat{k}}\mathbf{\Lambda}) - r_n(\mathbf{T}^0\mathbf{\Lambda}) \\
&\qquad +\mathfrak{c}_1(\xi_1,\ldots,\xi_n)\epsilon + 4x_{\widehat{k},\epsilon}\Big) + \mathfrak{c}_2\epsilon \\
&= \left(1 - \mathfrak{c}_{2.2.8}\frac{\lambda}{n}\right)^{-1}\Big(\min_{k\in\mathcal{K}}\{r_n(\widehat{\mathbf{T}}_k\mathbf{\Lambda}) - r_n(\mathbf{T}^0\mathbf{\Lambda}) + \text{pen}(k)\} \\
&\qquad +\mathfrak{c}_1(\xi_1,\ldots,\xi_n)\epsilon + 4x_{\widehat{k},\epsilon} - \text{pen}(\widehat{k})\Big) + \mathfrak{c}_2\epsilon \\
&\leqslant \frac{1}{1 - \mathfrak{c}_{2.2.8}\lambda n^{-1}}\min_{k\in\mathcal{K}}\Big\{(1 + \mathfrak{c}_{2.2.8}\lambda n^{-1})\|(\widehat{\mathbf{T}}_k - \mathbf{T}^0)\mathbf{\Lambda}\|_{\mathcal{F},\Pi}^2
\end{aligned}$$

$$
\begin{aligned}
&\qquad +4x_{k,\epsilon} + \mathrm{pen}(k)\Big\} + \frac{\mathfrak{m}_n + \mathfrak{c}_1(\xi_1,\dots,\xi_n)\epsilon}{1 - \mathfrak{c}_{2.2.8}\lambda n^{-1}} + \mathfrak{c}_2\epsilon \\
&\leqslant \; 2\min_{k\in\mathcal{K}}\{3/2\|(\widehat{\mathbf{T}}_k - \mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} + 2\mathrm{pen}(k)\} \\
&\qquad +4\mathfrak{m}_n + (2\mathfrak{c}_1(\xi_1,\dots,\xi_n) + \mathfrak{c}_2)\epsilon \qquad (2.30)
\end{aligned}
$$

with

$$
\mathfrak{c}_1(\xi_1,\dots,\xi_n) := 2\mathfrak{m}_{\mathbf{\Lambda}}\left(\frac{1}{n}\sum_{i=1}^n |\xi_i| + \mathfrak{m}_\varepsilon + 2\mathfrak{m}_0\right) \text{ and } \mathfrak{c}_2 = 4\mathfrak{m}_0\mathfrak{m}_{\mathbf{\Lambda}}.
$$

Moreover, by following the proof of Theorem 2.2.11 and Theorem 2.2.5 on the same event $\Omega_{\lambda,\epsilon}$,

$$
\begin{aligned}
&\|(\widehat{\mathbf{T}}_k - \mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} \\
&\qquad\qquad \leqslant 3\min_{\mathbf{T}\in\mathcal{S}_k}\|(\mathbf{T} - \mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} + \mathfrak{c}_{2.2.5}\left[k(d+\tau)\frac{\log(n)}{n} + \frac{1}{n}\log\left(\frac{2}{\alpha}|\mathcal{K}|\right)\right]
\end{aligned}
$$

for every $k \in \mathcal{K}$. Therefore, thanks to (2.26), (2.27) and (2.30), with probability larger than $1 - 2\alpha$,

$$
\begin{aligned}
&\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^0\|^2_{\mathcal{F},\Pi} \\
&\leqslant 4\min_{k\in\mathcal{K}}\left\{3\min_{\mathbf{T}\in\mathcal{S}_k}\|(\mathbf{T} - \mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} + (\mathfrak{c}_{2.2.5} + 16\mathfrak{c}_{\mathrm{pen}})k(d+\tau)\frac{\log(n)}{n}\right\} \\
&+ \frac{4\mathfrak{c}_{2.2.5} + 16\mathfrak{c}_{\mathrm{pen}}}{n}\log\left(\frac{2}{\alpha}|\mathcal{K}|\right) + 9\mathfrak{m}_0\frac{d^{1/2}\tau^{1/2}}{\mathfrak{m}_{\mathbf{\Lambda}}(\tau)n^2}\left[\mathfrak{c}_{2.2.11}{}_2 + 8\mathfrak{m}_{\mathbf{\Lambda}}\mathfrak{c}_\xi\log\left(\frac{1}{\alpha}\right)\right].
\end{aligned}
$$

To end the proof, let us replace α by $\alpha/2$ and note that $d^{1/2}\tau^{1/2}/(\mathfrak{m}_{\mathbf{\Lambda}}(\tau)n^2) \leqslant 1/n$ because $n \geqslant \max(d,\tau)$ and $\mathfrak{m}_{\mathbf{\Lambda}}(\tau) \geqslant 1$.

Chapitre 3

Nadaraya-Watson Estimator for i.i.d. Paths of Diffusions Processes

Let us start the chapter by recalling the model and the definitions introduced in the introduction. Consider the stochastic differential equation

$$X_t = x_0 + \int_0^t b(X_s)ds + \int_0^t \sigma(X_s)dW_s, \tag{3.1}$$

where $b, \sigma : \mathbb{R} \to \mathbb{R}$ are two continuous functions and $W = (W_t)_{t\in[0,T]}$ is a Brownian motion. Let $\mathcal{I} : (x, w) \mapsto \mathcal{I}(x, w)$ be the Itô map for Equation (3.1) and, for $N \in \mathbb{N}^*$ copies $W^1, \ldots, W^N$ of W, consider $X^i = \mathcal{I}(x_0, W^i)$ for every $i \in \{1, \ldots, N\}$.

As explained in the introduction, the estimation of the drift function b from continuous-time and discrete-time observations of $(X^1, \ldots, X^N)$ is a functional data analysis problem already investigated in the parametric framework (see Ditlevsen and De Gaetano [114], Overgaard et al. [125], Picchini, De Gaetano and Ditlevsen [126], Picchini and Ditlevsen [127], Comte, Genon-Catalot and Samson [105], Delattre and Lavielle [110], Delattre, Genon-Catalot and Samson [109], Dion and Genon-Catalot [113], Delattre, Genon-Catalot and Larédo [108], etc.) and more recently in the nonparametric framework (see Comte and Genon-Catalot [102, 103], Della Maestra and Hoffmann [111] and Denis et al. [112]). In [111], the authors also study a Nadaraya-Watson type estimator, presented bellow, of the drift function in McKean-Vlasov models.

Under the appropriate conditions on b and σ recalled at Section 3.1, the distribution of X_t has a density $p_t(x_0, .)$ for every $t \in (0, T]$, and then one can define

$$f(x) := \frac{1}{T - t_0} \int_{t_0}^T p_t(x_0, x)dt \ ; \ x \in \mathbb{R}$$

for any $t_0 > 0$. Clearly, f is a density function :

$$\int_{-\infty}^{\infty} f(x)dx = \frac{1}{T - t_0} \int_{t_0}^T \int_{-\infty}^{\infty} p_t(x_0, x)dxdt = 1.$$

Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel (i.e. an integrable function such that $\int K = 1$) and consider $K_h(x) := h^{-1}K(h^{-1}x)$ with $h \in (0, 1]$. In the spirit of Comte and Genon-Catalot [102, 103], our paper deals first with the continuous-time Nadaraya-Watson estimator

$$\widehat{b}_{N,h}(x) := \frac{\widehat{bf}_{N,h}(x)}{\widehat{f}_{N,h}(x)} \tag{3.2}$$

of the drift function b, where

$$\widehat{f}_{N,h}(x) := \frac{1}{N(T-t_0)} \sum_{i=1}^{N} \int_{t_0}^{T} K_h(X_t^i - x)dt \tag{3.3}$$

is an estimator of f and

$$\widehat{bf}_{N,h}(x) := \frac{1}{N(T-t_0)} \sum_{i=1}^{N} \int_{t_0}^{T} K_h(X_t^i - x)dX_t^i \tag{3.4}$$

is an estimator of bf. From independent copies of X continuously observed on $[0,T]$, $\widehat{b}_{N,h}$ is a natural extension of the Nadaraya-Watson estimator already well-studied in the regression framework (see Comte [52], Chapter 4 or Györfi et al. [116], Chapter 5). The paper also deals with a discrete-time approximate of the previous Nadaraya-Watson estimator :

$$\widehat{b}_{n,N,h}(x) := \frac{\widehat{bf}_{n,N,h}(x)}{\widehat{f}_{n,N,h}(x)}, \tag{3.5}$$

where

$$\widehat{f}_{n,N,h}(x) := \frac{1}{nN} \sum_{i=1}^{N} \sum_{j=0}^{n-1} K_h(X_{t_j}^i - x) \tag{3.6}$$

is an estimator of f,

$$\widehat{bf}_{n,N,h}(x) := \frac{1}{N(T-t_0)} \sum_{i=1}^{N} \sum_{j=0}^{n-1} K_h(X_{t_j}^i - x)(X_{t_{j+1}}^i - X_{t_j}^i) \tag{3.7}$$

is an estimator of bf, and $(t_0, t_1, \ldots, t_n)$ is the dissection of $[t_0, T]$ such that $t_j = t_0 + (T-t_0)j/n$ for every $j \in \{1, \ldots, n\}$. Finally, our paper deals with a risk bound on the adaptive double bandwidths Nadaraya-Watson's estimator

$$\widehat{b}_{N,\widehat{h},\widehat{h}'}(x) := \frac{\widehat{bf}_{N,\widehat{h}}(x)}{\widehat{f}_{N,\widehat{h}'}(x)},$$

where $\widehat{h}$ (resp. $\widehat{h}'$) is selected via a penalized comparison to overfitting (PCO) type criterion for its numerator (resp. denominator). However, in the nonparametric regression framework, it is established in Comte and Marie [106] that the leave-one-out cross-validation (looCV) bandwidth selection method for Nadaraya-Watson's estimator is numerically more satisfactory than two alternative procedures based on Goldenshluger-Lepski's method and on the PCO method. For this reason, an extension of the looCV method to $\widehat{b}_{n,N,h}$ is also provided, with numerical experiments, even if it seems difficult to establish a risk bound on the associated adaptive estimator.

Now, let us compare $\widehat{b}_{N,h}$ with the estimator of Della Maestra and Hoffmann [111] restricted to our framework :

$$\widehat{b}_{N,\mathbf{h}}(x) := \frac{\displaystyle\sum_{i=1}^{N} \int_0^T L_{h_1}(\tau - t) K_{h_2}(X_t^i - x) dX_t^i}{\displaystyle\sum_{i=1}^{N} K_{h_3}(X_\tau^i - x)},$$

where $L : \mathbb{R} \to \mathbb{R}$ is another kernel, $h_1, h_2, h_3 \in (0,1]$, $\mathbf{h} = (h_1, h_2, h_3)$ and $\tau \in (0,T)$. In [111], the authors provide a nice risk bound on the adaptive estimator obtained by selecting (h_1, h_2) (resp. h_3) via a Goldenshluger-Lepski type procedure on the numerator (resp. the denominator) of $\widehat{b}_{N,\mathbf{h}}$. As mentioned above, in Comte and Marie [106], it has been established that in the nonparametric regression framework, this approach is numerically less satisfactory than the looCV method. However, the looCV method provided in our paper doesn't extend to $\widehat{b}_{N,\mathbf{h}}$ because it cannot be written easily as a linear combination. Note also that even if it is numerically less satisfactory than the looCV method provided at Subsection 3.4.2, the PCO type method provided in our paper at Subsection 3.4.1 is easier to implement and numerically faster than a Goldenshluger-Lepski type method because, as in the nonparametric regression framework, the criterion to minimize depends on one variable instead of two, and because there is no constant to calibrate. For technical reasons explained at Section 3.5, the condition $(Nh^3)^{-1} \leqslant 1$ is required on the bandwidths collection to establish a risk bound on our PCO based adaptive estimator of bf, when Della Maestra and Hoffmann only need the condition $\log(N)^2(Nh)^{-1} \leqslant 1$ to establish a risk bound on their Goldenshluger-Lepski based adaptive estimator of $bp_\tau(x_0,.)$ in [111]. However, Remark 3.4.4 explains why the condition $(Nh^3)^{-1} \leqslant 1$ on the bandwidths collection is not that uncomfortable. Finally, under similar conditions on b, σ and K, the rate of convergence of our continuous-time Nadaraya-Watson estimator is of same order than the rate of convergence of the estimator of Della Maestra and Hoffmann [111] in the nonadaptive case. There is no discrete-time approximate of $\widehat{b}_{N,\mathbf{h}}$ studied in [111].

Finally, even if they deal with a different type of nonparametric estimators, let us say few words on the recent papers of Comte and Genon-Catalot [102] and Denis et al. [112]. On the one hand, in [102], the authors extend to the diffusion processes framework, for continuous-time observations, the least squares projection estimator already well studied in the regression framework (see Cohen et al. [100], Comte and Genon-Catalot [101], etc.). In particular, they provide a model selection procedure and establish a risk bound on the associated adaptive estimator. As explained at Section 3.2, in the nonadaptive case, the variance term of their estimator is comparable with the variance term of $\widehat{b}_{N,h}$, but as in the nonparametric regression framework, the rate of convergence of the least squares projection estimator depends on the regularity space associated to the projection basis. On the other hand, in [112], the authors focus on a projection least squares estimator computed from discrete-time observations and on a B-spline space. They provide a model selection procedure and prove both upper and lower bounds on the associated adaptive estimator.

This chapter is organized as follow. Section 3.1 deals with the existence and the regularity of the density $p_t(x_0,.)$ of X_t for every $t \in (0,T]$, and with a Nikol'skii type condition fulfilled by f. Section 3.2 deals with a risk bound on the continuous-time Nadaraya-Watson estimator and Section 3.3 with a risk bound on its discrete-time approximate. Finally, Section 3.4 provides extensions of the PCO and looCV methods for the Nadaraya-Watson estimator studied in this paper. Some numerical experiments on the looCV based adaptive Nadaraya-Watson estimator are also provided. The proofs are postponed to Appendix 3.6.

Notations and basic definitions :

- For every $A, B \in \mathbb{R}$ such that $A < B$, $C^0([A,B];\mathbb{R})$ is equipped with the uniform norm $\|.\|_{\infty,A,B}$, and $C^0(\mathbb{R})$ is equipped with the uniform (semi-)norm $\|.\|_\infty$.

- For every $p \in \overline{\mathbb{N}}$, $C_b^p(\mathbb{R}) := \cap_{j=0}^p \{\varphi \in C^p(\mathbb{R}) : \varphi^{(j)} \text{ is bounded}\}$.
- For every $p \geqslant 1$, $\mathbb{L}^p(\mathbb{R}, dx)$ is equipped with its usual norm $\|.\|_p$ such that

$$\|\varphi\|_p := \left(\int_{-\infty}^{\infty} \varphi(x)^p dx\right)^{1/p} ; \forall \varphi \in \mathbb{L}^p(\mathbb{R}, dx).$$

- $\mathbb{H}^2$ is the space of the processes $(Y_t)_{t\in[0,T]}$, adapted to the filtration generated by W, such that

$$\int_0^T \mathbb{E}(Y_t^2) dt < \infty.$$

- For a given kernel δ, the usual scalar product on $\mathbb{L}^2(\mathbb{R}, \delta(x)dx)$ is denoted by $\langle ., . \rangle_{2,\delta}$ and the associated norm by $\|.\|_{2,\delta}$.

3.1 Preliminaries : regularity of the density and estimates

This section deals with the existence and the regularity of the density $p_t(x_0, .)$ of X_t for every $t \in (0, T]$, with the Kusuoka-Stroock bounds on $(t, x) \mapsto p_t(x_0, x)$ and its derivatives, and then with a Nikol'skii type condition fulfilled by f.

In the sequel, in order to ensure the existence and the uniqueness of the (strong) solution to Equation (3.1), b and σ fulfill the following regularity assumption.

Assumption 3.1.1. — *The functions b and σ are Lipschitz continuous.*

Now, assume that the solution X to Equation (3.1) fulfills the following assumption.

Assumption 3.1.2. — *There exists $\beta \in \mathbb{N}^*$ such that, for any $t \in (0, T]$, the distribution of X_t has a β times continuously derivable density $p_t(x_0, .)$. Moreover, for every $x \in \mathbb{R}$,*

$$0 < p_t(x_0, x) \leqslant \frac{\mathfrak{c}_{3.1.2,1}}{t^{1/2}} \exp\left[-\mathfrak{m}_{3.1.2,1} \frac{(x - x_0)^2}{t}\right]$$

and

$$|\partial_x^\ell p_t(x_0, x)| \leqslant \frac{\mathfrak{c}_{3.1.2,2}(\ell)}{t^{q_2(\ell)}} \exp\left[-\mathfrak{m}_{3.1.2,2}(\ell) \frac{(x - x_0)^2}{t}\right] ; \forall \ell \in \{1, \ldots, \beta\},$$

where all the constants are positive, depend on T, but not on t and x.

At Section 3.3, the following assumption on X is also required.

Assumption 3.1.3. — *For any $x \in \mathbb{R}$, the function $t \in (0, T] \mapsto p_t(x_0, x)$ is continuously derivable. Moreover,*

$$|\partial_t p_t(x_0, x)| \leqslant \frac{\mathfrak{c}_{3.1.3,3}}{t^{q_3}} \exp\left[-\mathfrak{m}_{3.1.3,3} \frac{(x - x_0)^2}{t}\right] ; \forall t \in (0, T],$$

where $\mathfrak{c}_{3.1.3,3}$, $\mathfrak{m}_{3.1.3,3}$ and q_3 are three positive constants depending on T but not on t and x.

Let us provide some examples of diffusion processes categories satisfying Assumptions 3.1.2 and/or 3.1.3.

Examples :

(1) Assume that the functions b and σ belong to $C_{\mathrm{b}}^{\infty}(\mathbb{R})$, and that there exists $\alpha > 0$ such that

$$|\sigma(x)| > \alpha \ ; \ \forall x \in \mathbb{R}. \tag{3.8}$$

Then, by Kusuoka and Stroock [121], Corollary 3.25, X fulfills Assumptions 3.1.2 and 3.1.3.

(2) Assume that b is Lipschitz continuous (but not bounded) and that $\sigma \in C_{\mathrm{b}}^{1}(\mathbb{R})$. Assume also that σ satisfies the non-degeneracy condition (3.8) and that σ' is Hölder continuous. Then, by Menozzi et al. [124], Theorem 1.2, X fulfills Assumption 3.1.2 with $\beta = 1$ (but not necessarily Assumption 3.1.3). Note that the conditions required to apply Menozzi et al. [124], Theorem 1.2 are fulfilled by the so-called Ornstein-Uhlenbeck process, that is the solution to the Langevin equation :

$$X_t = x_0 - \theta \int_0^t X_s ds + \sigma W_t \ ; \ t \in \mathbb{R}_+, \tag{3.9}$$

where $\theta, \sigma > 0$ and $x_0 \in \mathbb{R}_+$. In this special case, since it is well-known that the solution to Equation (3.9) is a Gaussian process such that

$$\mathbb{E}(X_t) = x_0 e^{-\theta t} \quad \text{and} \quad \mathrm{var}(X_t) = \frac{\sigma^2}{2\theta}(1 - e^{-2\theta t}) \ ; \ \forall t \in [0, T],$$

one can show that X also fulfills Assumption 3.1.3.

Remark 3.1.4. — *Under Assumptions 3.1.1 and 3.1.2, for any $p \geqslant 1$ and any continuous function $\varphi : \mathbb{R} \to \mathbb{R}$ having polynomial growth, $t \in [0, T] \mapsto \mathbb{E}(|\varphi(X_t)|^p)$ is bounded. Indeed, for any $t \in [0, T]$,*

$$\begin{aligned} \mathbb{E}(|\varphi(X_t)|^p) &\leqslant \mathfrak{c}_1(1 + \mathbb{E}(|X_t|^{pq})) = \mathfrak{c}_1 \int_{-\infty}^{\infty} (1 + |x|^{pq}) p_t(x_0, x) dx \\ &\leqslant \mathfrak{c}_1 \mathfrak{c}_{3.1.2,1} \int_{-\infty}^{\infty} (1 + |t^{1/2}x + x_0|^{pq}) e^{-\mathfrak{m}_{3.1.2,1} x^2} dx \leqslant \mathfrak{c}_2(1 \vee T^{pq/2}) \end{aligned}$$

where

$$\mathfrak{c}_2 = \mathfrak{c}_1 \mathfrak{c}_{3.1.2,1} \int_{-\infty}^{\infty} [1 + (|x| + |x_0|)^{pq}] e^{-\mathfrak{m}_{3.1.2,1} x^2} dx$$

and the constants $\mathfrak{c}_1, q > 0$ only depend on φ. Moreover,

$$\begin{aligned} \|\varphi\|_{p,f} &:= \int_{-\infty}^{\infty} |\varphi(x)|^p f(x) dx \\ &= \frac{1}{T - t_0} \int_{t_0}^{T} \mathbb{E}(|\varphi(X_t)|^p) dt \leqslant \mathfrak{c}_2(1 \vee T^{pq/2}). \end{aligned}$$

Then, $\varphi \in \mathbb{L}^p(\mathbb{R}, f(x)dx)$ and $\|\varphi\|_{p,f}$ is bounded by a constant which doesn't depend on t_0. In particular, the remark applies to b and σ with $q = 1$ by Assumption 3.1.1.

Finally, let us show that f fulfills a Nikol'skii type condition.

Corollary 3.1.5. — *Under Assumption 3.1.1, $f(x) > 0$ for every $x \in \mathbb{R}$. Moreover, under Assumptions 3.1.1 and 3.1.2, there exists $\mathfrak{c}_{3.1.5} > 0$, depending on T but not on t_0, such that for every $\ell \in \{0, \ldots, \beta - 1\}$ and $\theta \in \mathbb{R}$,*

$$\int_{-\infty}^{\infty} [f^{(\ell)}(x + \theta) - f^{(\ell)}(x)]^2 dx \leqslant \frac{\mathfrak{c}_{3.1.5}}{t_0^{2q_2(\ell+1)}} (\theta^2 + |\theta|^3).$$

Remark 3.1.6. — *Assumption 3.1.2, Assumption 3.1.3 and Corollary 3.1.5 are crucial in the sequel, but t_0 has to be chosen carefully to get reasonable risk bounds on the estimators $\widehat{b}_{N,h}$ and $\widehat{b}_{n,N,h}$. Indeed, the behavior of the Kusuoka-Stroock bounds on $(t,x) \mapsto p_t(x_0,x)$ and its derivatives is singular at point $(0,x_0)$. This is due to the fact that the distribution of X at time 0 is a Dirac measure while that it has a smooth density with respect to Lebesgue's measure for every $t \in (0,T]$. Moreover, since X is not a stationary process in general, the Kusuoka-Stroock bounds on $(t,x) \mapsto p_t(x_0,x)$ and its derivatives explode when $T \to \infty$. The same difficulty appears with the estimators studied in Comte and Genon-Catalot [102] and in Della Maestra and Hoffmann [111]. So, it is recommended to take T as small as possible in practice. In the sequel, only the dependence in t_0 is tracked in the risk bounds derived from Assumption 3.1.2, Assumption 3.1.3 and Corollary 3.1.5 because it is specific to our approach. Finally, these risk bounds only depend on t_0 through a multiplicative constant of order $1/\min\{t_0^\alpha, T-t_0\}$ ($\alpha > 0$). So, to take $t_0 \in [1, T-1]$ when $T > 1$ gives constants not depending on t_0.*

3.2 Risk bound on the continuous-time Nadaraya-Watson estimator

This section deals with risk bounds on $\widehat{f}_{N,h}$, $\widehat{bf}_{N,h}$, and then on the Nadaraya-Watson estimator $\widehat{b}_{N,h}$.

In the sequel, the kernel K fulfills the following usual assumptions.

Assumption 3.2.1. — *The kernel K is symmetric, continuous and belongs to $\mathbb{L}^2(\mathbb{R}, dx)$.*

Assumption 3.2.2. — *There exists $\upsilon \in \mathbb{N}^*$ such that*

$$\int_{-\infty}^{\infty} |z^{\upsilon+1} K(z)| dz < \infty \quad \text{and} \quad \int_{-\infty}^{\infty} z^\ell K(z) dz = 0 \ ; \ \forall \ell \in \{1, \dots, \upsilon\}.$$

About the construction of kernels fulfilling both Assumptions 3.2.1 and 3.2.2, the reader can refer to Comte [52], Proposition 2.10. The following proposition provides a risk bound on $\widehat{f}_{N,h}$ (see (3.3)).

Proposition 3.2.3. — *Under Assumptions 3.1.1, 3.1.2, 3.2.1 and 3.2.2 with $\upsilon = \beta$,*

$$\mathbb{E}(\|\widehat{f}_{N,h} - f\|_2^2) \leqslant \mathfrak{c}_{3.2.3}(t_0) h^{2\beta} + \frac{\|K\|_2^2}{Nh}$$

with

$$\mathfrak{c}_{3.2.3}(t_0) = \frac{\mathfrak{c}_{3.1.5}}{|(\beta-2)!|^2 t_0^{2q_2(\beta)}} \left(\int_{-\infty}^{\infty} |z|^\beta (1+|z|^{1/2}) |K(z)| dz \right)^2.$$

Note that thanks to Proposition 3.2.3, the bias-variance tradeoff is reached by (the risk bound on) $\widehat{f}_{N,h}$ when h is of order $N^{-1/(2\beta+1)}$, leading to a rate of order $N^{-2\beta/(2\beta+1)}$. Moreover, by Remark 3.1.6, to take $t_0 \geqslant 1$ when $T > 1$ gives

$$\mathbb{E}(\|\widehat{f}_{N,h} - f\|_2^2) \leqslant \mathfrak{c}_{3.2.3} h^{2\beta} + \frac{\|K\|_2^2}{Nh} \quad \text{with} \quad \mathfrak{c}_{3.2.3} = \frac{\mathfrak{c}_{3.1.5}}{|(\beta-2)!|^2} \left(\int_{-\infty}^{\infty} |z|^\beta (1+|z|^{1/2}) |K(z)| dz \right)^2.$$

Note also that in the risk bound on $\widehat{f}_{N,h}$ of Proposition 3.2.3, only the control of the bias term depends on T, trough the constant $\mathfrak{c}_{3.1.5}$, depending itself on the constants

$\mathfrak{c}_{3.1.2,2}(\ell)$, $\ell \in \{1,\dots,\beta\}$, involved in the Kusuoka-Strook bounds (see Assumption 3.1.2). Indeed, except in the special case of the Ornstein-Uhlenbeck process which is stationary, for all the examples of diffusion processes fulfilling Assumption 3.1.2 (see Kusuoka and Stroock [121] and Menozzi et al. [124]), the constants $\mathfrak{c}_{3.1.2,2}(\ell)$, $\ell \in \{1,\dots,\beta\}$, depend on T. The variance term doesn't depend on time at all.

In the sequel, $\mathbb{L}^2(\mathbb{R}, f(x)dx)$ is equipped with the f-weighted norm $\|.\|_{2,f}$ defined at the end of the introduction section. Let us recall that by Remark 3.1.4, for every $\varphi \in \mathbb{L}^2(\mathbb{R}, f(x)dx)$, $\|\varphi\|_{2,f}$ is bounded by a constant which doesn't depend on t_0.

The following proposition provides a risk bound on $\widehat{bf}_{N,h}$ (see (3.4)).

Proposition 3.2.4. — *Under Assumptions 3.1.1 and 3.2.1,*

$$\mathbb{E}(\|\widehat{bf}_{N,h} - bf\|_2^2) \leqslant \|(bf)_h - bf\|_2^2 + \frac{\mathfrak{c}_{3.2.4}(t_0)}{Nh}$$

with $(bf)_h := K_h * (bf)$ *and*

$$\mathfrak{c}_{3.2.4}(t_0) = 2\|K\|_2^2 \left(\|b\|_{2,f}^2 + \frac{1}{T-t_0}\|\sigma\|_{2,f}^2 \right).$$

Assume that bf is $\gamma \in \mathbb{N}^*$ times continuously derivable and that there exists $\varphi \in \mathbb{L}^1(\mathbb{R}, |z|^{\gamma-1}K(z)dz)$ such that, for every $\theta \in \mathbb{R}$ and $h \in (0,1]$,

$$\int_{-\infty}^{\infty} [(bf)^{(\gamma-1)}(x+h\theta) - (bf)^{(\gamma-1)}(x)]^2 dx \leqslant \varphi(\theta)h^2. \tag{3.10}$$

If in addition K fulfills Assumption 3.2.2 with $\upsilon = \gamma$, then $\|(bf)_h - bf\|_2^2$ is of order $h^{2\gamma}$, and by Proposition 3.2.4, the bias-variance tradeoff is reached by $\widehat{bf}_{N,h}$ when h is of order $N^{-1/(2\gamma+1)}$, leading to the rate $N^{-2\gamma/(2\gamma+1)}$. Moreover, by Remark 3.1.6, to take $t_0 \leqslant T-1$ when $T > 1$ gives

$$\mathbb{E}(\|\widehat{bf}_{N,h} - bf\|_2^2) \leqslant \|(bf)_h - bf\|_2^2 + \frac{\mathfrak{c}_{3.2.4}}{Nh} \quad \text{with} \quad \mathfrak{c}_{3.2.4} = 2\|K\|_2^2(\|b\|_{2,f}^2 + \|\sigma\|_{2,f}^2).$$

Note also that the variance term in this risk bound doesn't depend on T.

Finally, Propositions 3.2.3 and 3.2.4 allow to provide a risk bound on a truncated version of the Nadaraya-Watson estimator $\widehat{b}_{N,h}$ (see (3.2)).

Proposition 3.2.5. — *Consider the 2 bandwidths (truncated) Nadaraya-Watson (2bNW) estimator*

$$\widehat{b}_{N,h,h'}(x) := \frac{\widehat{bf}_{N,h}(x)}{\widehat{f}_{N,h'}(x)} \mathbf{1}_{\widehat{f}_{N,h'}(x) > m/2} \quad \textit{with} \quad h, h' > 0,$$

and assume that $f(x) > m > 0$ *for every* $x \in [A,B]$ *(*$m \in (0,1]$ *and* $A, B \in \mathbb{R}$ *such that* $A < B$*). Under Assumptions 3.1.1, 3.1.2, 3.2.1 and 3.2.2 with* $\upsilon = \beta$*,*

$$\mathbb{E}(\|\widehat{b}_{N,h,h'} - b\|_{f,A,B}^2) \leqslant \frac{\mathfrak{c}_{3.2.5}}{m^2} \left[\|(bf)_h - bf\|_2^2 + \frac{\mathfrak{c}_{3.2.4}(t_0)}{Nh} + 2\|b\|_{2,f}^2 \left(\mathfrak{c}_{3.2.3}(t_0)(h')^{2\beta} + \frac{\|K\|_2^2}{Nh'} \right) \right]$$

with $\mathfrak{c}_{3.2.5} := 8(\|f\|_\infty \vee \|b^2 f\|_\infty)$ *and* $\|\varphi\|_{f,A,B} := \|\varphi \mathbf{1}_{[A,B]}\|_{2,f}$ *for every* $\varphi \in \mathbb{L}^2(\mathbb{R}, f(x)dx)$.

Proposition 3.2.5 says that the risk of $\widehat{b}_{N,h,h'}$ can be controlled by the sum of those of $\widehat{bf}_{N,h}$ and $\widehat{f}_{N,h'}$ up to a multiplicative constant. Now, if K fulfills Assumption 3.2.2 with $\upsilon = \beta \vee \gamma$, and if bf satisfies Condition (3.10), then the risk bound on $\widehat{b}_{N,h,h'}$ is of order $h^{2\gamma} + (h')^{2\beta} + 1/(Nh) + 1/(Nh')$, and the bias-variance tradeoff is reached when h (resp. h') is of order $N^{-1/(2\gamma+1)}$ (resp. $N^{-1/(2\beta+1)}$), leading to the rate

$$N^{-2\left[\left(\frac{\gamma}{2\gamma+1}\right)\wedge\left(\frac{\beta}{2\beta+1}\right)\right]} = N^{-\frac{2(\beta\wedge\gamma)}{2(\beta\wedge\gamma)+1}},$$

which is of same order than the rate of the nonadaptive version of the estimator of Della Maestra and Hoffmann [111] (see their Theorem 15). Note also that to consider the 2bNW estimator is crucial to extend the PCO method to our framework in the spirit of Comte and Marie [106] (see Subsection 3.4.1). However, by taking $h = h'$ of order $N^{-1/(2(\beta\wedge\gamma)+1)}$, the bias-variance tradeoff is reached by the 1 bandwidth (truncated) Nadaraya-Watson estimator with the same rate. Finally, if $h = h'$ and σ is bounded, then the variance term in the risk bound of Proposition 3.2.5 is comparable to the variance term in the risk bound obtained by Comte and Genon-Catalot in [102] for their least squares projection estimator (see [102], Propositions 2.1 and 2.2). Indeed, for a d-dimensional projection space, the variance term in the risk bound of Comte and Genon-Catalot [102], Proposition 2.1 is of order d/N which is comparable to $1/(Nh)$. The rate of convergence of their least squares projection estimator depends on the regularity space associated to the projection basis but, as in the nonparametric regression framework, not on the regularity of f.
The limitation of our Proposition 3.2.5 is that m is unknown in general and must be replaced by an estimator as well. Most of the time, as stated in Comte [52], Chapter 4, the minimum of an estimator of f is taken to choose m in practice :

$$\widehat{m}_{N,h'} = \min\{\widehat{f}_{N,h'}(x)\ ;\ x \in [A,B]\}$$

for instance. A more naive way to solve this difficulty in practice is to take

$$m = m_N = \mathfrak{c} N^{-\frac{\varepsilon}{2}\cdot\frac{2(\beta\wedge\gamma)}{2(\beta\wedge\gamma)+1}} \xrightarrow[N\to\infty]{} 0,$$

where $\mathfrak{c} > 0$ is a fixed constant and $\varepsilon \in (0,1)$ is chosen as close as possible to 0. Under Assumption 3.1.2, by Corollary 3.1.5,

$$\exists N_0 \in \mathbb{N} : \forall N > N_0,\ \forall x \in [A,B],\ f(x) > m_N.$$

So, by Proposition 3.2.5, when h (resp. h') is of order $N^{-1/(2\gamma+1)}$ (resp. $N^{-1/(2\beta+1)}$), $\widehat{b}_{N,h,h'}$ converges with the slightly degraded rate

$$N^{-(1-\varepsilon)\frac{2(\beta\wedge\gamma)}{2(\beta\wedge\gamma)+1}}.$$

This last comment remains true for Proposition 3.3.5 and Corollary 3.4.3.

3.3 Risk bound on the discrete-time approximate Nadaraya-Watson estimator

This section deals with risk bounds on $\widehat{f}_{n,N,h}$, $\widehat{bf}_{n,N,h}$, and then on the approximate Nadaraya-Watson estimator $\widehat{b}_{n,N,h}$.

In the sequel, in addition to Assumptions 3.2.1 and 3.2.2, K fulfills the following one.

Assumption 3.3.1. — *The kernel K is two times continuously derivable on $\mathbb{R}$ and $K', K'' \in \mathbb{L}^2(\mathbb{R}, dx)$.*

Compactly supported kernels belonging to $C^2(\mathbb{R})$ or Gaussian kernels fulfill Assumption 3.3.1. The following proposition provides a risk bound on $\widehat{f}_{n,N,h}$ (see (3.6)).

Proposition 3.3.2. — *Under Assumptions 3.1.1, 3.1.2, 3.1.3, 3.2.1, 3.2.2 with $\upsilon = \beta$, and 3.3.1, if*

$$\frac{1}{nh^2} \leqslant 1,$$

then there exists a constant $\mathfrak{c}_{3.3.2} > 0$, not depending on h, N, n and t_0, such that

$$\mathbb{E}(\|\widehat{f}_{n,N,h} - f\|_2^2) \leqslant \frac{\mathfrak{c}_{3.3.2}}{\min\{t_0^{2q_2(\beta)}, t_0^{2q_3}\}} \left(h^{2\beta} + \frac{1}{Nh} + \frac{1}{n^2}\right) + \frac{1}{Nnh^3}.$$

Assume that $\beta = 1$ (extreme case) and h is of order $N^{-1/3}$. As mentioned at Section 3.2, under this condition, the bias-variance tradeoff is reached by the continuous-time estimator of f. Then, the approximation error of $\widehat{f}_{n,N,h}$ is of order $1/n$, which is the order of the variance of the Brownian motion increments along the dissection $(t_0, t_1, \ldots, t_n)$ of $[t_0, T]$. For this reason, the risk bound established in Proposition 3.3.2 is satisfactory. Moreover, by Remark 3.1.6, to take $t_0 \geqslant 1$ when $T > 1$ gives

$$\mathbb{E}(\|\widehat{f}_{n,N,h} - f\|_2^2) \leqslant \mathfrak{c}_{3.3.2} \left(h^{2\beta} + \frac{1}{Nh} + \frac{1}{n^2}\right) + \frac{1}{Nnh^3}.$$

The following proposition provides a risk bound on $\widehat{bf}_{n,N,h}$ (see (3.7)).

Proposition 3.3.3. — *Consider $\varepsilon \in (0, 1)$. Under Assumptions 3.1.1, 3.1.2, 3.1.3, 3.2.1 and 3.3.1, if*

$$\frac{1}{nh^{2-\varepsilon}} \leqslant 1,$$

the kernel K belongs to $\mathbb{L}^4(\mathbb{R}, dx)$ and $z \mapsto zK'(z)$ belongs to $\mathbb{L}^2(\mathbb{R}, dx)$, then there exist a constant $\mathfrak{c}_{3.3.3} > 0$, not depending on ε, h, N, n and t_0, and a constant $\mathfrak{c}_{3.3.3}(\varepsilon) > 0$, depending on ε but not on h, N, n and t_0, such that

$$\mathbb{E}(\|\widehat{bf}_{n,N,h} - bf\|_2^2) \leqslant \frac{\mathfrak{c}_{3.3.3}}{\min\{t_0^{1/2}, t_0^{2q_3}, T - t_0\}} \left(\|(bf)_h - bf\|_2^2 + \frac{1}{Nh} + \frac{1}{n}\right) + \frac{\mathfrak{c}_{3.3.3}(\varepsilon)}{\min\{1, t_0^{(1-\varepsilon)/2}\}} \cdot \frac{1}{Nnh^{3+\varepsilon}}.$$

Remark 3.3.4. — *Note that if b and σ are bounded, Proposition 3.3.3 can be improved. Precisely, with $\varepsilon = 0$ and without the additional conditions $K \in \mathbb{L}^4(\mathbb{R}, dx)$ and $z \mapsto zK'(z)$ belongs to $\mathbb{L}^2(\mathbb{R}, dx)$, the risk bound on $\widehat{bf}_{n,N,h}$ is of same order than in Proposition 3.3.2 (see Remark 3.6.3 for details).*

Assume that bf fulfills Condition (3.10) with $\gamma = 1$ (extreme case), and that h is of order $N^{-1/3}$. Then, for $\varepsilon > 0$ as close as possible to 0, the approximation error of $\widehat{bf}_{n,N,h}$ is of order $N^{\varepsilon/3}/n$. If in addition b and σ are bounded, thanks to Remark 3.3.4, with $\varepsilon = 0$ and without the additional conditions $K \in \mathbb{L}^4(\mathbb{R}, dx)$ and $z \mapsto zK'(z)$ belongs to $\mathbb{L}^2(\mathbb{R}, dx)$, then the approximation error of $\widehat{bf}_{n,N,h}$ is of order $1/n$ as the error of $\widehat{f}_{n,N,h}$. Moreover, by Remark 3.1.6, to take $t_0 \in [1, T-1]$ when $T > 1$ gives

$$\mathbb{E}(\|\widehat{bf}_{n,N,h} - bf\|_2^2) \leqslant \mathfrak{c}_{3.3.3} \left(\|(bf)_h - bf\|_2^2 + \frac{1}{Nh} + \frac{1}{n}\right) + \frac{\mathfrak{c}_{3.3.3}(\varepsilon)}{Nnh^{3+\varepsilon}}.$$

Finally, Propositions 3.3.2 and 3.3.3 allow to provide a risk bound on a truncated version the approximate Nadaraya-Watson estimator $\widehat{b}_{n,N,h}$ (see (3.5)).

Proposition 3.3.5. — *Consider $\varepsilon > 0$, $m \in (0,1]$, and assume that $f(x) > m > 0$ for every $x \in [A,B]$ ($A, B \in \mathbb{R}$ such that $A < B$). Under the assumptions of Proposition 3.3.3 and, in addition, Assumptions 3.1.3 and 3.3.1, there exist a constant $\mathfrak{c}_{3.3.5} > 0$, not depending on ε, A, B, h, N, n and t_0, and a constant $\mathfrak{c}_{3.3.5}(\varepsilon) > 0$, depending on ε but not on A, B, h, N, n and t_0, such that*

$$\mathbb{E}(\|\widetilde{b}_{n,N,h} - b\|^2_{f,A,B}) \leqslant \frac{\mathfrak{c}_{3.2.5}}{m^2 \min\{1, t_0^{(1-\varepsilon)/2}, t_0^{1/2}, t_0^{2q_2(\beta)}, t_0^{2q_3}, T-t_0\}} \times \left[\mathfrak{c}_{3.3.5}\left(\|(bf)_h - bf\|_2^2 + h^{2\beta} + \frac{1}{Nh} + \frac{1}{n}\right) + \frac{\mathfrak{c}_{3.3.5}(\varepsilon)}{Nnh^{3+\varepsilon}}\right]$$

with $\widetilde{b}_{n,N,h}(.) := \widehat{b}_{n,N,h}(.)\mathbf{1}_{\widehat{f}_{n,N,h}(.) > m/2}$.

The proof of Proposition 3.3.5 given Propositions 3.3.2 and 3.3.3 is almost the same than the proof of Proposition 3.2.5 given Propositions 3.2.3 and 3.2.4. Of course one can establish a risk bound on the discrete-time approximate 2bNW estimator, but to focus on the 1 bandwidth estimator is clearer and sufficient to introduce the looCV selection method based on discrete-time observations of $X^1, \ldots, X^N$ at Subsection 3.4.2. Now, assume that bf satisfies Condition (3.10) with $\gamma = \beta$, and that K fulfills Assumption 3.2.2 with $\upsilon = \beta$. Then, $\|(bf)_h - bf\|_2^2$ is of order $h^{2\beta}$. For the sake of simplicity, assume also that b is bounded, and then let's take $\varepsilon = 0$ in Proposition 3.3.5. First, note that the minimization problem

$$\min_{h \in (0,\infty)} \left\{ h^{2\beta} + \frac{1}{Nh} + \frac{1}{n} + \frac{1}{Nnh^3} \right\}$$

has unfortunately no explicit solutions. However, let us provide an upper-bound on the rate of our discrete-time estimator. Since $(nh^2)^{-1} \leqslant 1$, Proposition 3.3.5 says that the risk of $\widetilde{b}_{n,N,h}$ is at most of order $h^{2\beta} + 1/(Nh) + 1/n$. So, the optimal bandwidth for this bound is of order $N^{-1/(2\beta+1)}$, leading to the rate

$$N^{-\frac{2\beta}{2\beta+1}} + \frac{1}{n}.$$

Moreover, by taking a bandwidth of order $N^{-1/(2\beta+1)}$ such that $(nh^2)^{-1} \leqslant 1$, N is at most of order $n^{(2\beta+1)/2}$. So, clearly, the more f and bf are regular, the more N can be chosen freely with respect to n, and if N is of order $n^{(2\beta+1)/2}$, then the risk of $\widetilde{b}_{n,N,h}$ is at most of order $1/n$. Finally, note that if $\beta = 1$, for a bandwidth of order $N^{-1/3}$ such that $(nh^2)^{-1} \leqslant 1$, then

$$1/n \leqslant h^2 \propto N^{-2/3},$$

and the rate of $\widetilde{b}_{n,N,h}$ is of order $N^{-2/3}$ (the optimal rate).

3.4 Bandwidth selection and numerical experiments

This section deals with extensions of the PCO (see Lacour et al. [123]) and looCV methods to the Nadaraya-Watson estimator studied in this paper (see Subsections 3.4.2 and 3.4.1). Subsection 3.4.3 deals with some numerical experiments on the looCV based adaptive Nadaraya-Watson estimator which is, as explained in Comte and Marie [106] in the nonparametric regression framework, numerically more satisfactory than the PCO based one. However, and this is its main advantage, the PCO based adaptive Nadaraya-Watson estimator offers theoretical guarantees : an oracle inequality is established in Subsection 3.4.1. Note also that the PCO method is easier to implement and numerically faster than the Goldenshluger-Lepski method which has been extended by Della Maestra and Hoffmann in [111] for their estimator of the drift function in McKean-Vlasov models.

3.4.1 An extension of the Penalized Comparison to Overfitting method

Let $\mathcal{H}_N$ (resp. $\mathcal{H}'_N$) be a finite subset of $[h_0, 1]$ (resp. $[h'_0, 1]$), where $h_0 > 0$ and $(Nh_0^3)^{-1} \leqslant 1$ (resp. $h'_0 > 0$ and $(Nh'_0)^{-1} \leqslant 1$). Consider an additional kernel δ,

$$\widehat{h} \in \arg\min_{h \in \mathcal{H}_N} \{\|\widehat{bf}_{N,h} - \widehat{bf}_{N,h_0}\|^2_{2,\delta} + \mathrm{pen}(h)\} \tag{3.11}$$

with

$$\mathrm{pen}(h) := \frac{2}{(T-t_0)^2 N^2} \sum_{i=1}^{N} \left\langle \int_{t_0}^{T} K_h(X_s^i - \cdot) dX_s^i, \int_{t_0}^{T} K_{h_0}(X_s^i - \cdot) dX_s^i \right\rangle_{2,\delta} ; \forall h \in \mathcal{H}_N, \tag{3.12}$$

and

$$\widehat{h}' \in \arg\min_{h \in \mathcal{H}'_N} \{\|\widehat{f}_{N,h} - \widehat{f}_{N,h'_0}\|_2^2 + \mathrm{pen}'(h)\} \tag{3.13}$$

with

$$\mathrm{pen}'(h) := \frac{2}{(T-t_0)^2 N^2} \sum_{i=1}^{N} \left\langle \int_{t_0}^{T} K_h(X_s^i - \cdot) ds, \int_{t_0}^{T} K_{h'_0}(X_s^i - \cdot) ds \right\rangle_{2} ; \forall h \in \mathcal{H}'_N.$$

This subsection deals with risk bounds on the adaptive estimators $\widehat{bf}_{N,\widehat{h}}(.)$ (see (3.3)), $\widehat{f}_{N,\widehat{h}'}(.)$ (see (3.4)) and

$$\widehat{b}_{N,\widehat{h},\widehat{h}'}(x) = \frac{\widehat{bf}_{N,\widehat{h}}(x)}{\widehat{f}_{N,\widehat{h}'}(x)} \mathbf{1}_{\widehat{f}_{N,\widehat{h}'}(x) > m/2} ; x \in [A, B]$$

with the notations of Proposition 3.2.5. In the sequel, K, δ and σ fulfill the following technical assumption.

Assumption 3.4.1. — *The kernels K and δ are continuously derivable on $\mathbb{R}$, the derivative of K belongs to $\mathbb{L}^2(\mathbb{R}, dx)$, δ is positive and its derivative is bounded, and σ is bounded.*

Moreover, recall that under Assumptions 3.1.1 and 3.1.2, b^2 and σ^2 belong to $\mathbb{L}^1(\mathbb{R}, f(x)dx)$ (see Remark 3.1.4).

Theorem 3.4.2. — *Under Assumptions 3.1.1, 3.1.2, 3.2.1 and 3.4.1,*

(1) There exist two deterministic constants $\mathfrak{c}_{3.4.2,1}, \mathfrak{c}_{3.4.2,2} > 0$, not depending on N, such that for every $\vartheta \in (0,1)$ and $\lambda > 0$, with probability larger than $1 - \mathfrak{c}_{3.4.2,1}|\mathcal{H}_N|e^{-\lambda}$,

$$\|\widehat{bf}_{N,\widehat{h}} - bf\|^2_{2,\delta} \leqslant (1+\vartheta) \min_{h \in \mathcal{H}_N} \|\widehat{bf}_{N,h} - bf\|^2_{2,\delta} + \frac{\mathfrak{c}_{3.4.2,2}}{\vartheta} \left[\|(bf)_{h_0} - bf\|^2_{2,\delta} + \frac{(1+\lambda)^3}{N} \right].$$

(2) There exist two deterministic constants $\bar{\mathfrak{c}}_{3.4.2,1}, \bar{\mathfrak{c}}_{3.4.2,2} > 0$, not depending on N, such that for every $\vartheta \in (0,1)$ and $\lambda > 0$, with probability larger than $1 - \bar{\mathfrak{c}}_{3.4.2,1}|\mathcal{H}_N|e^{-\lambda}$,

$$\|\widehat{f}_{N,\widehat{h}'} - f\|_2^2 \leqslant (1+\vartheta) \min_{h' \in \mathcal{H}_N} \|\widehat{f}_{N,h'} - f\|_2^2 + \frac{\bar{\mathfrak{c}}_{3.4.2,2}}{\vartheta} \left[\|f_{h'_0} - f\|_2^2 + \frac{(1+\lambda)^3}{N} \right].$$

Corollary 3.4.3. — *Under Assumptions 3.1.1, 3.1.2, 3.2.1 and 3.4.1, if $f(x), \delta(x) > m > 0$ for every $x \in [A,B]$ ($m \in (0,1]$ and $A, B \in \mathbb{R}$ such that $A < B$), then there exists a deterministic constant $\mathfrak{c}_{3.4.3} > 0$, not depending on N, A and B, such that for every $\vartheta \in (0,1)$,*

$$\mathbb{E}(\|\widehat{b}_{N,\widehat{h},\widehat{h}'} - b\|^2_{f,A,B}) \leqslant \frac{2\mathfrak{c}_{3.2.5}(1 \vee \|\delta\|_\infty)}{m^3}\left[(1+\vartheta)\min_{(h,h')\in\mathcal{H}_N\times\mathcal{H}'_N}\{\mathbb{E}(\|\widehat{bf}_{N,h} - bf\|_2^2) + \mathbb{E}(\|\widehat{f}_{N,h'} - f\|_2^2)\} + \frac{\mathfrak{c}_{3.4.3}}{\vartheta}\left(\|(bf)_{h_0} - bf\|_2^2 + \|f_{h'_0} - f\|_2^2 + \frac{1}{N}\right)\right].$$

Corollary 3.4.3 says that the risk of the adaptive estimator $\widehat{b}_{N,\widehat{h},\widehat{h}'}$ is controlled by the sum of the minimal risks of

$$\widehat{bf}_{N,h} \quad \text{and} \quad \widehat{f}_{N,h'} \; ; (h,h') \in \mathcal{H}_N,$$

up to a multiplicative constant and a negligible additive term.

Remark 3.4.4. — *The condition $(Nh_0^3)^{-1} \leqslant 1$ on the bandwidths collection $\mathcal{H}_N$ is quite uncomfortable but not that much because if bf satisfies Condition (3.10) with $\gamma = \beta \geqslant 2$, then the (unknown) bandwidth h^* of order $N^{-1/(2\beta+1)}$ such that our estimator of bf reaches the bias-variance tradeoff (see Section 3.2) possibly belongs to $\mathcal{H}_N$. Indeed, there exists an unknown constant $\mathfrak{c}^* > 0$ such that $h^* = \mathfrak{c}^* N^{-1/(2\beta+1)}$, and then*

$$\frac{1}{N(h^*)^3} = (\mathfrak{c}^*)^{-3} N^{\frac{2}{2\beta+1}(1-\beta)} \leqslant 1$$

for N large enough. Moreover, the proof of Proposition 3.4.2 remains true by replacing the condition $(Nh_0^3)^{-1} \leqslant 1$ by $(Nh_0^3)^{-1} \leqslant \mathfrak{m}$ with $\mathfrak{m} > 0$. So, even for $\beta = 1$, $(N(h^*)^3)^{-1} \leqslant (\mathfrak{c}^*)^{-3}$ and then h^* possibly belongs to $\mathcal{H}_N$ when $\mathfrak{m}$ is large enough.

Remark 3.4.5. — *A nice choice for δ is the standard normal density :*

$$\delta(x) := \frac{e^{-x^2/2}}{\sqrt{2\pi}} \; ; \forall x \in \mathbb{R}.$$

First, δ obviously fulfills Assumption 3.4.1. Moreover, $\|\delta\|_\infty \leqslant 1$. Finally, by assuming that $f(x) > m_1$ for every $x \in [A,B]$ ($m_1 \in (0,1]$ and $A, B \in \mathbb{R}$ such that $A < B$), since δ is continuous and positive on $\mathbb{R}$ ($\text{supp}(\delta) = \mathbb{R}$), necessarily there exists $m_2 > 0$ such that $\delta(x) > m_2$ for every $x \in [A,B]$. So, $f(x), \delta(x) > m = m_1 \wedge m_2 > 0$ for every $x \in [A,B]$. Therefore, under Assumptions 3.1.1, 3.1.2, 3.2.1 and 3.4.1, by Corollary 3.4.3,

$$\mathbb{E}(\|\widehat{b}_{N,\widehat{h},\widehat{h}'} - b\|^2_{f,A,B}) \leqslant \frac{2\mathfrak{c}_{3.2.5}}{m^3}\left[(1+\vartheta)\min_{(h,h')\in\mathcal{H}_N\times\mathcal{H}'_N}\{\mathbb{E}(\|\widehat{bf}_{N,h} - bf\|_2^2) + \mathbb{E}(\|\widehat{f}_{N,h'} - f\|_2^2)\} + \frac{\mathfrak{c}_{3.4.3}}{\vartheta}\left(\|(bf)_{h_0} - bf\|_2^2 + \|f_{h'_0} - f\|_2^2 + \frac{1}{N}\right)\right]$$

for every $\vartheta \in (0,1)$.

3.4.2 An extension of the leave-one-out cross-validation method

First of all, note that the estimator $\widehat{b}_{n,N,h}$ (see (3.5)) can be written the following way :

$$\widehat{b}_{n,N,h}(x) = \sum_{i=1}^{N}\sum_{j=0}^{n-1} \omega_j^i(x)(X^i_{t^i_{j+1}} - X_{t^i_j})$$

with

$$\omega_j^i(x) := \frac{K_h(X_{t_j}^i - x)}{\sum_{k=1}^{N}\sum_{\ell=0}^{n-1} K_h(X_{t_\ell}^k - x)(t_{\ell+1} - t_\ell)} \; ; \forall (j,i) \in \{0,\dots,n-1\} \times \{1,\dots,N\},$$

satisfying

$$\sum_{i=1}^{N}\sum_{j=0}^{n-1} \omega_j^i(x)(t_{j+1} - t_j) = 1.$$

This nice (weighted) representation of $\widehat{b}_{n,N,h}(x)$ allows us to consider the following extension of the well-known looCV criterion in our framework :

$$\mathrm{CV}(h) := \sum_{i=1}^{N}\left[\sum_{j=0}^{n-1} \widehat{b}_{n,N,h}^{-i}(X_{t_j}^i)^2(t_{j+1} - t_j) - 2\sum_{j=0}^{n-1} \widehat{b}_{n,N,h}^{-i}(X_{t_j}^i)(X_{t_{j+1}}^i - X_{t_j}^i)\right]$$

with

$$\widehat{b}_{n,N,h}^{-i}(x) := \sum_{k\in\{1,\dots,N\}\setminus\{i\}} \sum_{j=0}^{n-1} \omega_j^k(x)(X_{t_{j+1}}^k - X_{t_j}^k) \; ; \forall i \in \{1,\dots,N\}.$$

Let us explain heuristically this extension of the looCV criterion. By assuming that $dX_t = Y_t dt$, Equation (3.1) leads to the regression model

$$Y_{t_j} = b(X_{t_j}) + \varepsilon_{t_j} \quad \text{with} \quad \int_0^{t_j} \varepsilon_s ds = \int_0^{t_j} \sigma(X_s) dW_s.$$

Then, a natural extension of the looCV criterion is

$$\begin{aligned} \mathrm{CV}^*(h) &:= \sum_{i=1}^{N}\sum_{j=0}^{n-1} (Y_{t_j}^i - \widehat{b}_{n,N,h}^{-i}(X_{t_j}^i))^2(t_{j+1} - t_j) \\ &\approx \mathrm{CV}(h) + \sum_{i=1}^{N}\sum_{j=0}^{n-1} (Y_{t_j}^i)^2(t_{j+1} - t_j) \end{aligned}$$

because $Y_{t_j}(t_{j+1} - t_j) \approx X_{t_{j+1}} - X_{t_j}$ thanks to the assumption $dX_t = Y_t dt$. Of course $\mathrm{CV}^*(h)$ is not satisfactory because the last term of its previous decomposition doesn't exist, but since this term doesn't depend on h, to minimize $\mathrm{CV}^*(.)$ is almost equivalent to minimize $\mathrm{CV}(.)$ which only involves quantities existing without the condition $dX_t = Y_t dt$.

3.4.3 Numerical experiments

Some numerical experiments on our estimation method are presented in this subsection. The discrete-time approximate Nadaraya-Watson (NW) (see (3.5)) estimator is computed on 4 datasets generated by SDEs with various types of vector fields. In each case, the bandwidth of the NW estimator is selected via the looCV method introduced at Subsection 3.4.2. On the one hand, two models with the same linear drift function are considered, but with an additive noise for the first one and a multiplicative noise for the second one :

1. The so-called Langevin equation, that is

$$X_t = x_0 - \int_0^t X_s ds + 0.1 \cdot W_t.$$

2. The hyperbolic diffusion process, that is

$$X_t = x_0 - \int_0^t X_s ds + 0.1 \int_0^t \sqrt{1 + X_s^2} dW_s.$$

On the other hand, two models having the same non-linear drift function involving sin(.) are considered, but here again with an additive noise for the first one and a multiplicative noise for the second one :

3. The third model is defined by

$$X_t = x_0 - \int_0^t (X_s + \sin(4X_s)) ds + 0.1 \cdot W_t.$$

4. The fourth model is defined by

$$X_t = x_0 - \int_0^t (X_s + \sin(4X_s)) ds + 0.1 \int_0^t (2 + \cos(X_s)) dW_s.$$

The models and the estimator are implemented by taking $N = 200$, $n = 50$, $T = 5$, $x_0 = 2$, $t_0 = 1$ and K the Gaussian kernel $z \mapsto (2\pi)^{-1/2} e^{-z^2/2}$. For Models 1 and 2, the estimator of the drift function is computed for the bandwidths set

$$\mathcal{H}_1 := \{0.02k \ ; \ k = 1, \dots, 10\},$$

and for Models 3 and 4, it is computed for the bandwidths set

$$\mathcal{H}_2 := \{0.01k \ ; \ k = 1, \dots, 10\}.$$

Each set of bandwidths has been chosen after testing different values of h, to see with which ones the estimation performs better. To choose smaller values in the second set of bandwidths allows to check that the looCV method does not systematically select the smallest bandwidth for Models 3 and 4.

For each of the previous models, on Figures 3.1, 3.2, 3.3 and 3.4 respectively, the true drift function (in red) and the looCV adaptive NW estimator (in blue) are plotted on the left-hand side, and the beam of proposals is plotted in green on the right-hand side. On Figures 3.1 and 3.2, one can see that the drift function is well estimated by the looCV adaptive NW estimator, with a MSE equal to $2.95 \cdot 10^{-4}$ for the Langevin equation and to $8.31 \cdot 10^{-4}$ for the hyperbolic diffusion process. As presumed, the multiplicative noise in Model 2 slightly degrades the MSE. Note that when the bandwidth is too small, the estimation degrades, but the looCV method selects a higher value of h which performs better on the estimation. This means that, as expected, the looCV method selects a reasonable approximation of the bandwidth for which the NW estimator reaches the bias-variance tradeoff. On Figures 3.3 and 3.4, one can see that the drift function of Models 3 and 4 is still well estimated by our looCV adaptive NW estimator. However, note that there is a significant degradation of the MSE, which is equal to $2.89 \cdot 10^{-3}$ for Model 3 and to $9.26 \cdot 10^{-3}$ for Model 4. This is probably related to the *nonlinearity* of the drift function to estimate. Once again, to consider a multiplicative noise in Model 4 degrades the estimation quality with respect to Model 3. As for Models 1 and 2, note that the looCV does not systematically select the smallest bandwidth.

For Model 1, at levels $n = 10, 20, \dots, 100$, Figure 3.5 shows the evolution of the MSE of the looCV adaptive NW estimator as a function of N. For this study, the value of N

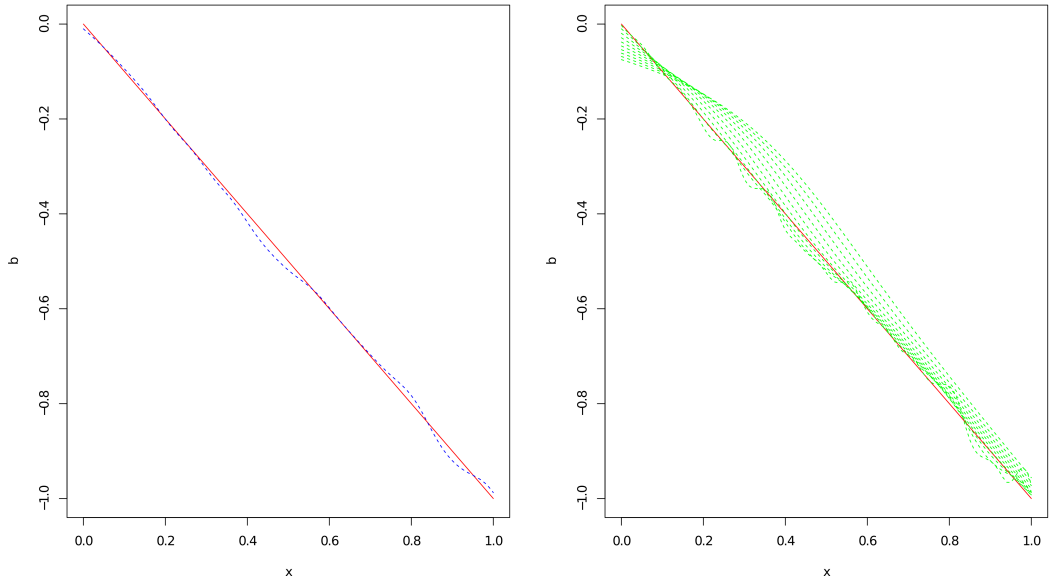

FIGURE 3.1 – LooCV NW estimation for Model 1 (Langevin equation), $\widehat{h} = 0.04$.

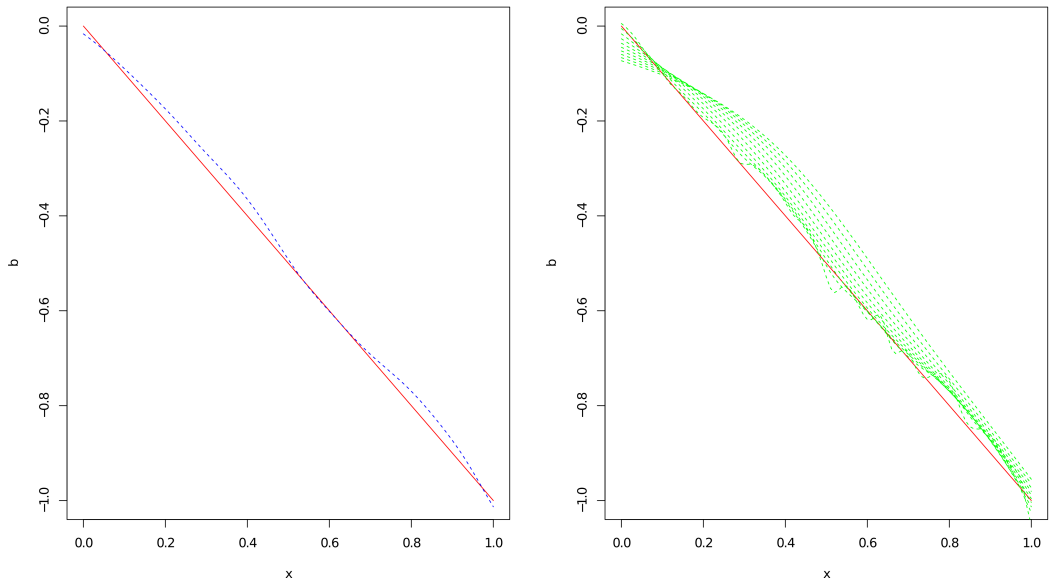

FIGURE 3.2 – LooCV NW estimation for Model 2 (hyperbolic diffusion process), $\widehat{h} = 0.04$.

ranges from 20 to 200. Figure 3.5 shows that the MSE of our adaptive estimator remains low regardless to the value of (n, N) (from $4.50 \cdot 10^{-5}$ to $9.01 \cdot 10^{-3}$), decreases when N increases (for each n), and decreases when n increases (for a fixed N). This is consistent with the risk bounds of Section 4. Note also that for $N \geqslant 70$, there is no significant gain to take n larger than 30. For Model 3, Figure 3.6 shows the evolution of the MSE of the looCV adaptive NW estimator as a function of N and leads to the same conclusions than for Model 1. Note anyway that due to the *nonlinearity* of b, the MSE of our adaptive estimator reaches higher values (from $2.67 \cdot 10^{-4}$ to $4.49 \cdot 10^{-2}$) than for Model 1. Again,

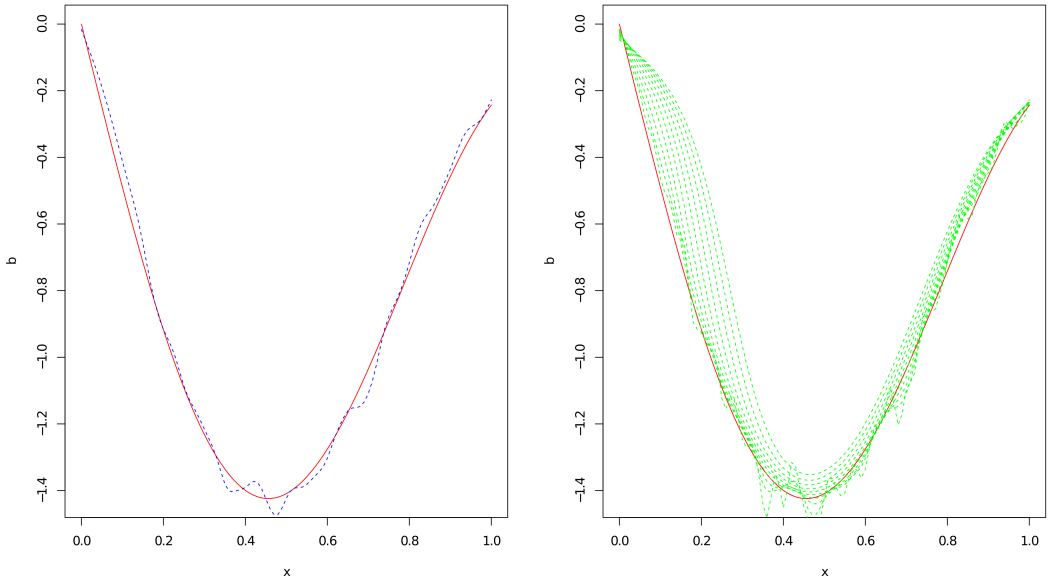

FIGURE 3.3 – LooCV NW estimation for Model 3, $\widehat{h} = 0.02$.

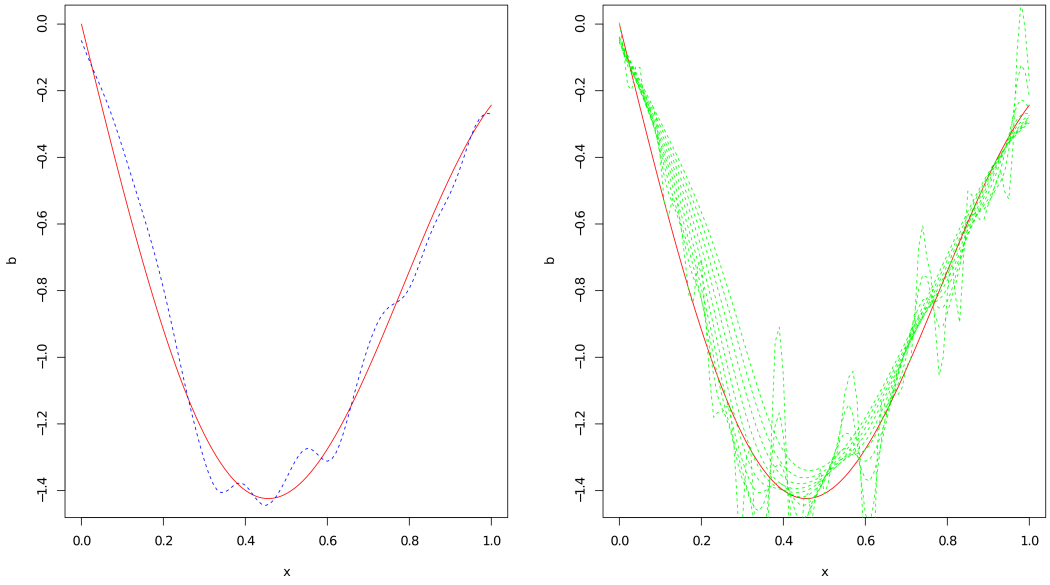

FIGURE 3.4 – LooCV NW estimation for Model 4, $\widehat{h} = 0.06$.

there is no significant gain to take n larger than 30, and above all larger than 70.

Finally, for each model, Table 3.1 gathers the mean MSE of 100 looCV NW estimations of the drift function as well as the mean MSE of the corresponding 100 oracle estimations. The mean MSEs are globally low, but significantly higher for the models with a nonlinear drift function (Models 3 and 4) than for the models with a linear one (Models 1 and 2). Moreover, for each drift function, the mean MSE is slightly degraded for the models with a multiplicative noise (Models 2 and 4) with respect to the models with an additive one (Models 1 and 3). Note also that for each model, the mean MSE of

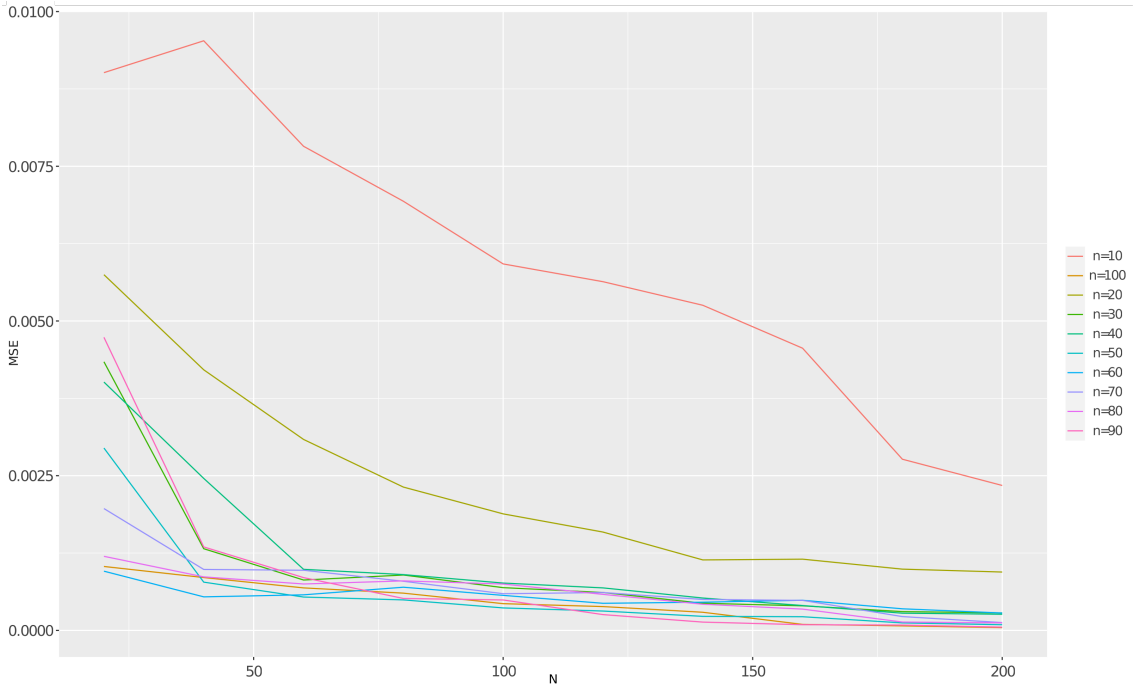

FIGURE 3.5 – MSE of the looCV estimator with respect to N and n for Model 1.

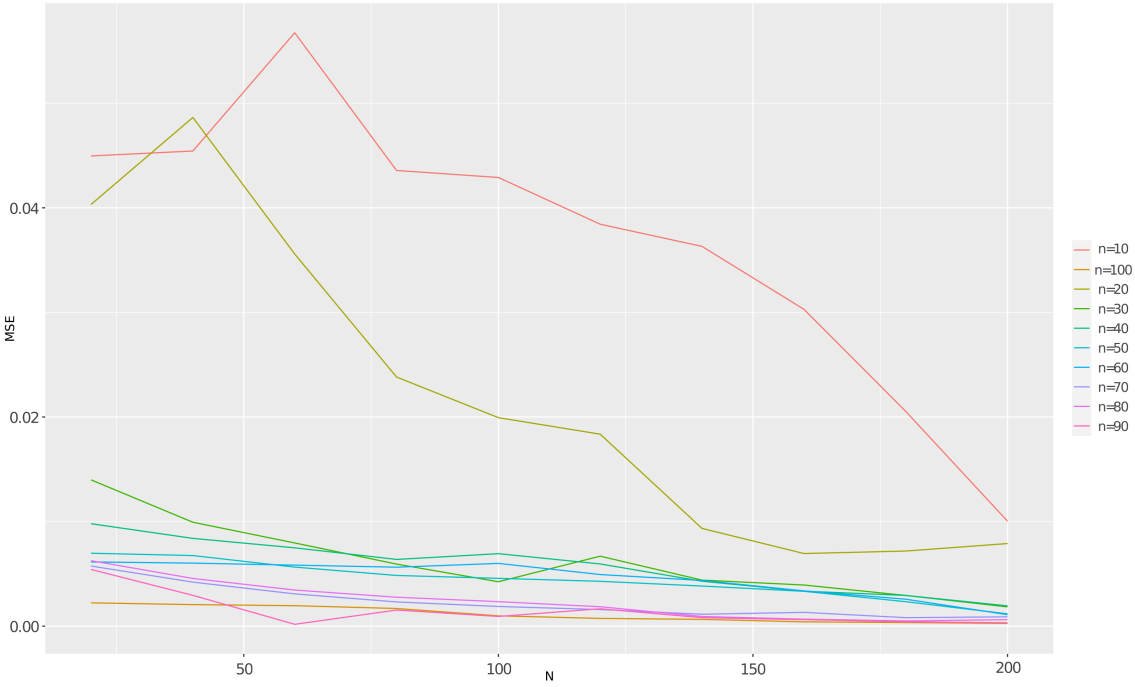

FIGURE 3.6 – MSE of the looCV estimator with respect to N and n for Model 3.

the looCV estimations is close to the mean MSE of the corresponding oracle estimations. This means that our looCV method performs well in practice.

Remark 3.4.6. — *Note that to take $t_0 \geqslant 1$ (here $t_0 = 1$) is recommended even in numerical experiments. Indeed, for instance, the mean MSE of 10 looCV estimations for*

	looCV	Oracle
Model 1	$3.03 \cdot 10^{-4}$	$2.67 \cdot 10^{-4}$
Model 2	$6.52 \cdot 10^{-4}$	$4.96 \cdot 10^{-4}$
Model 3	$2.45 \cdot 10^{-3}$	$1.99 \cdot 10^{-3}$
Model 4	$9.15 \cdot 10^{-3}$	$6.02 \cdot 10^{-3}$

TABLE 3.1 – Mean MSEs of 100 looCV adaptive NW estimations compared to the oracle estimations.

Model 1 is significantly lower with $t_0 = 1$ *(*$2.49 \cdot 10^{-4}$*) than with* $t_0 = 0$ *(*$3.82 \cdot 10^{-3}$*).*

3.5 Concluding remarks

In this paper, first, a risk bound on our continuous-time Nadaraya-Watson estimator of b has been established. This bound is satisfactory because it leads to a rate of same order than in the classic nonparametric regression framework (see Comte [52], Chapter 4), and of same order than in Della Maestra and Hoffmann [111] for their estimator of the drift function in McKean-Vlasov models. Then, a risk bound on a discrete-time approximate estimator of b has been established too. The bound is satisfactory when b and σ are bounded, but a bit degraded when b is unbounded. To improve this bound will be the subject of future investigations.

In a second part, two bandwidth selection methods are provided. The first one is an extension of the PCO method to the 2bNW estimator of b in the spirit of Comte and Marie [106]. An oracle inequality is established but under the condition $(Nh^3)^{-1} \leqslant 1$ (instead of $(Nh)^{-1} \leqslant 1$) on the bandwidths collection. Unfortunately, it seems difficult to bypass this condition because of some constants involved in Bernstein's inequality and in the concentration inequality for U-statistics of Giné and Nickl [115] (see Subsection 3.6.7), but as explained at Remark 3.4.4 this condition is not so bad. The second bandwidth selection method is an extension of the looCV procedure for the discrete-time approximate estimator written has a convex combination. As in the nonparametric regression framework, this method is numerically satisfactory but it seems difficult to establish a theoretical risk bound on the associated adaptive estimator.

Finally, the estimation of b has been only investigated in the case of one-dimensional diffusion processes because of its simplicity, but by following the same ideas than Halconruy and Marie used in the nonparametric regression framework in [117], the major part of the results of the present paper should be extendable to multidimensional diffusion processes.

3.6 Proofs

3.6.1 Proof of Corollary 3.1.5

First of all, since $p_t(x_0, x) > 0$ for every $(t, x) \in (0, T] \times \mathbb{R}$,

$$f(x) = \frac{1}{T - t_0} \int_{t_0}^{T} p_t(x_0, x) dt > 0$$

for every $x \in \mathbb{R}$. Consider $\ell \in \{0, \dots, \beta - 1\}$ and $\theta \in \mathbb{R}_+$. Thanks to the bound on $(t, x) \mapsto \partial_x^{\ell+1} p_t(x_0, x)$ given in Assumption 3.1.2,

$$\begin{aligned}
\|f^{(\ell)}(\cdot + \theta) - f^{(\ell)}\|_2^2 &= \int_{-\infty}^{\infty} [f^{(\ell)}(x + x_0 + \theta) - f^{(\ell)}(x + x_0)]^2 dx \\
&\leqslant \frac{1}{T - t_0} \int_{t_0}^{T} \int_{-\infty}^{\infty} (\partial_2^{\ell} p_t(x_0, x + x_0 + \theta) - \partial_2^{\ell} p_t(x_0, x + x_0))^2 dx dt \\
&\leqslant \frac{\theta^2}{T - t_0} \int_{t_0}^{T} \int_{-\infty}^{\infty} \sup_{z \in [x, x+\theta]} |\partial_2^{\ell+1} p_t(x_0, z + x_0)|^2 dx dt \\
&\leqslant \mathfrak{c}_{3.1.2,2}(\ell + 1)^2 \frac{\theta^2}{T - t_0} \int_{t_0}^{T} \frac{1}{t^{2q_2(\ell+1)}} \int_{-\infty}^{\infty} \sup_{z \in [x, x+\theta]} \exp\left(-2\mathfrak{m}_{3.1.2,2}(\ell + 1)\frac{z^2}{t}\right) dx dt \\
&= \mathfrak{c}_{3.1.2,2}(\ell + 1)^2 \frac{\theta^2}{T - t_0} \int_{t_0}^{T} \frac{1}{t^{2q_2(\ell+1)}} \times \\
&\qquad \left[\int_{-\infty}^{-\theta} \exp\left(-2\mathfrak{m}_{3.1.2,2}(\ell + 1)\frac{(x + \theta)^2}{t}\right) dx + \theta + \int_{0}^{\infty} \exp\left(-2\mathfrak{m}_{3.1.2,2}(\ell + 1)\frac{x^2}{t}\right) dx\right] dt \\
&\leqslant \frac{1}{t_0^{2q_2(\ell+1)}} \left[\mathfrak{c}_1 \theta^2 + \theta^3 \max_{k \in \{0, \dots, \beta-1\}} \mathfrak{c}_{3.1.2,2}(k + 1)^2\right]
\end{aligned}$$

with

$$\mathfrak{c}_1 = 2 \max_{k \in \{0, \dots, \beta-1\}} \left\{\mathfrak{c}_{3.1.2,2}(k + 1)^2 \int_{0}^{\infty} \exp\left(-2\mathfrak{m}_{3.1.2,2}(k + 1)\frac{x^2}{T}\right) dx\right\},$$

and the same way,

$$\begin{aligned}
\|f^{(\ell)}(\cdot - \theta) - f^{(\ell)}\|_2^2 &\leqslant \frac{\theta^2}{T - t_0} \int_{t_0}^{T} \int_{-\infty}^{\infty} \sup_{z \in [x-\theta, x]} |\partial_2^{\ell+1} p_t(x_0, z + x_0)|^2 dx dt \\
&\leqslant \mathfrak{c}_{3.1.2,2}(\ell + 1)^2 \frac{\theta^2}{T - t_0} \int_{t_0}^{T} \frac{1}{t^{2q_2(\ell+1)}} \times \\
&\qquad \left[\int_{-\infty}^{0} \exp\left(-2\mathfrak{m}_{3.1.2,2}(\ell + 1)\frac{x^2}{t}\right) dx + \theta + \int_{\theta}^{\infty} \exp\left(-2\mathfrak{m}_{3.1.2,2}(\ell + 1)\frac{(x - \theta)^2}{t}\right) dx\right] dt \\
&\leqslant \frac{1}{t_0^{2q_2(\ell+1)}} \left[\mathfrak{c}_1 \theta^2 + \theta^3 \max_{k \in \{0, \dots, \beta-1\}} \mathfrak{c}_{3.1.2,2}(k + 1)^2\right].
\end{aligned}$$

This concludes the proof.

3.6.2 Proof of Proposition 3.2.3

First of all, the bias of $\widehat{f}_{N,h}(x)$ is denoted by $\mathfrak{b}(x)$ and its variance by $\mathfrak{v}(x)$. Moreover, let us recall the bias-variance decomposition of the $\mathbb{L}^2$-risk of $\widehat{f}_{N,h}$:

$$\mathbb{E}(\|\widehat{f}_{N,h} - f\|_2^2) = \int_{-\infty}^{\infty} \mathfrak{b}(x)^2 dx + \int_{-\infty}^{\infty} \mathfrak{v}(x) dx.$$

On the one hand, let us find a suitable bound on the integrated variance of $\widehat{f}_{N,h}$. Since $X^1, \dots, X^N$ are i.i.d. copies of X, and thanks to Jensen's inequality,

$$\mathfrak{v}(x) \quad = \quad \mathrm{var}\left(\frac{1}{N(T - t_0)} \sum_{i=1}^{N} \int_{t_0}^{T} K_h(X_t^i - x) dt\right)$$

$$= \frac{1}{N(T-t_0)^2}\mathrm{var}\left(\int_{t_0}^T K_h(X_t-x)dt\right) \leqslant \frac{1}{N}\mathbb{E}\left[\left(\int_{t_0}^T K_h(X_t-x)\frac{dt}{T-t_0}\right)^2\right]$$
$$\leqslant \frac{1}{N(T-t_0)}\int_{t_0}^T \mathbb{E}(K_h(X_t-x)^2)dt = \frac{1}{N}\int_{-\infty}^{\infty} K_h(z-x)^2 f(z)dz.$$

Thus, since K is symmetric,

$$\begin{aligned}\int_{-\infty}^{\infty}\mathfrak{v}(x)dx &\leqslant \frac{1}{N}\int_{-\infty}^{\infty} f(z)\int_{-\infty}^{\infty} K_h(z-x)^2 dxdz \\ &= \frac{1}{Nh}\left(\int_{-\infty}^{\infty} f(z)dz\right)\left(\int_{-\infty}^{\infty} K(x)^2 dx\right) = \frac{\|K\|_2^2}{Nh}.\end{aligned}$$

On the other hand, let us find a suitable bound on the integrated squared-bias of $\widehat{f}_{N,h}(x)$. Since $X^1,\ldots,X^N$ are i.i.d. copies of X,

$$\begin{aligned}\mathfrak{b}(x) &= \frac{1}{T-t_0}\int_{t_0}^T \mathbb{E}(K_h(X_t-x))dt - f(x) \\ &= \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{z-x}{h}\right)f(z)dz - f(x) \\ &= \int_{-\infty}^{\infty} K(z)(f(hz+x)-f(x))dz.\end{aligned}$$

First, assume that $\beta = 1$. By Assumption 3.2.2, the generalized Minkowski inequality and Corollary 3.1.5,

$$\begin{aligned}\int_{-\infty}^{\infty}\mathfrak{b}(x)^2dx &\leqslant \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} K(z)(f(hz+x)-f(x))dz\right)^2 dx \\ &\leqslant \left[\int_{-\infty}^{\infty} K(z)\left(\int_{-\infty}^{\infty}(f(hz+x)-f(x))^2dx\right)^{1/2}dz\right]^2 \leqslant \mathfrak{c}_1(t_0)h^2\end{aligned}$$

with

$$\mathfrak{c}_1(t_0) = \frac{\mathfrak{c}_{3.1.5}}{t_0^{2q_2(1)}}\left(\int_{-\infty}^{\infty}|z|(1+|z|^{1/2})|K(z)|dz\right)^2.$$

Now, assume that $\beta \geqslant 2$. By the Taylor formula with integral remainder, for every $z\in\mathbb{R}$,

$$f(hz+x)-f(x) = \mathbf{1}_{\beta\geqslant 3}\sum_{\ell=1}^{\beta-2}\frac{(hz)^\ell}{\ell!}f^{(\ell)}(x) + \frac{(hz)^{\beta-1}}{(\beta-2)!}\int_0^1(1-\tau)^{\beta-2}f^{(\beta-1)}(\tau hz+x)d\tau.$$

Then, by Assumption 3.2.2, the generalized Minkowski inequality (two times) and Corollary 3.1.5,

$$\begin{aligned}\int_{-\infty}^{\infty}\mathfrak{b}(x)^2dx &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} K(z)(f(hz+x)-f(x))dz\right)^2 dx \\ &= \frac{h^{2(\beta-1)}}{|(\beta-2)!|^2}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} z^{\beta-1}K(z)\int_0^1(1-\tau)^{\beta-2}[f^{(\beta-1)}(\tau hz+x)-f^{(\beta-1)}(x)]d\tau dz\right)^2 dx \\ &\leqslant \frac{h^{2(\beta-1)}}{|(\beta-2)!|^2}\times \\ &\quad\left[\int_{-\infty}^{\infty}|z|^{\beta-1}|K(z)|\int_0^1(1-\tau)^{\beta-2}\left(\int_{-\infty}^{\infty}[f^{(\beta-1)}(\tau hz+x)-f^{(\beta-1)}(x)]^2dx\right)^{1/2}d\tau dz\right]^2\end{aligned}$$

$$\leqslant \frac{\mathfrak{c}_{3.1.5}h^{2\beta}}{|(\beta-2)!|^2 t_0^{2q_2(\beta)}}\left(\int_{-\infty}^{\infty}|z|^{\beta-1}|K(z)|\int_0^1(1-\tau)^{\beta-2}[\tau|z|+(\tau|z|)^{3/2}]d\tau dz\right)^2 \leqslant \mathfrak{c}_2(t_0)h^{2\beta}$$

with

$$\mathfrak{c}_2(t_0)=\frac{\mathfrak{c}_{3.1.5}}{|(\beta-2)!|^2 t_0^{2q_2(\beta)}}\left(\int_{-\infty}^{\infty}|z|^{\beta}(1+|z|^{1/2})|K(z)|dz\right)^2.$$

This concludes the proof.

3.6.3 Proof of Proposition 3.2.4

First of all,

$$\mathbb{E}(\|\widehat{bf}_{N,h}-bf\|_2^2)=\int_{-\infty}^{\infty}\mathfrak{b}(x)^2dx+\int_{-\infty}^{\infty}\mathfrak{v}(x)dx$$

where $\mathfrak{b}(x)$ (resp. $\mathfrak{v}(x)$) is the bias (resp. the variance) term of $\widehat{bf}_{N,h}(x)$ for any $x\in\mathbb{R}$. On the one hand, let us find a suitable bound on the integrated variance of $\widehat{bf}_{N,h}$. Since $X^1,\dots,X^N$ are i.i.d. copies of X,

$$\begin{aligned}\mathfrak{v}(x) &= \mathrm{var}\left(\frac{1}{N(T-t_0)}\sum_{i=1}^N\int_{t_0}^T K_h(X_t^i-x)dX_t^i\right)\leqslant\frac{1}{N(T-t_0)^2}\mathbb{E}\left[\left(\int_{t_0}^T K_h(X_t-x)dX_t\right)^2\right]\\ &\leqslant \frac{2}{N}\mathbb{E}\left[\left(\int_{t_0}^T K_h(X_t-x)b(X_t)\frac{dt}{T-t_0}\right)^2+\frac{1}{(T-t_0)^2}\left(\int_{t_0}^T K_h(X_t-x)\sigma(X_t)dW_t\right)^2\right].\end{aligned}$$

In the right-hand side of the previous inequality, Jensen's inequality on the first term and the isometry property for Itô's integral on the second one give

$$\begin{aligned}\mathfrak{v}(x) &\leqslant \frac{2}{N(T-t_0)}\int_{t_0}^T\mathbb{E}[K_h(X_t-x)^2b(X_t)^2]dt+\frac{2}{N(T-t_0)^2}\int_{t_0}^T\mathbb{E}[K_h(X_t-x)^2\sigma(X_t)^2]dt\\ &= \frac{2}{N}\int_{-\infty}^{\infty}K_h(z-x)^2b(z)^2f(z)dz+\frac{2}{N(T-t_0)}\int_{-\infty}^{\infty}K_h(z-x)^2\sigma(z)^2f(z)dz.\end{aligned}$$

Moreover, K is symmetric and $K\in\mathbb{L}^2(\mathbb{R},dx)$ by Assumption 3.2.1, and $b,\sigma\in\mathbb{L}^2(\mathbb{R},f(x)dx)$ by Remark 3.1.4. Then,

$$\begin{aligned}\int_{-\infty}^{\infty}\mathfrak{v}(x)dx &\leqslant \frac{2}{N}\int_{\mathbb{R}^2}K_h(z-x)^2b(z)^2f(z)dzdx+\frac{2}{N(T-t_0)}\int_{\mathbb{R}^2}K_h(z-x)^2\sigma(z)^2f(z)dzdx\\ &= \frac{2}{Nh}\int_{-\infty}^{\infty}b(z)^2f(z)\int_{-\infty}^{\infty}K(x)^2dxdz+\frac{2}{N(T-t_0)h}\int_{-\infty}^{\infty}\sigma(z)^2f(z)\int_{-\infty}^{\infty}K(x)^2dxdz\\ &\leqslant \frac{2\|K\|_2^2}{Nh}\left(\int_{-\infty}^{\infty}b(z)^2f(z)dz+\frac{1}{T-t_0}\int_{-\infty}^{\infty}\sigma(z)^2f(z)dz\right).\end{aligned}$$

On the other hand, let us find a suitable bound on the integrated squared-bias of $\widehat{bf}_{N,h}(x)$. Again, since $X^1,\dots,X^N$ are i.i.d. copies of X, and since Itô's integral restricted to $\mathbb{H}^2$ is a martingale-valued map,

$$\begin{aligned}\mathfrak{b}(x) &= \mathbb{E}\left[\frac{1}{N(T-t_0)}\sum_{i=1}^N\int_{t_0}^T K_h(X_t^i-x)dX_t^i\right]-b(x)f(x)\\ &= \frac{1}{T-t_0}\mathbb{E}\left(\int_{t_0}^T K_h(X_t-x)dX_t\right)-b(x)f(x)\end{aligned}$$

$$= \frac{1}{T-t_0}\left[\mathbb{E}\left(\int_{t_0}^{T} K_h(X_t-x)b(X_t)dt\right)+\mathbb{E}\left(\int_{t_0}^{T} K_h(X_t-x)\sigma(X_t)dW_t\right)\right]-b(x)f(x)$$
$$= \frac{1}{T-t_0}\int_{t_0}^{T}\mathbb{E}(K_h(X_t-x)b(X_t))dt-b(x)f(x)=\int_{-\infty}^{\infty}K_h(z-x)b(z)f(z)dz-b(x)f(x).$$

Then,

$$\mathfrak{b}(x)^2=((bf)_h-bf)(x)^2 \quad \text{with} \quad (bf)_h=K_h*(bf).$$

Therefore, since f is bounded and b belongs to $\mathbb{L}^2(\mathbb{R},f(x)dx)$ by Remark 3.1.4,

$$\int_{-\infty}^{\infty}\mathfrak{b}(x)^2dx=\|bf-(bf)_h\|_2^2.$$

This concludes the proof.

3.6.4 Proof of Proposition 3.2.5

First of all,

$$\widehat{b}_{N,h,h'}-b=\left[\frac{\widehat{bf}_{N,h}-bf}{\widehat{f}_{N,h'}}+\left(\frac{1}{\widehat{f}_{N,h'}}-\frac{1}{f}\right)bf\right]\mathbf{1}_{\widehat{f}_{N,h'}(.)>m/2}-b\mathbf{1}_{\widehat{f}_{N,h'}(.)\leqslant m/2}.$$

Then,

$$\|\widehat{b}_{N,h,h'}-b\|_{f,A,B}^2 \leqslant 2\left\|\left[\frac{\widehat{bf}_{N,h}-bf}{\widehat{f}_{N,h'}}+\left(\frac{1}{\widehat{f}_{N,h'}}-\frac{1}{f}\right)bf\right]\mathbf{1}_{\widehat{f}_{N,h'}(.)>m/2}\right\|_{f,A,B}^2+2\|b\mathbf{1}_{\widehat{f}_{N,h'}(.)\leqslant m/2}\|_{f,A,B}^2.$$

Moreover, for any $x\in[A,B]$, since $f(x)>m$, for every $\omega\in\{\widehat{f}_{N,h'}(.)\leqslant m/2\}$,

$$|f(x)-\widehat{f}_{N,h'}(x,\omega)|\geqslant f(x)-\widehat{f}_{N,h'}(x,\omega)>m-\frac{m}{2}=\frac{m}{2}.$$

Thus,

$$\begin{aligned}\|\widehat{b}_{N,h,h'}-b\|_{f,A,B}^2 &\leqslant \frac{8}{m^2}\|\widehat{bf}_{N,h}-bf\|_{2,f}^2+\frac{8}{m^2}\|(f-\widehat{f}_{N,h'})b\|_{f,A,B}^2+2\|b\mathbf{1}_{|f(.)-\widehat{f}_{N,h'}(.)|>m/2}\|_{f,A,B}^2\\ &\leqslant \frac{8}{m^2}\int_{-\infty}^{\infty}(\widehat{bf}_{N,h}-bf)(x)^2f(x)dx\\ &\qquad+\frac{8}{m^2}\int_{A}^{B}(f(x)-\widehat{f}_{N,h'}(x))^2b(x)^2f(x)dx\\ &\qquad+2\int_{A}^{B}b(x)^2f(x)\mathbf{1}_{|f(x)-\widehat{f}_{N,h'}(x)|>m/2}dx.\end{aligned}$$

Since f has a sub-Gaussian tail by Assumption 3.1.2, and since b has at most linear growth because it is Lipschitz continuous from $\mathbb{R}$ into itself (see Assumption 3.1.1), b^2f is bounded on $\mathbb{R}$. So,

$$\begin{aligned}\|\widehat{b}_{N,h,h'}-b\|_{f,A,B}^2 &\leqslant \frac{8\|f\|_\infty}{m^2}\|\widehat{bf}_{N,h}-bf\|_2^2\\ &\qquad+\frac{8\|b^2f\|_\infty}{m^2}\|\widehat{f}_{N,h'}-f\|_2^2+2\|b^2f\|_\infty\int_{-\infty}^{\infty}\mathbf{1}_{|f(x)-\widehat{f}_{N,h'}(x)|>m/2}dx.\end{aligned}$$

Therefore, thanks to Markov's inequality,

$$\mathbb{E}(\|\widehat{b}_{N,h,h'}-b\|_{f,A,B}^2) \leqslant \frac{8\|f\|_\infty}{m^2}\mathbb{E}(\|\widehat{bf}_{N,h}-bf\|_2^2)$$

$$+\frac{8\|b^2f\|_\infty}{m^2}\mathbb{E}(\|\widehat{f}_{N,h'}-f\|_2^2)+\frac{8\|b^2f\|_\infty}{m^2}\int_{-\infty}^{\infty}\mathbb{E}(|f(x)-\widehat{f}_{N,h'}(x)|^2)dx$$
$$\leqslant \frac{8(\|f\|_\infty\vee\|b^2f\|_\infty)}{m^2}[\mathbb{E}(\|\widehat{bf}_{N,h}-bf\|_2^2)+2\mathbb{E}(\|\widehat{f}_{N,h'}-f\|_2^2)].$$

Propositions 3.2.4 and 3.2.3 allow to conclude.

3.6.5 Proof of Proposition 3.3.2

First of all, note that

$$\begin{aligned}\mathbb{E}(\|\widehat{f}_{n,N,h}-f\|_2^2) &\leqslant 2\mathbb{E}(\|\widehat{f}_{N,h}-f\|_2^2)+2\mathbb{E}(\|\widehat{f}_{N,h}-\widehat{f}_{n,N,h}\|_2^2)\\ &\leqslant 2\left[\mathfrak{c}_{3.2.3}(t_0)h^{2\beta}+\frac{1}{Nh}+\int_{-\infty}^{\infty}\mathbb{E}(\widehat{f}_{N,h}(x)-\widehat{f}_{n,N,h}(x))^2dx\right.\\ &\qquad\left.+\int_{-\infty}^{\infty}\mathrm{var}(\widehat{f}_{N,h}(x)-\widehat{f}_{n,N,h}(x))dx\right]\end{aligned}$$

by Proposition 3.2.3, and note also that

$$\widehat{f}_{N,h}(x)-\widehat{f}_{n,N,h}(x)=\frac{1}{N(T-t_0)}\sum_{i=1}^{N}\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}}(K_{h,x}(X_t^i)-K_{h,x}(X_{t_j}^i))dt$$

with $K_{h,x}(.):=K_h(\cdot-x)$. On the one hand, for every $s,u\in[t_0,T]$ such that $s\leqslant u$, by Itô's formula, Jensen's inequality, the isometry property for Itô's integral and Remark 3.1.4,

$$\begin{aligned}\int_{-\infty}^{\infty}\mathbb{E}[(K_{h,x}(X_u)-K_{h,x}(X_s))^2]dx &= \int_{-\infty}^{\infty}\mathbb{E}\left[\left(\int_s^u K'_{h,x}(X_t)dX_t+\frac{1}{2}\int_s^u K''_{h,x}(X_t)d\langle X\rangle_t\right)^2\right]dx\\ &\leqslant \mathfrak{c}_1\int_{-\infty}^{\infty}\left[\mathbb{E}\left[\left(\int_s^u K'_{h,x}(X_t)b(X_t)dt\right)^2\right]\right.\\ &\qquad+\mathbb{E}\left[\left(\int_s^u K''_{h,x}(X_t)\sigma(X_t)^2dt\right)^2\right]\\ &\qquad\left.+\mathbb{E}\left[\left(\int_s^u K'_{h,x}(X_t)\sigma(X_t)dW_t\right)^2\right]\right]dx\\ &\leqslant \mathfrak{c}_1\left[(u-s)\int_s^u\mathbb{E}\left(b(X_t)^2\int_{-\infty}^{\infty}K'_{h,x}(X_t)^2dx\right)dt\right.\\ &\qquad+(u-s)\int_s^u\mathbb{E}\left(\sigma(X_t)^4\int_{-\infty}^{\infty}K''_{h,x}(X_t)^2dx\right)dt\\ &\qquad\left.+\int_s^u\mathbb{E}\left(\sigma(X_t)^2\int_{-\infty}^{\infty}K'_{h,x}(X_t)^2dx\right)dt\right]\\ &\leqslant \mathfrak{c}_2\left[\frac{(u-s)^2}{h^3}+\frac{(u-s)^2}{h^5}+\frac{u-s}{h^3}\right]\end{aligned}$$

where $\mathfrak{c}_1$ and $\mathfrak{c}_2$ are two positive constants not depending on s, u, h, N, n and t_0. Then,

$$\begin{aligned}&\int_{-\infty}^{\infty}\mathrm{var}(\widehat{f}_{n,N,h}(x)-\widehat{f}_{N,h}(x))dx\\ &\qquad=\frac{1}{N(T-t_0)^2}\int_{-\infty}^{\infty}\mathrm{var}\left[\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}}(K_{h,x}(X_t)-K_{h,x}(X_{t_j}))dt\right]dx\end{aligned}$$

$$\leqslant \frac{1}{N(T-t_0)}\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}}\int_{-\infty}^{\infty}\mathbb{E}[(K_{h,x}(X_t)-K_{h,x}(X_{t_j}))^2]dxdt \leqslant \frac{\mathfrak{c}_3}{Nnh^3}$$

where the constant $\mathfrak{c}_3 > 0$ is not depending on h, N, n and t_0. On the other hand, by Assumption 3.1.3,

$$\begin{aligned}
|\mathbb{E}(\widehat{f}_{N,h}(x)-\widehat{f}_{n,N,h}(x))| &\leqslant \frac{1}{T-t_0}\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}}|\mathbb{E}(K_{h,x}(X_t))-\mathbb{E}(K_{h,x}(X_{t_j}))|dt \\
&\leqslant \frac{1}{T-t_0}\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}}\int_{-\infty}^{\infty}|K_h(z-x)|\cdot|p_t(x_0,z)-p_{t_j}(x_0,z)|dzdt \\
&\leqslant \frac{1}{T-t_0}\sum_{j=0}^{n-1}\left[\int_{t_j}^{t_{j+1}}(t-t_j)dt\right] \\
&\qquad\qquad \times\left[\int_{-\infty}^{\infty}|K_h(z-x)|\sup_{u\in[t_0,T]}|\partial_u p_u(x_0,z)|dz\right] \\
&\leqslant \mathfrak{c}_{3.1.3,3}\frac{T-t_0}{nt_0^{q_3}}\int_{-\infty}^{\infty}|K(z)|\exp\left[-\mathfrak{m}_{3.1.3,3}\frac{(hz+x-x_0)^2}{T}\right]dz.
\end{aligned}$$

Then, by Jensen's inequality,

$$\begin{aligned}
\int_{-\infty}^{\infty}\mathbb{E}(\widehat{f}_{N,h}(x)-\widehat{f}_{n,N,h}(x))^2dx &\leqslant \frac{\mathfrak{c}_5}{n^2t_0^{2q_3}}\int_{-\infty}^{\infty}|K(z)|\int_{-\infty}^{\infty}\exp\left[-2\mathfrak{m}_{3.1.3,3}\frac{(hz+x-x_0)^2}{T}\right]dxdz \\
&= \frac{\mathfrak{c}_6}{n^2t_0^{2q_3}}
\end{aligned}$$

where

$$\mathfrak{c}_6 = \mathfrak{c}_5\|K\|_1\int_{-\infty}^{\infty}\exp\left[-2\mathfrak{m}_{3.1.3,3}\frac{(x-x_0)^2}{T}\right]dx$$

and the constant $\mathfrak{c}_5 > 0$ is not depending on h, N, n and t_0. This concludes the proof.

3.6.6 Proof of Proposition 3.3.3

The proof of Proposition 3.3.3 relies on the two following technical lemmas.

Lemma 3.6.1. — *Consider a symmetric and continuous function $\varphi_1 : \mathbb{R}\to\mathbb{R}$ such that $\overline{\varphi}_1 : z \mapsto z\varphi_1(z)$ belongs to $\mathbb{L}^2(\mathbb{R},dx)$. Consider also $\varphi_2, \psi \in C^0(\mathbb{R})$ having polynomial growth. Under Assumptions 3.1.1 and 3.1.2, for every $p > 0$, there exists a constant $\mathfrak{c}_{3.6.1}(p) > 0$, not depending on φ_1 and t_0, such that for every $s,t \in [t_0,T]$ satisfying $s < t$,*

$$\begin{aligned}
&\int_{-\infty}^{\infty}\mathbb{E}\left[\left(\int_s^t\varphi_1(x-X_u)\varphi_2(X_u)dW_u\right)^2\psi(X_t)^2\right]dx \\
&\qquad\leqslant \mathfrak{c}_{3.6.1}(p)(t-s)\left[\|\varphi_1\|_2^2+\|\overline{\varphi}_1\|_2^2+\frac{1}{t_0^{1/(2p)}}\left(\int_{-\infty}^{\infty}\varphi_1(z)^{2p}dz\right)^{1/p}\right].
\end{aligned}$$

Lemma 3.6.2. — *Consider $\varphi \in C^0(\mathbb{R})$. Under Assumptions 3.1.1 and 3.1.2, for every $s,t \in [t_0,T]$ such that $s < t$,*

$$\int_{-\infty}^{\infty}\mathbb{E}(K_{h,x}(X_s)\varphi(X_s,X_t))^2dx \leqslant \frac{\mathfrak{c}_{3.1.2,1}\|K\|_1^2}{t_0^{1/2}}\mathbb{E}[\varphi(X_s,X_t)^2].$$

The proof of Lemma 3.6.1 (resp. Lemma 3.6.2) is postponed to Subsubection 3.6.6.1 (resp. Subsubsection 3.6.6.2).

First of all, note that

$$\begin{aligned}\mathbb{E}(\|\widehat{bf}_{n,N,h} - bf\|_2^2) &\leqslant 2\mathbb{E}(\|\widehat{bf}_{N,h} - bf\|_2^2) + 2\mathbb{E}(\|\widehat{bf}_{N,h} - \widehat{bf}_{n,N,h}\|_2^2) \\ &\leqslant 2\left[\|(bf)_h - bf\|_2^2 + \frac{\mathfrak{q}_{3.2.4}(t_0)}{Nh} + \int_{-\infty}^{\infty}\mathbb{E}(\widehat{bf}_{N,h}(x) - \widehat{bf}_{n,N,h}(x))^2 dx \right. \\ &\qquad \left. + \int_{-\infty}^{\infty}\mathrm{var}(\widehat{bf}_{N,h}(x) - \widehat{bf}_{n,N,h}(x))dx\right] \\ &=: 2\left[\|(bf)_h - bf\|_2^2 + \frac{\mathfrak{q}_{3.2.4}(t_0)}{Nh} + B_{n,N,h} + V_{n,N,h}\right]\end{aligned}$$

by Proposition 3.2.4, and note also that

$$\widehat{bf}_{N,h}(x) - \widehat{bf}_{n,N,h}(x) = \frac{1}{N(T-t_0)}\sum_{i=1}^{N}\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}}(K_{h,x}(X_t^i) - K_{h,x}(X_{t_j}^i))dX_t^i.$$

The proof is dissected in two steps. The term $V_{n,N,h}$ is controlled in the first step, and then $B_{n,N,h}$ is controlled in the second one.

Step 1. First of all, by Jensen's inequality,

$$\begin{aligned}V_{n,N,h} &= \frac{1}{N(T-t_0)^2}\int_{-\infty}^{\infty}\mathrm{var}\left[\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}}(K_{h,x}(X_t) - K_{h,x}(X_{t_j}))dX_t\right]dx \\ &\leqslant \frac{2}{N(T-t_0)}\int_{-\infty}^{\infty}\left[\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}}\mathbb{E}[(K_{h,x}(X_t) - K_{h,x}(X_{t_j}))^2 b(X_t)^2]dt\right]dx + V_{n,N,h}^{\sigma}\end{aligned}$$

with

$$V_{n,N,h}^{\sigma} := \frac{2}{N(T-t_0)^2}\int_{-\infty}^{\infty}\mathbb{E}\left[\left(\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}}(K_{h,x}(X_t) - K_{h,x}(X_{t_j}))\sigma(X_t)dW_t\right)^2\right]dx.$$

In order to control $V_{n,N,h}$ as in the proof of Proposition 3.3.2, a preliminary bound on $V_{n,N,h}^{\sigma}$ has to be established via the isometry property of Itô's integral :

$$\begin{aligned}V_{n,N,h}^{\sigma} &= \frac{2}{N(T-t_0)^2}\int_{-\infty}^{\infty}\mathbb{E}\left[\left(\int_{t_0}^{T}\left(\sum_{j=0}^{n-1}(K_{h,x}(X_t) - K_{h,x}(X_{t_j}))\sigma(X_t)\mathbf{1}_{[t_j,t_{j+1}]}(t)\right)dW_t\right)^2\right]dx \\ &= \frac{2}{N(T-t_0)^2}\int_{-\infty}^{\infty}\int_{t_0}^{T}\mathbb{E}\left[\left(\sum_{j=0}^{n-1}(K_{h,x}(X_t) - K_{h,x}(X_{t_j}))\sigma(X_t)\mathbf{1}_{[t_j,t_{j+1}]}(t)\right)^2\right]dtdx \\ &= \frac{2}{N(T-t_0)}\int_{-\infty}^{\infty}\left(\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}}\mathbb{E}[(K_{h,x}(X_t) - K_{h,x}(X_{t_j}))^2\sigma(X_t)^2]dt\right)dx.\end{aligned}$$

Then,

$$V_{n,N,h} \leqslant \frac{2}{N(T-t_0)}\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}}\int_{-\infty}^{\infty}\mathbb{E}[(K_{h,x}(X_t) - K_{h,x}(X_{t_j}))^2 b(X_t)^2]dxdt$$

$$+\frac{2}{N(T-t_0)}\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}}\int_{-\infty}^{\infty}\mathbb{E}[(K_{h,x}(X_t)-K_{h,x}(X_{t_j}))^2\sigma(X_t)^2]dxdt.$$

For $\varphi = b$ or $\varphi = \sigma$, by Itô's formula,

$$\begin{aligned}
&\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}}\int_{-\infty}^{\infty}\mathbb{E}[(K_{h,x}(X_t)-K_{h,x}(X_{t_j}))^2\varphi(X_t)^2]dxdt\\
&\quad\leqslant \mathfrak{c}_1\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}}\int_{-\infty}^{\infty}\left[\mathbb{E}\left[\left(\int_{t_j}^{t}K'_{h,x}(X_u)b(X_u)du\right)^2\varphi(X_t)^2\right]\right.\\
&\qquad+\mathbb{E}\left[\left(\int_{t_j}^{t}K''_{h,x}(X_u)\sigma(X_u)^2du\right)^2\varphi(X_t)^2\right]\\
&\qquad\left.+\mathbb{E}\left[\left(\int_{t_j}^{t}K'_{h,x}(X_u)\sigma(X_u)dW_u\right)^2\varphi(X_t)^2\right]\right]dxdt
\end{aligned}$$

where the constant $\mathfrak{c}_1 > 0$ is not depending on φ, h, N, n and t_0. Moreover, for every $j \in \{0,\dots,n-1\}$ and $t \in [t_j, t_{j+1}]$, by Lemma 3.6.1 with $p = 1/(1-\varepsilon)$,

$$\begin{aligned}
\int_{-\infty}^{\infty}\mathbb{E}\left[\left(\int_{t_j}^{t}K'_{h,x}(X_u)\sigma(X_u)dW_u\right)^2\varphi(X_t)^2\right]dx &\leqslant \mathfrak{c}_{3.6.1}(p)(t-t_j)\int_{-\infty}^{\infty}K'_h(z)^2dz\\
&\quad+\mathfrak{c}_{3.6.1}(p)(t-t_j)\int_{-\infty}^{\infty}z^2K'_h(z)^2dz\\
&\quad+\frac{\mathfrak{c}_{3.6.1}(p)}{t_0^{1/(2p)}}(t-t_j)\left(\int_{-\infty}^{\infty}K'_h(z)^{2p}dz\right)^{1/p}\\
&\leqslant \mathfrak{c}_2(\varepsilon)(t-t_j)\left[1+\frac{1}{h^3}+\frac{t_0^{-(1-\varepsilon)/2}}{h^{3+\varepsilon}}\right]
\end{aligned}$$

where the constant $\mathfrak{c}_2(\varepsilon) > 0$ depends on ε, but not on φ, j, t, h, N, n and t_0. Thus, by Jensen's inequality and Remark 3.1.4,

$$\begin{aligned}
&\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}}\int_{-\infty}^{\infty}\mathbb{E}[(K_{h,x}(X_t)-K_{h,x}(X_{t_j}))^2\varphi(X_t)^2]dxdt\\
&\quad\leqslant \mathfrak{c}_3(\varepsilon)\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}}\left[(t-t_j)\int_{t_j}^{t}\mathbb{E}\left(\varphi(X_t)^2b(X_u)^2\int_{-\infty}^{\infty}K'_{h,x}(X_u)^2dx\right)du\right.\\
&\qquad\left.+(t-t_j)\int_{t_j}^{t}\mathbb{E}\left(\varphi(X_t)^2\sigma(X_u)^4\int_{-\infty}^{\infty}K''_{h,x}(X_u)^2dx\right)du+(t-t_j)\left[1+\frac{1}{h^3}+\frac{t_0^{-(1-\varepsilon)/2}}{h^{3+\varepsilon}}\right]\right]dt\\
&\quad\leqslant \frac{\mathfrak{c}_4(\varepsilon)}{\min\{1,t_0^{(1-\varepsilon)/2}\}}(T-t_0)^3\left(\frac{1}{n^2h^3}+\frac{1}{n^2h^5}+\frac{1}{nh^{3+\varepsilon}}\right)
\end{aligned}$$

where $\mathfrak{c}_3(\varepsilon)$ and $\mathfrak{c}_4(\varepsilon)$ are two positive constants depending on ε, but not on φ, h, N, n and t_0. Therefore,

$$V_{n,N,h} \leqslant \frac{\mathfrak{c}_5(\varepsilon)}{\min\{1,t_0^{(1-\varepsilon)/2}\}}\cdot\frac{1}{Nnh^{3+\varepsilon}}$$

where the constant $\mathfrak{c}_5(\varepsilon) > 0$ depends on ε, but not on h, N, n and t_0.

Step 2. First of all, since Itô's integral restricted to $\mathbb{H}^2$ is a martingale-valued map, since $K_{h,x}$ is a kernel, by Lemma 3.6.2, by Assumptions 3.1.2 and 3.1.3, and since b is Lipschitz continuous,

$$\begin{aligned}
&\int_{-\infty}^{\infty} \mathbb{E}(\widehat{bf}_{N,h}(x) - \widehat{bf}_{n,N,h}(x))^2 dx \\
&\quad = \int_{-\infty}^{\infty} \left(\frac{1}{T-t_0} \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} \mathbb{E}((K_{h,x}(X_t) - K_{h,x}(X_{t_j}))b(X_t))dt \right. \\
&\qquad \left. + \frac{1}{T-t_0} \sum_{j=0}^{n-1} \mathbb{E}\left[\int_{t_j}^{t_{j+1}} (K_{h,x}(X_t) - K_{h,x}(X_{t_j}))\sigma(X_t) dW_t \right] \right)^2 dx \\
&\quad \leqslant \frac{2}{T-t_0} \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} \int_{-\infty}^{\infty} \mathbb{E}(K_{h,x}(X_t)b(X_t) - K_{h,x}(X_{t_j})b(X_{t_j}))^2 dx dt \\
&\qquad + \frac{2}{T-t_0} \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} \int_{-\infty}^{\infty} \mathbb{E}(|K_{h,x}(X_{t_j})| \cdot |b(X_t) - b(X_{t_j})|)^2 dx dt \\
&\quad \leqslant \frac{2\|K\|_1}{T-t_0} \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} |K_{h,x}(z)| dx \right) b(z)^2 (p_t(x_0,z) - p_{t_j}(x_0,z))^2 dz dt \\
&\qquad + \frac{2\mathfrak{c}_{3.1.2,1}\|K\|_1^2}{t_0^{1/2}(T-t_0)} \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} \mathbb{E}[(b(X_t) - b(X_{t_j}))^2] dt \\
&\quad \leqslant \frac{4\|K\|_1^2}{T-t_0} \sum_{j=0}^{n-1} \left[\int_{t_j}^{t_{j+1}} (t-t_j)^2 dt \right] \left[\int_{-\infty}^{\infty} b(z)^2 \sup_{u \in [t_0,T]} |\partial_u p_u(x_0,z)|^2 dz \right] \\
&\qquad + \frac{2\mathfrak{c}_{3.1.2,1}\|K\|_1^2}{t_0^{1/2}(T-t_0)} \|b'\|_\infty^2 \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} \mathbb{E}[(X_t - X_{t_j})^2] dt \\
&\quad \leqslant \mathfrak{c}_6 \left(\frac{1}{t_0^{2q_3} n^2} + \frac{1}{t_0^{1/2}(T-t_0)} \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} \mathbb{E}[(X_t - X_{t_j})^2] dt \right)
\end{aligned}$$

where the constant $\mathfrak{c}_6 > 0$ is not depending on h, N, n and t_0. Moreover, for any $j \in \{0, \ldots, n-1\}$ and $t \in [t_j, t_{j+1}]$,

$$X_t - X_{t_j} = \int_{t_j}^{t} b(X_u) du + \int_{t_j}^{t} \sigma(X_u) dW_u$$

and then, by Jensen's inequality, the isometry property of Itô's integral and Remark 3.1.4,

$$\begin{aligned}
\mathbb{E}[(X_t - X_{t_j})^2] &\leqslant (t-t_j) \int_{t_j}^{t} \mathbb{E}(b(X_u)^2) du + \int_{t_j}^{t} \mathbb{E}(\sigma(X_u)^2) du \\
&\leqslant (t-t_j)^2 \sup_{u \in [t_0,T]} \mathbb{E}(b(X_u)^2) + (t-t_j) \sup_{u \in [t_0,T]} \mathbb{E}(\sigma(X_u)^2) \leqslant \mathfrak{c}_7 (t-t_j)
\end{aligned}$$

where the constant $\mathfrak{c}_7 > 0$ is not depending on j, t, h, N, n and t_0. Therefore,

$$\int_{-\infty}^{\infty} \mathbb{E}(\widehat{bf}_{N,h}(x) - \widehat{bf}_{n,N,h}(x))^2 dx \leqslant \frac{\mathfrak{c}_8}{\min\{t_0^{1/2}, t_0^{2q_3}\}} \left(\frac{1}{n^2} + \frac{1}{n} \right)$$

where the constant $\mathfrak{c}_8 > 0$ is not depending on n, N, h and t_0.

Remark 3.6.3. — *Assume that b and σ are bounded. Then, in Step 1, for $\varphi = b$ or $\varphi = \sigma$,*

$$\begin{aligned}
&\sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} \int_{-\infty}^{\infty} \mathbb{E}[(K_{h,x}(X_t) - K_{h,x}(X_{t_j}))^2 \varphi(X_t)^2] dx dt \\
&\qquad \leqslant \mathfrak{c}_1 \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} \int_{-\infty}^{\infty} \left[\mathbb{E}\left[\left(\int_{t_j}^{t} K'_{h,x}(X_u) b(X_u) du \right)^2 \varphi(X_t)^2 \right] \right. \\
&\qquad\qquad + \mathbb{E}\left[\left(\int_{t_j}^{t} K''_{h,x}(X_u) \sigma(X_u)^2 du \right)^2 \varphi(X_t)^2 \right] \\
&\qquad\qquad \left. + \mathbb{E}\left[\left(\int_{t_j}^{t} K'_{h,x}(X_u) \sigma(X_u) dW_u \right)^2 \varphi(X_t)^2 \right] \right] dx dt \\
&\qquad \leqslant \mathfrak{c}_1 \|\varphi\|_\infty^2 \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} \int_{-\infty}^{\infty} \left[\mathbb{E}\left[\left(\int_{t_j}^{t} K'_{h,x}(X_u) b(X_u) du \right)^2 \right] \right. \\
&\qquad\qquad \left. + \mathbb{E}\left[\left(\int_{t_j}^{t} K''_{h,x}(X_u) \sigma(X_u)^2 du \right)^2 \right] + \mathbb{E}\left[\left(\int_{t_j}^{t} K'_{h,x}(X_u) \sigma(X_u) dW_u \right)^2 \right] \right] dx dt.
\end{aligned}$$

So, in this special case, the bound on $V_{n,N,h}$ is established by using the exact same arguments than in the proof of Proposition 3.3.2. In particular, one can take $\varepsilon = 0$, the additional conditions $K \in \mathbb{L}^4(\mathbb{R}, dx)$ and $z \mapsto zK(z)$ belongs to $\mathbb{L}^2(\mathbb{R}, dx)$ are not required, and the bound on $V_{n,N,h}$ is of order $1/(Nnh^3)$ and doesn't depend on t_0. When $\varphi = b$ or $\varphi = \sigma$ is not bounded, since $\varphi(X_t)$ is not $\sigma(W_u)$-measurable for every $u \in [t_j, t)$ ($j \in \{0, \ldots, n-1\}$), the Hölder inequality has to be used to get a suitable bound on

$$\int_{-\infty}^{\infty} \mathbb{E}\left[\left(\int_{t_j}^{t} K'_{h,x}(X_u) \sigma(X_u) dW_u \right)^2 \varphi(X_t)^2 \right] dx$$

(see the proof of Lemma 3.6.1), and for this reason the variance term in the bound of Proposition 3.3.3 is of order $1/(Nnh^{3+\varepsilon})$ instead of $1/(Nnh^3)$ as when b and σ are bounded.

3.6.6.1 Proof of Lemma 3.6.1

Consider $\varphi(x, z) := \varphi_1(x - z)\varphi_2(z)$ for every $z \in \mathbb{R}$, $q > 0$ such that $1/p + 1/q = 1$, and $s, t \in [0, T]$ such that $s < t$. First of all, by the isometry property of Itô's integral, Burkholder-Davis-Gundy's inequality, Hölder's inequality, Markov's inequality, Remark 3.1.4, and the generalized Minkowski inequality,

$$\begin{aligned}
&\mathbb{E}\left[\left(\int_s^t \varphi(x, X_u) dW_u \right)^2 \psi(X_t)^2 \right] \\
&\qquad \leqslant x^2 \mathbb{E}\left[\left(\int_s^t \varphi(x, X_u) dW_u \right)^2 \mathbf{1}_{\psi(X_t)^2 \leqslant x^2} \right] \\
&\qquad\qquad + \mathbb{E}\left[\left(\int_s^t \varphi(x, X_u) dW_u \right)^{2p} \right]^{1/p} \mathbb{E}(\psi(X_t)^{4q})^{1/(2q)} \mathbb{P}(\psi(X_t)^2 > x^2)^{1/(2q)}
\end{aligned}$$

$$\leqslant x^2\int_s^t \mathbb{E}[\varphi(x,X_u)^2]du + \mathfrak{c}_1(p)\mathbb{E}\left[\left(\int_s^t \varphi(x,X_u)^2du\right)^p\right]^{1/p}\mathbb{E}(\psi(X_t)^{4q})^{1/(2q)}$$
$$\times\left[\frac{\mathbb{E}(\psi(X_t)^2)^{1/(2q)}}{x^{1/q}}\mathbf{1}_{[-1,1]}(x)+\frac{\mathbb{E}(\psi(X_t)^{4q})^{1/(2q)}}{x^2}\mathbf{1}_{\mathbb{R}\setminus[-1,1]}(x)\right]$$
$$\leqslant x^2\int_s^t \mathbb{E}[\varphi(x,X_u)^2]du + \mathfrak{c}_2(p)\left(\int_s^t \mathbb{E}[\varphi(x,X_u)^{2p}]^{1/p}du\right)\left(\frac{1}{x^{1/q}}\mathbf{1}_{[-1,1]}(x)+\frac{1}{x^2}\mathbf{1}_{\mathbb{R}\setminus[-1,1]}(x)\right)$$

where $\mathfrak{c}_1(p)$ and $\mathfrak{c}_2(p)$ are two positive constants depending on p, but not on x, s, t, φ and t_0. On the one hand, since φ_2 has polynomial growth, by Remark 3.1.4, for every $u\in[s,t]$,

$$\begin{aligned}\int_{-\infty}^{\infty} x^2\mathbb{E}[\varphi(x,X_u)^2]dx &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^2\varphi_1(x-z)^2\varphi_2(z)^2p_u(x_0,z)dxdz\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+z)^2\varphi_1(x)^2\varphi_2(z)^2p_u(x_0,z)dxdz\\ &\leqslant 2\left(\int_{-\infty}^{\infty}x^2\varphi_1(x)^2dx\right)\left(\int_{-\infty}^{\infty}\varphi_2(z)^2p_u(x_0,z)dz\right)\\ &\qquad +2\left(\int_{-\infty}^{\infty}\varphi_1(x)^2dx\right)\left(\int_{-\infty}^{\infty}z^2\varphi_2(z)^2p_u(x_0,z)dz\right)\\ &\leqslant \mathfrak{c}_3(\|\overline{\varphi}_1\|_2^2+\|\varphi_1\|_2^2)\end{aligned}$$

where the constant $\mathfrak{c}_3>0$ is not depending on u, φ_1 and t_0. Then,

$$\int_{-\infty}^{\infty}x^2\int_s^t\mathbb{E}[\varphi(x,X_u)^2]dudx\leqslant\mathfrak{c}_3(t-s)(\|\overline{\varphi}_1\|_2^2+\|\varphi_1\|_2^2).$$

On the other hand, since φ_2 has polynomial growth, by Assumption 3.1.2, for every $x\in\mathbb{R}$,

$$\begin{aligned}\int_s^t\mathbb{E}[\varphi(x,X_u)^{2p}]^{1/p}du &= \int_s^t\left(\int_{-\infty}^{\infty}\varphi_1(z)^{2p}\varphi_2(z+x)^{2p}p_u(x_0,z+x)dz\right)^{1/p}du\\ &\leqslant \frac{\mathfrak{c}_4(p)}{t_0^{1/(2p)}}(t-s)\left(\int_{-\infty}^{\infty}\varphi_1(z)^{2p}dz\right)^{1/p}\end{aligned}$$

where the constant $\mathfrak{c}_4(p)>0$ depends on p, but not on x, s, t, φ_1 and t_0. Then,

$$\int_{-\infty}^{\infty}\left(\int_s^t\mathbb{E}[\varphi(x,X_u)^{2p}]^{1/p}du\right)\left(\frac{1}{x^{1/q}}\mathbf{1}_{[-1,1]}(x)+\frac{1}{x^2}\mathbf{1}_{\mathbb{R}\setminus[-1,1]}(x)\right)dx$$
$$\leqslant\frac{\mathfrak{c}_5(p)}{t_0^{1/(2p)}}(t-s)\left(\int_{-\infty}^{\infty}\varphi_1(z)^{2p}dz\right)^{1/p}$$

with

$$\mathfrak{c}_5(p)=\mathfrak{c}_4(p)\left(\int_{-1}^{1}\frac{dx}{x^{1/q}}+\int_{\mathbb{R}\setminus[-1,1]}\frac{dx}{x^2}\right)<\infty.$$

3.6.6.2 Proof of Lemma 3.6.2

Consider $s,t\in[t_0,T]$ such that $s<t$, and let $p_{s,t}$ (resp. $p_{t|s}$) be the density of (X_s,X_t) (resp. the conditional density of X_t with respect to X_s). Moreover, for the sake of readability, $p_s(x_0,.)$ is denoted by $p_s(.)$ in this proof. By Assumption 3.1.2,

$$\mathbb{E}(K_{h,x}(X_s)\varphi(X_s,X_t))^2 = \|K\|_1^2\left[\int_{-\infty}^{\infty}\frac{K_{h,x}(y)}{\|K_{h,x}\|_1}\left(\int_{-\infty}^{\infty}\varphi(y,z)p_{t|s}(z|y)dz\right)p_s(y)dy\right]^2$$

$$
\begin{aligned}
&\leqslant \|K\|_1 \int_{-\infty}^{\infty} |K_{h,x}(y)| \left(\int_{-\infty}^{\infty} \varphi(y,z) p_{t|s}(z|y) dz \right)^2 p_s(y)^2 dy \\
&\leqslant \|K\|_1 \sup_{s\in[t_0,T]} \left\{ \sup_{y\in\mathbb{R}} p_s(y) \right\} \int_{-\infty}^{\infty} |K_{h,x}(y)| \left(\int_{-\infty}^{\infty} \varphi(y,z)^2 p_{t|s}(z|y) dz \right) p_s(y) dy \\
&\leqslant \frac{\mathfrak{c}_{3.1.2,1}\|K\|_1}{t_0^{1/2}} \int_{-\infty}^{\infty} |K_{h,x}(y)| \int_{-\infty}^{\infty} \varphi(y,z)^2 p_{s,t}(y,z) dz dy.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\int_{-\infty}^{\infty} \mathbb{E}(K_{h,x}(X_s)\varphi(X_s,X_t))^2 dx &\leqslant \frac{\mathfrak{c}_{3.1.2,1}\|K\|_1}{t_0^{1/2}} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} |K_{h,x}(y)| dx \right) \left(\int_{-\infty}^{\infty} \varphi(y,z)^2 p_{s,t}(y,z) dz \right) dy \\
&= \frac{\mathfrak{c}_{3.1.2,1}\|K\|_1^2}{t_0^{1/2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \varphi(y,z)^2 p_{s,t}(y,z) dz dy \\
&= \frac{\mathfrak{c}_{3.1.2,1}\|K\|_1^2}{t_0^{1/2}} \mathbb{E}[\varphi(X_s,X_t)^2].
\end{aligned}
$$

3.6.7 Proof of Theorem 3.4.2

Throughout this subsection, $\mathcal{K}$ is a primitive function of the kernel K. The proof of Theorem 3.4.2,(1) relies on the following technical lemmas proved at the end of this subsection. The proof of Theorem 3.4.2,(2) is left to the reader because it is similar but simpler than the proof of Theorem 3.4.2,(1) detailed in this subsection.

Lemma 3.6.4. — *Under Assumptions 3.1.1, 3.1.2, 3.2.1 and 3.4.1,*

$$
\frac{1}{T-t_0} \int_{t_0}^{T} K_h(X_t - x) dX_t = \Phi_h(X,x) \ ; \ \forall x \in \mathbb{R}, \ \forall h > 0,
$$

where $(x,h,\varphi) \mapsto \Phi_h(\varphi,x)$ *is the map from* $\mathbb{R} \times (0,\infty) \times C^0([t_0,T];\mathbb{R})$ *into* $\mathbb{R}$ *defined by*

$$
\Phi_h(\varphi,x) := \frac{1}{T-t_0} \left[\mathcal{K}\left(\frac{\varphi(T)-x}{h} \right) - \mathcal{K}\left(\frac{\varphi(t_0)-x}{h} \right) - \frac{1}{2h^2} \int_{t_0}^{T} K'\left(\frac{\varphi(t)-x}{h} \right) \sigma(\varphi(t))^2 dt \right]
$$

for every $x \in \mathbb{R}$, $h > 0$ *and* $\varphi \in C^0([t_0,T];\mathbb{R})$. *Moreover,*

(1) For every $x \in \mathbb{R}$, $h > 0$ *and* $\varphi \in C^0([t_0,T];\mathbb{R})$,

$$
|\Phi_h(\varphi,x)| \leqslant \frac{2\|\mathcal{K}\|_\infty}{T-t_0} + \frac{\|\sigma\|_\infty^2 \|K'\|_\infty}{2h^2}.
$$

(2) For every $h > 0$ *and* $\varphi \in C^0([t_0,T];\mathbb{R})$,

$$
\|\Phi_h(\varphi,.)\|_{2,\delta}^2 \leqslant \frac{6\|\mathcal{K}\|_\infty^2}{(T-t_0)^2} + \frac{\|\delta\|_\infty \|\sigma\|_\infty^4 \|K'\|_2^2}{h^3}.
$$

(3) There exists a deterministic constant $\mathfrak{c}_{3.6.4,1} > 0$ *such that, for every* $h, h' > 0$,

$$
\mathbb{E}(\langle \Phi_h(X^1,.), \Phi_{h'}(X^2,.) \rangle_{2,\delta}^2) \leqslant \mathfrak{c}_{3.6.4,1} \mathfrak{m}(h')
$$

with

$$
\mathfrak{m}(h') = \mathbb{E}(\|\Phi_{h'}(X,.)\|_{2,\delta}^2).
$$

(4) There exists a deterministic constant $\mathfrak{c}_{3.6.4,2} > 0$ *such that, for every* $h > 0$ *and* $\varphi \in \mathbb{L}^2(\mathbb{R}, dx)$,

$$\mathbb{E}(\langle \Phi_h(X, .), \varphi \rangle_{2,\delta}^2) \leqslant \mathfrak{c}_{3.6.4,2} \|\varphi\|_{2,\delta}^2.$$

(5) There exists a deterministic constant $\mathfrak{c}_{3.6.4,3} > 0$ *such that, for every* $h, h' \in \mathcal{H}_N$,

$$|\langle \Phi_h(X, .), (bf)_{h'} \rangle_{2,\delta}| \leqslant \mathfrak{c}_{3.6.4,3} \quad \text{a.s.}$$

Lemma 3.6.5. — *Consider*

$$U_{h,h'}(N) := \sum_{i \neq j} \langle \Phi_h(X^i, .) - (bf)_h, \Phi_{h'}(X^j, .) - (bf)_{h'} \rangle_{2,\delta} \; ; \forall h, h' \in \mathcal{H}_N. \tag{3.14}$$

Under Assumptions 3.1.1, 3.1.2, 3.2.1 and 3.4.1, there exists a deterministic constant $\mathfrak{c}_{3.6.5} > 0$, *not depending on* N, *such that for every* $\theta \in (0,1)$ *and* $\lambda > 0$, *with probability larger than* $1 - 5.4|\mathcal{H}_N|e^{-\lambda}$,

$$\sup_{h \in \mathcal{H}_N} \left\{ \frac{|U_{h,h_0}(N)|}{N^2} - \frac{\theta \mathfrak{m}(h)}{N} \right\} \leqslant \frac{\mathfrak{c}_{3.6.5}(1+\lambda)^3}{\theta N}$$

and

$$\sup_{h \in \mathcal{H}_N} \left\{ \frac{|U_{h,h}(N)|}{N^2} - \frac{\theta \mathfrak{m}(h)}{N} \right\} \leqslant \frac{\mathfrak{c}_{3.6.5}(1+\lambda)^3}{\theta N}.$$

Lemma 3.6.6. — *Consider*

$$V_h(N) := \frac{1}{N} \sum_{i=1}^{N} \|\Phi_h(X^i, .) - (bf)_h\|_{2,\delta}^2 \; ; \forall h \in \mathcal{H}_N.$$

Under Assumptions 3.1.1, 3.1.2, 3.2.1 and 3.4.1, there exists a deterministic constant $\mathfrak{c}_{3.6.6} > 0$, *not depending on* N, *such that for every* $\theta \in (0,1)$ *and* $\lambda > 0$, *with probability larger than* $1 - 2|\mathcal{H}_N|e^{-\lambda}$,

$$\sup_{h \in \mathcal{H}_N} \left\{ \frac{1}{N} |V_h(N) - \mathfrak{m}(h)| - \frac{\theta \mathfrak{m}(h)}{N} \right\} \leqslant \frac{\mathfrak{c}_{3.6.6}(1+\lambda)}{\theta N}.$$

Lemma 3.6.7. — *Consider*

$$W_{h,h'}(N) := \langle \widehat{bf}_{N,h} - (bf)_h, (bf)_{h'} - bf \rangle_{2,\delta} \; ; \forall h, h' \in \mathcal{H}_N. \tag{3.15}$$

Under Assumptions 3.1.1, 3.1.2, 3.2.1 and 3.4.1, there exists a deterministic constant $\mathfrak{c}_{3.6.7} > 0$, *not depending on* N, *such that for every* $\theta \in (0,1)$ *and* $\lambda > 0$, *with probability larger than* $1 - 2|\mathcal{H}_N|e^{-\lambda}$,

$$\begin{aligned}
\sup_{h \in \mathcal{H}_N} \{|W_{h,h_0}(N)| - \theta \|(bf)_{h_0} - bf\|_{2,\delta}^2\} &\leqslant \frac{\mathfrak{c}_{3.6.7}(1+\lambda)^2}{\theta N}, \\
\sup_{h \in \mathcal{H}_N} \{|W_{h_0,h}(N)| - \theta \|(bf)_h - bf\|_{2,\delta}^2\} &\leqslant \frac{\mathfrak{c}_{3.6.7}(1+\lambda)^2}{\theta N} \quad \textit{and} \\
\sup_{h \in \mathcal{H}_N} \{|W_{h,h}(N)| - \theta \|(bf)_h - bf\|_{2,\delta}^2\} &\leqslant \frac{\mathfrak{c}_{3.6.7}(1+\lambda)^2}{\theta N}.
\end{aligned}$$

3.6.7.1 Steps of the proof

The proof of Theorem 3.4.2,(1) is dissected in four steps.

Step 1. This first step provides a suitable decomposition of $\|\widehat{bf}_{N,\widehat{h}} - bf\|_{2,\delta}^2$. First,

$$\begin{aligned}\|\widehat{bf}_{N,\widehat{h}} - bf\|_{2,\delta}^2 &= \|\widehat{bf}_{N,\widehat{h}} - \widehat{bf}_{N,h_0}\|_{2,\delta}^2 + \|\widehat{bf}_{N,h_0} - bf\|_{2,\delta}^2 \\ &\quad -2\langle \widehat{bf}_{N,h_0} - \widehat{bf}_{N,\widehat{h}}, \widehat{bf}_{N,h_0} - bf\rangle_{2,\delta}.\end{aligned}$$

Then, by (3.11) and the definition of pen(.) (see (3.12)), for any $h \in \mathcal{H}_N$,

$$\begin{aligned}\|\widehat{bf}_{N,\widehat{h}} - bf\|_{2,\delta}^2 &\leqslant \|\widehat{bf}_{N,h} - \widehat{bf}_{N,h_0}\|_{2,\delta}^2 + \mathrm{pen}(h) - \mathrm{pen}(\widehat{h}) \\ &\quad +\|\widehat{bf}_{N,h_0} - bf\|_{2,\delta}^2 - 2\langle \widehat{bf}_{N,h_0} - \widehat{bf}_{N,\widehat{h}}, \widehat{bf}_{N,h_0} - bf\rangle_{2,\delta} \\ &\leqslant \|\widehat{bf}_{N,h} - bf\|_{2,\delta}^2 + \mathrm{pen}(h) - \mathrm{pen}(\widehat{h}) \\ &\quad +\|\widehat{bf}_{N,h_0} - bf\|_{2,\delta}^2 - 2\langle \widehat{bf}_{N,h} - \widehat{bf}_{N,\widehat{h}}, \widehat{bf}_{N,h_0} - bf\rangle_{2,\delta} \\ &= \|\widehat{bf}_{N,h} - bf\|_{2,\delta}^2 - \psi_N(h) + \psi_N(\widehat{h}) \qquad (3.16)\end{aligned}$$

where

$$\psi_N(h) := 2\langle \widehat{bf}_{N,h} - bf, \widehat{bf}_{N,h_0} - bf\rangle_{2,\delta} - \mathrm{pen}(h).$$

Let's complete the decomposition of $\|\widehat{bf}_{N,\widehat{h}} - bf\|_{2,\delta}^2$ by writing

$$\psi_N(h) = 2(\psi_{1,N}(h) + \psi_{2,N}(h) + \psi_{3,N}(h)),$$

where

$$\begin{aligned}\psi_{1,N}(h) &:= \frac{1}{(T-t_0)^2N^2}\sum_{i=1}^N \left\langle \int_{t_0}^T K_h(X_s^i - \cdot)dX_s^i, \int_{t_0}^T K_{h_0}(X_s^i - \cdot)dX_s^i \right\rangle_{2,\delta} + \frac{U_{h,h_0}(N)}{N^2} - \frac{1}{2}\mathrm{pen}(h) \\ &= \frac{U_{h,h_0}(N)}{N^2}, \\ \psi_{2,N}(h) &:= -\frac{1}{N^2}\left(\sum_{i=1}^N \left\langle \frac{1}{T-t_0}\int_{t_0}^T K_{h_0}(X_s^i - \cdot)dX_s^i, (bf)_h \right\rangle_{2,\delta} + \right. \\ &\quad \left. + \sum_{i=1}^N \left\langle \frac{1}{T-t_0}\int_{t_0}^T K_h(X_s^i - \cdot)dX_s^i, (bf)_{h_0} \right\rangle_{2,\delta}\right) + \frac{1}{N}\langle (bf)_{h_0}, (bf)_h\rangle_{2,\delta} \text{ and} \\ \psi_{3,N}(h) &:= W_{h,h_0}(N) + W_{h_0,h}(N) + \langle (bf)_h - bf, (bf)_{h_0} - bf\rangle_{2,\delta}.\end{aligned}$$

Step 2. This step deals with bounds on $\mathbb{E}(\psi_{j,N}(h))$ and $\mathbb{E}(\psi_{j,N}(\widehat{h}))$ for $j = 1, 2, 3$.

- By Lemma 3.6.5, for any $\lambda > 0$ and $\theta \in (0,1)$, with probability larger than $1 - 5.4|\mathcal{H}_N|e^{-\lambda}$,

$$|\psi_{1,N}(h)| \leqslant \frac{\theta\mathfrak{m}(h)}{N} + \frac{\mathfrak{c}_{3.6.5}(1+\lambda)^3}{\theta N} \quad \text{and} \quad |\psi_{1,N}(\widehat{h})| \leqslant \frac{\theta\mathfrak{m}(\widehat{h})}{N} + \frac{\mathfrak{c}_{3.6.5}(1+\lambda)^3}{\theta N}.$$

- On the one hand, for any $h, h' \in \mathcal{H}_N$, consider

$$\Psi_{2,N}(h,h') := \frac{1}{N}\sum_{i=1}^N \langle \Phi_h(X^i, .), (bf)_{h'}\rangle_{2,\delta}.$$

By Lemma 3.6.4,

$$|\Psi_{2,N}(h,h')| \leqslant \frac{1}{N}\sum_{i=1}^{N}\left|\int_{-\infty}^{\infty}\Phi_h(X^i,x)(bf)_{h'}(x)\delta(x)dx\right| \leqslant \mathfrak{c}_{3.6.4,3} \quad \text{a.s.}$$

On the other hand,

$$|\langle (bf)_h,(bf)_{h_0}\rangle_{2,\delta}| \leqslant \|\delta\|_\infty \|K_h*(bf)\|_\infty \|K_{h_0}*(bf)\|_1 \leqslant \|\delta\|_\infty \|K\|_1^2 \|bf\|_\infty \|bf\|_1.$$

Then, there exists a deterministic constant $\mathfrak{c}_1>0$, not depending on N and h, such that

$$|\psi_{2,N}(h)| \leqslant \frac{\mathfrak{c}_1}{N} \quad \text{and} \quad |\psi_{2,N}(\widehat{h})| \leqslant \sup_{h'\in\mathcal{H}_N}|\psi_{2,N}(h')| \leqslant \frac{\mathfrak{c}_1}{N} \quad \text{a.s.}$$

- By Lemma 3.6.7 and Cauchy-Schwarz's inequality, with probability larger that $1-|\mathcal{H}_N|e^{-\lambda}$,

$$\begin{aligned}|\psi_{3,N}(h)| &\leqslant \frac{\theta}{4}(\|(bf)_h-bf\|_{2,\delta}^2+\|(bf)_{h_0}-bf\|_{2,\delta}^2)+\frac{8\mathfrak{c}_{3.6.7}(1+\lambda)^2}{\theta N}\\ &\qquad +2\times\frac{1}{2^{1/2}}\left(\frac{\theta}{2}\right)^{1/2}\|(bf)_h-bf\|_{2,\delta}\times\frac{1}{2^{1/2}}\left(\frac{2}{\theta}\right)^{1/2}\|(bf)_{h_0}-bf\|_{2,\delta}\\ &\leqslant \frac{\theta}{2}\|(bf)_h-bf\|_{2,\delta}^2+\left(\frac{\theta}{4}+\frac{1}{\theta}\right)\|(bf)_{h_0}-bf\|_{2,\delta}^2+\frac{8\mathfrak{c}_{3.6.7}(1+\lambda)^2}{\theta N}\end{aligned}$$

and

$$|\psi_{3,N}(\widehat{h})| \leqslant \frac{\theta}{2}\|(bf)_{\widehat{h}}-bf\|_{2,\delta}^2+\left(\frac{\theta}{4}+\frac{1}{\theta}\right)\|(bf)_{h_0}-bf\|_{2,\delta}^2+\frac{8\mathfrak{c}_{3.6.7}(1+\lambda)^2}{\theta N}.$$

Step 3. Let us establish that there exist two deterministic constants $\mathfrak{c}_2,\bar{\mathfrak{c}}_2>0$, not depending on N and θ, such that with probability larger than $1-\bar{\mathfrak{c}}_2|\mathcal{H}_N|e^{-\lambda}$,

$$\sup_{h\in\mathcal{H}_N}\left\{\|\widehat{bf}_{N,h}-bf\|_{2,\delta}^2-(1+\theta)\left(\|(bf)_h-bf\|_{2,\delta}^2+\frac{\mathfrak{m}(h)}{N}\right)\right\} \leqslant \frac{\mathfrak{c}_2(1+\lambda)^3}{\theta N}$$

and

$$\sup_{h\in\mathcal{H}_N}\left\{\|(bf)_h-bf\|_{2,\delta}^2+\frac{\mathfrak{m}(h)}{N}-\frac{1}{1-\theta}\|\widehat{bf}_{N,h}-bf\|_{2,\delta}^2\right\} \leqslant \frac{\mathfrak{c}_2(1+\lambda)^3}{\theta(1-\theta)N}.$$

On the one hand, note that

$$\|\widehat{bf}_{N,h}-bf\|_{2,\delta}^2-(1+\theta)\left(\|(bf)_h-bf\|_{2,\delta}^2+\frac{\mathfrak{m}(h)}{N}\right)$$

can be written

$$\|\widehat{bf}_{N,h}-(bf)_h\|_{2,\delta}^2-\frac{(1+\theta)\mathfrak{m}(h)}{N}+2W_h(N)-\theta\|(bf)_h-bf\|_{2,\delta}^2,$$

where $W_h(N):=W_{h,h}(N)$ (see (3.15)). Moreover, for any $h\in\mathcal{H}_N$,

$$\|\widehat{bf}_{N,h}-(bf)_h\|_{2,\delta}^2=\frac{U_h(N)}{N^2}+\frac{V_h(N)}{N} \tag{3.17}$$

with $U_h(N)=U_{h,h}(N)$ (see (3.14)). So, with probability larger than $1-\bar{\mathfrak{c}}_2|\mathcal{H}_N|e^{-\lambda}$,

$$\sup_{h\in\mathcal{H}_N}\left\{\left|\|\widehat{bf}_{N,h}-(bf)_h\|_{2,\delta}^2-\frac{\mathfrak{m}(h)}{N}\right|-\frac{\theta\mathfrak{m}(h)}{N}\right\} \leqslant \frac{2(\mathfrak{c}_{3.6.5}+\mathfrak{c}_{3.6.6})(1+\lambda)^3}{\theta N}$$

by Lemmas 3.6.5 and 3.6.6, and then

$$\sup_{h\in\mathcal{H}_N}\left\{\|\widehat{bf}_{N,h}-bf\|_{2,\delta}^2-(1+\theta)\left(\|(bf)_h-bf\|_{2,\delta}^2+\frac{\mathfrak{m}(h)}{N}\right)\right\}\leqslant\frac{\mathfrak{c}_2(1+\lambda)^3}{\theta N}$$

by Lemma 3.6.7. On the other hand, for any $h\in\mathcal{H}_N$,

$$\|(bf)_h-bf\|_{2,\delta}^2=\|\widehat{bf}_{N,h}-bf\|_{2,\delta}^2-\|\widehat{bf}_{N,h}-(bf)_h\|_{2,\delta}^2-W_h(N).$$

Then,

$$(1-\theta)\left(\|(bf)_h-bf\|_{2,\delta}^2+\frac{\mathfrak{m}(h)}{N}\right)-\|\widehat{bf}_{N,h}-bf\|_{2,\delta}^2\leqslant|W_h(N)|-\theta\|(bf)_h-bf\|_{2,\delta}^2+\Lambda_h(N)-\frac{\theta\mathfrak{m}(h)}{N}$$

where

$$\Lambda_h(N):=\left|\|\widehat{bf}_{N,h}-(bf)_h\|_{2,\delta}^2-\frac{\mathfrak{m}(h)}{N}\right|.$$

By Equality (3.17),

$$\Lambda_h(N)=\left|\frac{U_h(N)}{N^2}+\frac{V_h(N)}{N}-\frac{\mathfrak{m}(h)}{N}\right|.$$

By Lemmas 3.6.6 and 3.6.5, there exist two deterministic constants $\mathfrak{c}_3,\bar{\mathfrak{c}}_3>0$, not depending N and θ, such that with probability larger than $1-\bar{\mathfrak{c}}_3|\mathcal{H}_N|e^{-\lambda}$,

$$\sup_{h\in\mathcal{H}_N}\left\{\Lambda_h(N)-\theta\frac{\mathfrak{m}(h)}{N}\right\}\leqslant\frac{\mathfrak{c}_3(1+\lambda)^3}{\theta N}.$$

By Lemma 3.6.7, with probability larger than $1-2|\mathcal{H}_N|e^{-\lambda}$,

$$\sup_{h\in\mathcal{H}_N}\{|W_h(N)|-\theta\|(bf)_h-bf\|_{2,\delta}^2\}\leqslant\frac{\mathfrak{c}_{3.6.7}(1+\lambda)^2}{\theta N}.$$

Therefore, with probability larger than $1-\bar{\mathfrak{c}}_2|\mathcal{H}_N|e^{-\lambda}$,

$$\sup_{h\in\mathcal{H}_N}\left\{\|(bf)_h-bf\|_{2,\delta}^2+\frac{\mathfrak{m}(h)}{N}-\frac{1}{1-\theta}\|\widehat{bf}_{N,h}-bf\|_{2,\delta}^2\right\}\leqslant\frac{\mathfrak{c}_2(1+\lambda)^3}{\theta(1-\theta)N}.$$

Step 4. By step 2, there exist two deterministic constants $\mathfrak{c}_4,\bar{\mathfrak{c}}_4>0$, not depending on N, θ, h and h_0, such that with probability larger than $1-\bar{\mathfrak{c}}_4|\mathcal{H}_N|e^{-\lambda}$,

$$|\psi_N(h)|\leqslant\theta\left(\|(bf)_h-bf\|_{2,\delta}^2+\frac{\mathfrak{m}(h)}{N}\right)+\left(\frac{\theta}{2}+\frac{2}{\theta}\right)\|(bf)_{h_0}-bf\|_{2,\delta}^2+\frac{\mathfrak{c}_4(1+\lambda)^3}{\theta N}$$

and

$$|\psi_N(\widehat{h})|\leqslant\theta\left(\|(bf)_{\widehat{h}}-bf\|_{2,\delta}^2+\frac{\mathfrak{m}(\widehat{h})}{N}\right)+\left(\frac{\theta}{2}+\frac{2}{\theta}\right)\|(bf)_{h_0}-bf\|_{2,\delta}^2+\frac{\mathfrak{c}_4(1+\lambda)^3}{\theta N}.$$

Then, by step 3, there exist two deterministic constants $\mathfrak{c}_5,\bar{\mathfrak{c}}_5>0$, not depending on N, θ, h and h_0, such that with probability larger than $1-\bar{\mathfrak{c}}_5|\mathcal{H}_N|e^{-\lambda}$,

$$|\psi_N(h)|\leqslant\frac{\theta}{1-\theta}\|\widehat{bf}_{N,h}-bf\|_{2,\delta}^2+\left(\frac{\theta}{2}+\frac{2}{\theta}\right)\|(bf)_{h_0}-bf\|_{2,\delta}^2+\mathfrak{c}_5\left(\frac{1}{\theta}+\frac{1}{1-\theta}\right)\frac{(1+\lambda)^3}{N}$$

and

$$|\psi_N(\widehat{h})|\leqslant\frac{\theta}{1-\theta}\|\widehat{bf}_{N,\widehat{h}}-bf\|_{2,\delta}^2+\left(\frac{\theta}{2}+\frac{2}{\theta}\right)\|(bf)_{h_0}-bf\|_{2,\delta}^2+\mathfrak{c}_5\left(\frac{1}{\theta}+\frac{1}{1-\theta}\right)\frac{(1+\lambda)^3}{N}.$$

By the decomposition (3.16), there exist two deterministic constants $\mathfrak{c}_6, \bar{\mathfrak{c}}_6 > 0$, not depending on N, θ, h and h_0, such that with probability larger than $1 - \bar{\mathfrak{c}}_6 |\mathcal{H}_N| e^{-\lambda}$,

$$\begin{aligned}
\|\widehat{bf}_{N,\widehat{h}} - bf\|_{2,\delta}^2 &\leqslant \|\widehat{bf}_{N,h} - bf\|_{2,\delta}^2 + |\psi_N(h)| + |\psi_N(\widehat{h})| \\
&\leqslant \left(1 + \frac{\theta}{1-\theta}\right) \|\widehat{bf}_{N,h} - bf\|_{2,\delta}^2 + \frac{\theta}{1-\theta} \|\widehat{bf}_{N,\widehat{h}} - bf\|_{2,\delta}^2 \\
&\qquad + \frac{\mathfrak{c}_6}{\theta} \|(bf)_{h_0} - bf\|_{2,\delta}^2 + \frac{\mathfrak{c}_6}{\theta(1-\theta)} \cdot \frac{(1+\lambda)^3}{N}.
\end{aligned}$$

This concludes the proof.

3.6.7.2 Proof of Lemma 3.6.4

First of all, for any $x \in \mathbb{R}$ and $h > 0$, by Itô's formula,

$$\mathcal{K}\left(\frac{X_T - x}{h}\right) = \mathcal{K}\left(\frac{X_{t_0} - x}{h}\right) + \int_{t_0}^T K_h(X_t - x) dX_t + \frac{1}{2h^2} \int_{t_0}^T K'\left(\frac{X_t - x}{h}\right) d\langle X \rangle_t.$$

So,

$$\begin{aligned}
\int_{t_0}^T K_h(X_t - x) dX_t &= \mathcal{K}\left(\frac{X_T - x}{h}\right) - \mathcal{K}\left(\frac{X_{t_0} - x}{h}\right) \\
&\qquad - \frac{1}{2h^2} \int_{t_0}^T K'\left(\frac{X_t - x}{h}\right) \sigma(X_t)^2 dt = (T - t_0) \Phi_h(X, x).
\end{aligned}$$

Lemma 3.6.4.(1) is a straightforward consequence of the previous equality and Lemma 3.6.4.(2) is easy to establish : for every $h > 0$ and $\varphi \in C^0([t_0, T]; \mathbb{R})$,

$$\begin{aligned}
(T - t_0)^2 \|\Phi_h(\varphi, .)\|_{2,\delta}^2 &\leqslant 2 \int_{-\infty}^{\infty} \mathcal{K}\left(\frac{\varphi(T) - x}{h}\right)^2 \delta(x) dx + 4 \int_{-\infty}^{\infty} \mathcal{K}\left(\frac{\varphi(t_0) - x}{h}\right)^2 \delta(x) dx \\
&\qquad + \frac{1}{h^4} \int_{-\infty}^{\infty} \left[\int_{t_0}^T K'\left(\frac{\varphi(t) - x}{h}\right) \sigma(\varphi(t))^2 dt\right]^2 \delta(x) dx \\
&\leqslant 6\|\mathcal{K}\|_\infty^2 + \frac{T - t_0}{h^4} \int_{t_0}^T \sigma(\varphi(t))^4 \int_{-\infty}^{\infty} K'\left(\frac{\varphi(t) - x}{h}\right)^2 \delta(x) dx dt \\
&\leqslant 6\|\mathcal{K}\|_\infty^2 + \frac{(T - t_0)^2 \|\delta\|_\infty \|\sigma\|_\infty^4 \|K'\|_2^2}{h^3}.
\end{aligned}$$

Let us prove Lemma 3.6.4.(3,4,5). First, for any $h, h' > 0$,

$$\begin{aligned}
&\mathbb{E}(\langle \Phi_h(X^1, .), \Phi_{h'}(X^2, .) \rangle_{2,\delta}^2) \\
&\qquad = \frac{1}{(T - t_0)^4} \mathbb{E}\left[\left(\int_{-\infty}^{\infty} \left(\int_{t_0}^T K_h(X_t^1 - x) dX_t^1\right) \left(\int_{t_0}^T K_{h'}(X_t^2 - x) dX_t^2\right) \delta(x) dx\right)^2\right] \\
&\qquad \leqslant \frac{2}{(T - t_0)^4} (\mathbb{E}(A_{h,h'}^2) + \mathbb{E}(B_{h,h'}^2))
\end{aligned}$$

with

$$\begin{aligned}
A_{h,h'} &:= \int_{-\infty}^{\infty} \left(\int_{t_0}^T K_h(X_t^1 - x) \sigma(X_t^1) dW_t^1\right) \left(\int_{t_0}^T K_{h'}(X_t^2 - x) dX_t^2\right) \delta(x) dx \text{ and} \\
B_{h,h'} &:= \int_{-\infty}^{\infty} \left(\int_{t_0}^T K_h(X_t^1 - x) b(X_t^1) dt\right) \left(\int_{t_0}^T K_{h'}(X_t^2 - x) dX_t^2\right) \delta(x) dx.
\end{aligned}$$

Bound on $\mathbb{E}(A^2_{h,h'})$. Since (X^1, W^1) and X^2 are independent,

$$\begin{aligned}\mathbb{E}(A^2_{h,h'}) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{E}\left[\left(\int_{t_0}^T K_h(X^1_t - x)\sigma(X^1_t)dW^1_t\right)\left(\int_{t_0}^T K_h(X^1_t - y)\sigma(X^1_t)dW^1_t\right)\right] \\ &\qquad \times \mathbb{E}\left[\left(\int_{t_0}^T K_{h'}(X^2_t - x)dX^2_t\right)\left(\int_{t_0}^T K_{h'}(X^2_t - y)dX^2_t\right)\right]\delta(x)\delta(y)dxdy.\end{aligned}$$

On the one hand, for every $x, y \in \mathbb{R}$, by the isometry property of Itô's integral and the definition of f,

$$\begin{aligned}&\mathbb{E}\left[\left(\int_{t_0}^T K_h(X^1_t - x)\sigma(X^1_t)dW^1_t\right)\left(\int_{t_0}^T K_h(X^1_t - y)\sigma(X^1_t)dW^1_t\right)\right] \\ &\quad = \int_{t_0}^T \mathbb{E}(K_h(X^1_t - x)K_h(X^1_t - y)\sigma(X^1_t)^2)dt = (T - t_0)\int_{-\infty}^{\infty} K_h(z - x)K_h(z - y)\sigma(z)^2 f(z)dz.\end{aligned}$$

Then,

$$\begin{aligned}\mathbb{E}(A^2_{h,h'}) &= (T - t_0)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K_h(z - x)K_h(z - y)\sigma(z)^2 f(z) \\ &\qquad \times \mathbb{E}\left[\left(\int_{t_0}^T K_{h'}(X^2_t - x)dX^2_t\right)\left(\int_{t_0}^T K_{h'}(X^2_t - y)dX^2_t\right)\right]\delta(x)\delta(y)dxdydz \\ &= (T - t_0)\int_{-\infty}^{\infty} \sigma(z)^2 f(z)\mathbb{E}\left[\left(\int_{-\infty}^{\infty} K_h(z - x)\delta(x)\int_{t_0}^T K_{h'}(X^2_t - x)dX^2_t dx\right)^2\right]dz.\end{aligned}$$

On the other hand, for every $z \in \mathbb{R}$, $x \mapsto |K_h(z - x)|/\|K\|_1$ is a density function. Then, by Jensen's inequality,

$$\begin{aligned}\mathbb{E}(A^2_{h,h'}) &\leqslant (T - t_0)\|K\|_1\int_{-\infty}^{\infty}\sigma(z)^2 f(z)\int_{-\infty}^{\infty}|K_h(z - x)|\delta(x)^2\mathbb{E}\left[\left(\int_{t_0}^T K_{h'}(X^2_t - x)dX^2_t\right)^2\right]dxdz \\ &\leqslant (T - t_0)\|\sigma^2 f\|_\infty\|K\|_1^2\|\delta\|_\infty\int_{-\infty}^{\infty}\delta(x)\mathbb{E}\left[\left(\int_{t_0}^T K_{h'}(X^2_t - x)dX^2_t\right)^2\right]dx \\ &\leqslant (T - t_0)^3\|\sigma^2 f\|_\infty\|K\|_1^2\|\delta\|_\infty\mathfrak{m}(h').\end{aligned}$$

Bound on $\mathbb{E}(B^2_{h,h'})$. Since $x \mapsto |K_h(X_t(\omega) - x)|/\|K\|_1$ is a density function for every $(t, \omega) \in [t_0, T] \times \Omega$, by Jensen's inequality,

$$\begin{aligned}\mathbb{E}(B^2_{h,h'}) &= \mathbb{E}\left[\left(\int_{t_0}^T\int_{-\infty}^{\infty} K_h(X^1_t - x)b(X^1_t)\delta(x)\int_{t_0}^T K_{h'}(X^2_s - x)dX^2_s dxdt\right)^2\right] \\ &\leqslant (T - t_0)\|K\|_1\int_{t_0}^T\int_{-\infty}^{\infty}\mathbb{E}(|K_h(X^1_t - x)|b(X^1_t)^2)\delta(x)^2\mathbb{E}\left[\left(\int_{t_0}^T K_{h'}(X^2_s - x)dX^2_s\right)^2\right]dxdt \\ &= (T - t_0)^2\|K\|_1\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}|K_h(z - x)|b(z)^2 f(z)dz\right)\delta(x)^2\mathbb{E}\left[\left(\int_{t_0}^T K_{h'}(X^2_s - x)dX^2_s\right)^2\right]dx \\ &\leqslant (T - t_0)^2\|b^2 f\|_\infty\|K\|_1^2\|\delta\|_\infty\int_{-\infty}^{\infty}\delta(x)\mathbb{E}\left[\left(\int_{t_0}^T K_{h'}(X^2_s - x)dX^2_s\right)^2\right]dx \\ &\leqslant (T - t_0)^4\|b^2 f\|_\infty\|K\|_1^2\|\delta\|_\infty\mathfrak{m}(h').\end{aligned}$$

Now, for any $h > 0$ and $\varphi \in \mathbb{L}^2(\mathbb{R}, dx)$,

$$\mathbb{E}(\langle\Phi_h(X, .), \varphi\rangle^2_{2,\delta}) = \frac{1}{(T - t_0)^2}\mathbb{E}\left[\left(\int_{-\infty}^{\infty}\varphi(x)\delta(x)\int_{t_0}^T K_h(X_t - x)dX_t dx\right)^2\right]$$

$$\leqslant \frac{2}{(T-t_0)^2}(\mathbb{E}(C_h^2)+\mathbb{E}(D_h^2))$$

with

$$\begin{aligned} C_h &:= \int_{-\infty}^{\infty}\varphi(x)\delta(x)\int_{t_0}^{T}K_h(X_t-x)\sigma(X_t)dW_t dx \text{ and} \\ D_h &:= \int_{-\infty}^{\infty}\varphi(x)\delta(x)\int_{t_0}^{T}K_h(X_t-x)b(X_t)dt dx. \end{aligned}$$

Bound on $\mathbb{E}(C_h^2)$. By the isometry property of Itô's integral and the definition of f,

$$\begin{aligned} \mathbb{E}(C_h^2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\varphi(x)\varphi(y)\delta(x)\delta(y)\int_{t_0}^{T}\mathbb{E}(K_h(X_t-x)K_h(X_t-y)\sigma(X_t)^2)dtdxdy \\ &= (T-t_0)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\varphi(x)\varphi(y)\delta(x)\delta(y)\int_{-\infty}^{\infty}K_h(z-x)K_h(z-y)\sigma(z)^2f(z)dzdxdy \\ &= (T-t_0)\int_{-\infty}^{\infty}(K_h*(\varphi\delta))(z)^2\sigma(z)^2f(z)dz \leqslant (T-t_0)\|\sigma^2 f\|_\infty\|K\|_1^2\|\delta\|_\infty\|\varphi\|_{2,\delta}^2. \end{aligned}$$

Bound on $\mathbb{E}(D_h^2)$. By the definition of f,

$$\begin{aligned} \mathbb{E}(D_h^2) &= \mathbb{E}\left[\left(\int_{t_0}^{T}b(X_t)\int_{-\infty}^{\infty}K_h(X_t-x)\varphi(x)\delta(x)dxdt\right)^2\right] \leqslant (T-t_0)\int_{t_0}^{T}\mathbb{E}(b(X_t)^2(K_h*(\varphi\delta))(X_t)^2)dt \\ &\leqslant (T-t_0)^2\int_{-\infty}^{\infty}(K_h*(\varphi\delta))(z)^2b(z)^2f(z)dz \leqslant (T-t_0)^2\|b^2f\|_\infty\|K\|_1^2\|\delta\|_\infty\|\varphi\|_{2,\delta}^2. \end{aligned}$$

Finally, since X is a semi-martingale and since the map $(t,\omega,x)\mapsto K_h(X_t(\omega)-x)(bf)_{h'}(x)\delta(x)$ is measurable and bounded for any $h,h'\in\mathcal{H}_N$, by the stochastic Fubini theorem and Itô's formula,

$$\begin{aligned} (T-t_0)\langle\Phi_h(X,.),(bf)_{h'}\rangle_{2,\delta} &= (T-t_0)\int_{-\infty}^{\infty}\Phi_h(X,x)(bf)_{h'}(x)\delta(x)dx \\ &= \int_{t_0}^{T}\int_{-\infty}^{\infty}K_h(X_t-x)(bf)_{h'}(x)\delta(x)dxdX_t \quad \text{a.s.} \\ &= \int_{t_0}^{T}[K_h*((bf)_{h'}\delta)](X_t)dX_t \\ &= \Psi_{h,h'}(X_T)-\Psi_{h,h'}(X_{t_0})-\frac{1}{2}\int_{t_0}^{T}\psi'_{h,h'}(X_t)\sigma(X_t)^2dt \end{aligned}$$

where

$$\psi_{h,h'} := K_h*((bf)_{h'}\delta), \quad \text{and} \quad \Psi_{h,h'} := \mathcal{K}(./h)*((bf)_{h'}\delta)$$

is a primitive function of $\psi_{h,h'}$. On the one hand,

$$\begin{aligned} \psi'_{h,h'} &= K_h*((bf)_{h'}\delta)' \\ &= K_h*((bf)_{h'}\delta')+K_h*((K_{h'}*(bf)')\delta) \\ &= K_h*((bf)_{h'}\delta')+K_h*((K_{h'}*(bf'))\delta)+K_h*((K_{h'}*(b'f))\delta). \end{aligned}$$

Then, since bf, bf' and $b'f$ are bounded under Assumption 3.1.2,

$$\begin{aligned} \|\psi'_{h,h'}\|_\infty &\leqslant \|K_h\|_1\|K_{h'}*(bf)\|_\infty\|\delta'\|_\infty+\|K_h\|_1\|K_{h'}*(bf')\|_\infty\|\delta\|_\infty+\|K_h\|_1\|K_{h'}*(b'f))\|_\infty\|\delta\|_\infty \\ &\leqslant \|K\|_1^2\|bf\|_\infty\|\delta'\|_\infty+\|K\|_1^2\|bf'\|_\infty\|\delta\|_\infty+\|K\|_1^2\|b'f\|_\infty\|\delta\|_\infty<\infty. \end{aligned}$$

On the other hand,

$$\|\Psi_{h,h'}\|_\infty \leqslant \|\mathcal{K}(./h)\|_\infty \|(K_{h'} * (bf))\delta\|_1 \leqslant \|\mathcal{K}\|_\infty \|\delta\|_\infty \|K\|_1 \|bf\|_1 < \infty.$$

This concludes the proof because

$$(T-t_0)|\langle \Phi_h(X,.),(bf)_{h'}\rangle_{2,\delta}| \leqslant 2\|\Psi_{h,h'}\|_\infty + \frac{T-t_0}{2}\|\sigma\|_\infty^2 \|\psi'_{h,h'}\|_\infty \quad \text{a.s.}$$

3.6.7.3 Proof of Lemma 3.6.5

For any $h, h' \in \mathcal{H}_N$,

$$U_{h,h'}(N) = \sum_{i \neq j} g_{h,h'}(X^i, X^j)$$

with, for every $\varphi_1, \varphi_2 \in E = C^0([0,T];\mathbb{R})$,

$$g_{h,h'}(\varphi_1,\varphi_2) := \langle \Phi_h(\varphi_1,.) - (bf)_h, \Phi_{h'}(\varphi_2,.) - (bf)_{h'}\rangle_{2,\delta}.$$

On the one hand, since $\mathbb{E}(g_{h,h'}(\varphi, X)) = 0$ for every $\varphi \in E$, by Giné and Nickl [115], Theorem 3.4.8, there exists a universal constant $\mathfrak{m} \geqslant 1$ such that for any $\lambda > 0$, with probability larger than $1 - 5.4e^{-\lambda}$,

$$\frac{|U_{h,h'}(N)|}{N^2} \leqslant \frac{\mathfrak{m}}{N^2}(\mathfrak{c}_{h,h'}(N)\lambda^{1/2} + \mathfrak{d}_{h,h'}(N)\lambda + \mathfrak{b}_{h,h'}(N)\lambda^{3/2} + \mathfrak{a}_{h,h'}(N)\lambda^2)$$

where the constants $\mathfrak{a}_{h,h'}(N)$, $\mathfrak{b}_{h,h'}(N)$, $\mathfrak{c}_{h,h'}(N)$ and $\mathfrak{d}_{h,h'}(N)$ are defined and controlled later. First, note that

$$U_{h,h'}(N) = \sum_{i \neq j} (\overline{g}_{h,h'}(X^i,X^j) - \widetilde{g}_{h,h'}(X^i) - \widetilde{g}_{h',h}(X^j) + \mathbb{E}(\overline{g}_{h,h'}(X^i,X^j))) \tag{3.18}$$

where, for every $\eta, \eta' \in \mathcal{H}_N$ and $\varphi_1, \varphi_2, \psi \in E$,

$$\overline{g}_{\eta,\eta'}(\varphi_1,\varphi_2) := \langle \Phi_\eta(\varphi_1,.), \Phi_{\eta'}(\varphi_2,.)\rangle_{2,\delta} \quad \text{and} \quad \widetilde{g}_{\eta,\eta'}(\psi) := \langle \Phi_\eta(\psi,.),(bf)_{\eta'}\rangle_{2,\delta} = \mathbb{E}(\overline{g}_{\eta,\eta'}(\psi,X)).$$

Let us now control $\mathfrak{a}_{h,h'}(N)$, $\mathfrak{b}_{h,h'}(N)$, $\mathfrak{c}_{h,h'}(N)$ and $\mathfrak{d}_{h,h'}(N)$:

- **The constant $\mathfrak{a}_{h,h'}(N)$.** Consider

$$\mathfrak{a}_{h,h'}(N) := \sup_{\varphi_1,\varphi_2 \in E} |g_{h,h'}(\varphi_1,\varphi_2)|.$$

By (3.18), Cauchy-Schwarz's inequality and Lemma 3.6.4,

$$\begin{aligned}
\mathfrak{a}_{h,h'}(N) &\leqslant 4 \sup_{\varphi_1,\varphi_2 \in E} |\langle \Phi_h(\varphi_1,.), \Phi_{h'}(\varphi_2,.)\rangle_{2,\delta}| \leqslant 4 \left(\sup_{\varphi_1 \in E} \|\Phi_h(\varphi_1,.)\|_{2,\delta}\right)\left(\sup_{\varphi_2 \in E} \|\Phi_{h'}(\varphi_2,.)\|_{2,\delta}\right) \\
&\leqslant 4\left[\frac{6\|\mathcal{K}\|_\infty^2}{(T-t_0)^2} + \frac{\|\delta\|_\infty \|\sigma\|_\infty^4 \|K'\|_2^2}{h^3}\right]^{1/2}\left[\frac{6\|\mathcal{K}\|_\infty^2}{(T-t_0)^2} + \frac{\|\delta\|_\infty \|\sigma\|_\infty^4 \|K'\|_2^2}{(h')^3}\right]^{1/2} \leqslant \frac{\mathfrak{c}_1}{h_0^3}
\end{aligned}$$

with

$$\mathfrak{c}_1 = 4\left[\frac{6\|\mathcal{K}\|_\infty^2}{(T-t_0)^2} + \|\delta\|_\infty \|\sigma\|_\infty^4 \|K'\|_2^2\right].$$

So, since $(Nh_0^3)^{-1} \leqslant 1$,

$$\frac{\mathfrak{a}_{h,h'}(N)\lambda^2}{N^2} \leqslant \frac{\mathfrak{c}_1\lambda^2}{N^2 h_0^3} \leqslant \frac{\mathfrak{c}_1\lambda^2}{N}.$$

- **The constant $\mathfrak{b}_{h,h'}(N)$.** Consider

$$\mathfrak{b}_{h,h'}(N)^2 := N \sup_{\varphi \in E} \mathbb{E}(g_{h,h'}(\varphi, X)^2).$$

By (3.18), Cauchy-Schwarz's inequality and Lemma 3.6.4,

$$\begin{aligned}\mathfrak{b}_{h,h'}(N)^2 &\leqslant 16N \sup_{\varphi \in E} \mathbb{E}(\langle \Phi_h(\varphi, .), \Phi_{h'}(X, .)\rangle_{2,\delta}^2) \\ &\leqslant 16N\mathbb{E}(\|\Phi_{h'}(X, .)\|_{2,\delta}^2) \sup_{\varphi \in E} \|\Phi_h(\varphi, .)\|_{2,\delta}^2 \leqslant \frac{\mathfrak{c}_2 \mathfrak{m}(h')N}{h^3} \quad \text{with} \quad \mathfrak{c}_2 = 4\mathfrak{c}_1.\end{aligned}$$

So, for any $\theta \in (0, 1)$, since $(Nh_0^3)^{-1} \leqslant 1$,

$$\begin{aligned}\frac{\mathfrak{b}_{h,h'}(N)\lambda^{3/2}}{N^2} &\leqslant 2\left(\frac{\theta}{3\mathfrak{m}}\right)^{1/2} \frac{\mathfrak{m}(h')^{1/2}}{Nh^{3/2}} \times \left(\frac{3\mathfrak{m}}{\theta}\right)^{1/2} \frac{\mathfrak{c}_2^{1/2}\lambda^{3/2}}{N^{1/2}} \\ &\leqslant \frac{\theta \mathfrak{m}(h')}{3\mathfrak{m}N^2h^3} + \frac{3\mathfrak{c}_2\mathfrak{m}\lambda^3}{\theta N} \leqslant \frac{\theta \mathfrak{m}(h')}{3\mathfrak{m}N} + \frac{3\mathfrak{c}_2\mathfrak{m}\lambda^3}{\theta N}.\end{aligned}$$

- **The constant $\mathfrak{c}_{h,h'}(N)$.** Consider

$$\mathfrak{c}_{h,h'}(N)^2 := N^2\mathbb{E}(g_{h,h'}(X^1, X^2)^2).$$

By (3.18) and Lemma 3.6.4,

$$\begin{aligned}\mathfrak{c}_{h,h'}(N)^2 &\leqslant 16N^2\mathbb{E}(\langle \Phi_h(X^1, .), \Phi_{h'}(X^2, .)\rangle_{2,\delta}^2) \\ &\leqslant \mathfrak{c}_3\mathfrak{m}(h')N^2 \quad \text{with} \quad \mathfrak{c}_3 = 16\mathfrak{c}_{3.6.4,1}.\end{aligned}$$

So, as previously,

$$\frac{\mathfrak{c}_{h,h'}(N)\lambda^{1/2}}{N^2} \leqslant \frac{\theta \mathfrak{m}(h')}{3\mathfrak{m}N} + \frac{3\mathfrak{c}_3\mathfrak{m}\lambda}{\theta N}.$$

- **The constant $\mathfrak{d}_{h,h'}(N)$.** Consider

$$\mathfrak{d}_{h,h'}(N) := \sup_{(a,b)\in\mathcal{A}} \mathbb{E}\left[\sum_{i<j} a_i(X^i)b_j(X^j)g_{h,h'}(X^i, X^j)\right],$$

where

$$\mathcal{A} := \left\{(a, b) : \sum_{i=1}^{N-1} \mathbb{E}(a_i(X^i)^2) \leqslant 1 \text{ and } \sum_{j=2}^{N} \mathbb{E}(b_j(X^j)^2) \leqslant 1\right\}.$$

By (3.18), Cauchy-Schwarz's inequality, Jensen's inequality and Lemma 3.6.4,

$$\begin{aligned}\mathfrak{d}_{h,h'}(N) &\leqslant 4 \sup_{(a,b)\in\mathcal{A}} \mathbb{E}\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} |a_i(X^i)b_j(X^j)\overline{g}_{h,h'}(X^i, X^j)|\right) \\ &\leqslant 4N\mathbb{E}(\langle \Phi_h(X^1, .), \Phi_{h'}(X^2, .)\rangle_{2,\delta}^2)^{1/2} \leqslant \mathfrak{c}_3^{1/2}\mathfrak{m}(h')^{1/2}N.\end{aligned}$$

So, as previously,

$$\frac{\mathfrak{d}_{h,h'}(N)\lambda}{N^2} \leqslant \frac{\theta \mathfrak{m}(h')}{3\mathfrak{m}N} + \frac{3\mathfrak{c}_3\mathfrak{m}\lambda^2}{\theta N}.$$

Therefore, there exists a deterministic constant $\mathfrak{c}_4 > 0$, not depending on N, h and h', such that with probability larger than $1 - 5.4e^{-\lambda}$,

$$\frac{|U_{h,h'}(N)|}{N^2} \leqslant \frac{\theta \mathfrak{m}(h')}{N} + \frac{\mathfrak{c}_4(1+\lambda)^3}{\theta N}.$$

In conclusion, with probability larger than $1 - 5.4|\mathcal{H}_N|e^{-\lambda}$,

$$\sup_{h \in \mathcal{H}_N} \left\{ \frac{|U_{h,h_0}(N)|}{N^2} - \frac{\theta \mathfrak{m}(h)}{N} \right\} \leqslant \frac{\mathfrak{c}_4(1+\lambda)^3}{\theta N}$$

and

$$\sup_{h \in \mathcal{H}_N} \left\{ \frac{|U_{h,h}(N)|}{N^2} - \frac{\theta \mathfrak{m}(h)}{N} \right\} \leqslant \frac{\mathfrak{c}_4(1+\lambda)^3}{\theta N}.$$

3.6.7.4 Proof of Lemma 3.6.6

First, the two following results are used several times in the sequel :

$$\begin{aligned} \|(bf)_h\|_{2,\delta}^2 &\leqslant \|\delta\|_\infty \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} K_h(y-x) b(y) f(y) dy \right)^2 dx \\ &\leqslant \|\delta\|_\infty \int_{-\infty}^{\infty} b(y)^2 f(y) \int_{-\infty}^{\infty} K_h(y-x)^2 dx dy \leqslant \frac{\|\delta\|_\infty \|K\|_2^2 \|b^2 f\|_1}{h} \end{aligned} \tag{3.19}$$

and

$$\begin{aligned} \mathbb{E}(V_h(N)) &= \mathbb{E}(\|\Phi_h(X,.) - (bf)_h\|_{2,\delta}^2) \\ &= \mathbb{E}(\|\Phi_h(X,.)\|_{2,\delta}^2) + \|(bf)_h\|_{2,\delta}^2 - 2\int_{-\infty}^{\infty} (bf)_h(x) \mathbb{E}(\Phi_h(X,x)) \delta(x) dx \\ &= \mathbb{E}(\|\Phi_h(X,.)\|_{2,\delta}^2) - \|(bf)_h\|_{2,\delta}^2. \end{aligned} \tag{3.20}$$

Consider

$$v_h(N) := V_h(N) - \mathbb{E}(V_h(N)) = \frac{1}{N} \sum_{i=1}^{N} (g_h(X^i) - \mathbb{E}(g_h(X^i)))$$

with

$$g_h(\varphi) := \|\Phi_h(\varphi,.) - (bf)_h\|_{2,\delta}^2 \; ; \forall \varphi \in E.$$

By Bernstein's inequality, for any $\lambda > 0$, with probability larger than $1 - 2e^{-\lambda}$,

$$|v_h(N)| \leqslant \sqrt{\frac{2\mathfrak{v}_h \lambda}{N}} + \frac{\mathfrak{c}_h \lambda}{N}$$

where

$$\mathfrak{c}_h = \frac{\|g_h\|_\infty}{3} \quad \text{and} \quad \mathfrak{v}_h = \mathbb{E}(g_h(X)^2).$$

Moreover, by Inequality (3.19) and Lemma 3.6.4,

$$\begin{aligned} \mathfrak{c}_h &= \frac{1}{3} \sup_{\varphi \in E} \|\Phi_h(\varphi,.) - (bf)_h\|_{2,\delta}^2 \leqslant \frac{2}{3} \left(\sup_{\varphi \in E} \|\Phi_h(\varphi,.)\|_{2,\delta}^2 + \|(bf)_h\|_{2,\delta}^2 \right) \\ &\leqslant \frac{\mathfrak{c}_1}{h^3} \quad \text{with} \quad \mathfrak{c}_1 = \frac{2}{3} \left[\frac{6\|\mathcal{K}\|_\infty^2}{(T-t_0)^2} + \|\delta\|_\infty \|\sigma\|_\infty^4 \|K'\|_2^2 + \|\delta\|_\infty \|K\|_2^2 \|b^2 f\|_1 \right] \end{aligned}$$

and, by Inequality (3.19), Equality (3.20) and Lemma 3.6.4,

$$\mathfrak{v}_h \leqslant \|g_h\|_\infty \mathbb{E}(V_h(N)) \leqslant \frac{3\mathfrak{c}_1}{h^3} (\mathbb{E}(\|\Phi_h(X,.)\|_{2,\delta}^2) - \|(bf)_h\|_{2,\delta}^2)$$

$$\leqslant \frac{\mathfrak{c}_2\mathfrak{m}(h)}{h^3} \quad \text{with} \quad \mathfrak{c}_2 = 3\mathfrak{c}_1.$$

Then, for any $\theta \in (0,1)$, since $(Nh_0^3)^{-1} \leqslant 1$, with probability larger than $1 - 2e^{-\lambda}$,

$$\begin{aligned} |v_h(N)| &\leqslant 2\sqrt{\frac{\mathfrak{c}_2\mathfrak{m}(h)\lambda}{Nh^3}} + \frac{\mathfrak{c}_1\lambda}{Nh^3} \\ &\leqslant \theta\mathfrak{m}(h) + \frac{(\mathfrak{c}_1+\mathfrak{c}_2)\lambda}{\theta N h^3} \leqslant \theta\mathfrak{m}(h) + \frac{(\mathfrak{c}_1+\mathfrak{c}_2)\lambda}{\theta}. \end{aligned}$$

So, with probability larger than $1 - 2|\mathcal{H}_N|e^{-\lambda}$,

$$\sup_{h\in\mathcal{H}_N}\left\{\frac{|v_h(N)|}{N} - \frac{\theta\mathfrak{m}(h)}{N}\right\} \leqslant \frac{(\mathfrak{c}_1+\mathfrak{c}_2)\lambda}{\theta N}.$$

Therefore, by Equality (3.20), with probability larger that $1 - 2|\mathcal{H}_N|e^{-\lambda}$,

$$\begin{aligned} &\sup_{h\in\mathcal{H}_N}\left\{\frac{1}{N}|V_h(N) - \mathbb{E}(\|\Phi_h(X,.)\|_{2,\delta}^2)| - \frac{\theta\mathfrak{m}(h)}{N}\right\} \\ &\qquad \leqslant \sup_{h\in\mathcal{H}_N}\left\{\frac{|v_h(N)|}{N} - \frac{\theta\mathfrak{m}(h)}{N}\right\} + \frac{1}{N}\|K_h * (bf)\|_{2,\delta}^2 \leqslant \frac{(\mathfrak{c}_1+\mathfrak{c}_2+\|\delta\|_\infty\|K\|_1^2\|bf\|_2^2)(1+\lambda)}{\theta N}. \end{aligned}$$

3.6.7.5 Proof of Lemma 3.6.7

For any $h, h' \in \mathcal{H}_N$,

$$W_{h,h'}(N) = \frac{1}{N}\sum_{i=1}^N (g_{h,h'}(X^i) - \mathbb{E}(g_{h,h'}(X^i)))$$

with, for every $\varphi \in E$,

$$g_{h,h'}(\varphi) := \langle \Phi_h(\varphi,.), (bf)_{h'} - bf\rangle_{2,\delta}.$$

By Bernstein's inequality, for any $\lambda > 0$, with probability larger than $1 - 2e^{-\lambda}$,

$$|W_{h,h'}(N)| \leqslant \sqrt{\frac{2\mathfrak{v}_{h,h'}\lambda}{N}} + \frac{\mathfrak{c}_{h,h'}\lambda}{N}$$

where

$$\mathfrak{c}_{h,h'} = \frac{\|g_{h,h'}\|_\infty}{3} \quad \text{and} \quad \mathfrak{v}_{h,h'} = \mathbb{E}(g_{h,h'}(X)^2).$$

Moreover, by Lemma 3.6.4,

$$\begin{aligned} \mathfrak{c}_{h,h'} &= \frac{1}{3}\sup_{\varphi\in E}|\langle \Phi_h(\varphi,.), (bf)_{h'} - bf\rangle_{2,\delta}| \leqslant \frac{1}{3}\|(bf)_{h'} - bf\|_{2,\delta}\sup_{\varphi\in E}\|\Phi_h(\varphi,.)\|_{2,\delta} \\ &\leqslant \frac{\mathfrak{c}_1}{h^{3/2}}\|(bf)_{h'} - bf\|_{2,\delta} \quad \text{with} \quad \mathfrak{c}_1 = \frac{1}{3}\left[\frac{6\|\mathcal{K}\|_\infty^2}{(T-t_0)^2} + \|\delta\|_\infty\|\sigma\|_\infty^4\|K'\|_2^2\right]^{1/2} \end{aligned}$$

and

$$\mathfrak{v}_{h,h'} \leqslant \mathbb{E}(\langle \Phi_h(X,.), (bf)_{h'} - bf\rangle_{2,\delta}^2) \leqslant \mathfrak{c}_{3.6.4,2}\|(bf)_{h'} - bf\|_{2,\delta}^2.$$

Then, for any $\theta \in (0,1)$, with probability larger than $1 - 2e^{-\lambda}$,

$$|W_{h,h'}(N)| \leqslant 2\sqrt{\frac{\mathfrak{c}_{3.6.4,2}\lambda}{N}\|(bf)_{h'} - bf\|_{2,\delta}^2} + \frac{\mathfrak{c}_1\lambda}{Nh^{3/2}}\|(bf)_{h'} - bf\|_{2,\delta}$$

$$\leqslant \quad \theta\|(bf)_{h'} - bf\|_{2,\delta}^2 + \frac{2\mathfrak{c}_{3.6.4,2}\lambda}{\theta N} + \frac{2\mathfrak{c}_1^2\lambda^2}{\theta N^2 h^3} \leqslant \theta\|(bf)_{h'} - bf\|_{2,\delta}^2 + \frac{2(\mathfrak{c}_{3.6.4,2} + \mathfrak{c}_1^2)(1+\lambda)^2}{\theta N}.$$

So, with probability larger than $1 - 2|\mathcal{H}_N|e^{-\lambda}$,

$$\begin{aligned}
\sup_{h\in\mathcal{H}_N}\{|W_{h,h_0}(N)| - \theta\|(bf)_{h_0} - bf\|_{2,\delta}^2\} &\leqslant \frac{\mathfrak{c}_{3.6.7}(1+\lambda)^2}{\theta N},\\
\sup_{h\in\mathcal{H}_N}\{|W_{h_0,h}(N)| - \theta\|(bf)_{h} - bf\|_{2,\delta}^2\} &\leqslant \frac{\mathfrak{c}_{3.6.7}(1+\lambda)^2}{\theta N} \text{ and}\\
\sup_{h\in\mathcal{H}_N}\{|W_{h,h}(N)| - \theta\|(bf)_{h} - bf\|_{2,\delta}^2\} &\leqslant \frac{\mathfrak{c}_{3.6.7}(1+\lambda)^2}{\theta N}.
\end{aligned}$$

3.6.8 Proof of Corollary 3.4.3

On the one hand, as in the proof of Proposition 3.2.5 and since $\delta(x) > m$ for every $x \in [A, B]$,

$$\begin{aligned}
\mathbb{E}(\|\widehat{b}_{N,\widehat{h},\widehat{h}'} - b\|_{f,A,B}^2) &\leqslant \frac{\mathfrak{c}_{3.2.5}}{m^2}[\mathbb{E}(\|\widehat{bf}_{N,\widehat{h}} - bf\|_{2,A,B}^2) + 2\mathbb{E}(\|\widehat{f}_{N,\widehat{h}'} - f\|_2^2)]\\
&\leqslant \frac{2\mathfrak{c}_{3.2.5}}{m^3}[\mathbb{E}(\|\widehat{bf}_{N,\widehat{h}} - bf\|_{2,\delta}^2) + \mathbb{E}(\|\widehat{f}_{N,\widehat{h}'} - f\|_2^2)].
\end{aligned}$$

On the other hand, by Theorem 3.4.2 and *union bounds*,

$$\begin{aligned}
\mathbb{E}(\|\widehat{bf}_{N,\widehat{h}} - bf\|_{2,\delta}^2) &\leqslant (1+\vartheta)\min_{h\in\mathcal{H}_N}\mathbb{E}(\|\widehat{bf}_{N,h} - bf\|_{2,\delta}^2) + \frac{\mathfrak{c}_{3.4.2,2}}{\vartheta}\left(\|(bf)_{h_0} - bf\|_{2,\delta}^2 + \frac{1}{N}\right)\\
&\leqslant (1 \vee \|\delta\|_\infty)\left[(1+\vartheta)\min_{h\in\mathcal{H}_N}\mathbb{E}(\|\widehat{bf}_{N,h} - bf\|_2^2) + \frac{\mathfrak{c}_{3.4.2,2}}{\vartheta}\left(\|(bf)_{h_0} - bf\|_2^2 + \frac{1}{N}\right)\right]
\end{aligned}$$

and

$$\mathbb{E}(\|\widehat{f}_{N,\widehat{h}'} - f\|_2^2) \leqslant (1+\vartheta)\min_{h'\in\mathcal{H}_N}\mathbb{E}(\|\widehat{f}_{N,h'} - f\|_2^2) + \frac{\bar{\mathfrak{c}}_{3.4.2,2}}{\vartheta}\left(\|f_{h_0'} - f\|_2^2 + \frac{1}{N}\right).$$

Therefore,

$$\begin{aligned}
\mathbb{E}(\|\widehat{b}_{N,\widehat{h},\widehat{h}'} - b\|_{f,A,B}^2) \leqslant \frac{2\mathfrak{c}_{3.2.5}(1 \vee \|\delta\|_\infty)}{m^3}\Bigg[&(1+\vartheta)\min_{(h,h')\in\mathcal{H}_N\times\mathcal{H}_N'}\{\mathbb{E}(\|\widehat{bf}_{N,h} - bf\|_2^2) + \mathbb{E}(\|\widehat{f}_{N,h'} - f\|_2^2)\}\\
&+\frac{\mathfrak{c}_{3.4.3}}{\vartheta}\left(\|(bf)_{h_0} - bf\|_2^2 + \|f_{h_0'} - f\|_2^2 + \frac{1}{N}\right)\Bigg].
\end{aligned}$$

Bibliographie

[1] D. Lamberton, B. Lapeyre. *Introduction au calcul stochastique appliqué à la finance.* Ellipses (2012).

[2] K. Øksendal, A. Sulem. *Stochastic differential equations : Theory and application.* Physics Reports (2000).

[3] M. Musiela, M. Rutkowski. *Stochastic differential equations in finance.* Springer (2005).

[4] J. Paulsson. *Stochastic models of gene expression.* Current Opinion in Genetics and Development (2005).

[5] R. H. Shumway, D. S. Stoffer. *Time series analysis and its applications : With R examples.* Springer (2016).

[6] P. J. Brockwell, R. A. Davis. *Introduction to time series and forecasting.* Springer (2016).

[7] H. Lütkepohl. *New introduction to multiple time series analysis.* Springer (2005).

[8] Estrella, Mishkin. *A new approach to forecasting recessions using the yield curve.* The Review of Economics and Statistics (1998).

[9] Shaman and al. *Real-time influenza forecast during the 2012-2013 season in the United States.* Nature Communications (2013).

[10] K. Fokianos, D. Tjøstheim. *Nonlinear poisson autoregression.* Annals of the Institute of Statistical Mathematics, 64(6) : 1205-1225 (2012).

[11] C. Francq, J. M. Zakoian. *GARCH models : structure, statistical inference and financial applications.* Wiley (2019).

[12] C. Gourieroux, A. Monfort. *Time Series and Dynamic Models.* Cambridge University Press (1997).

[13] J. Bennett, S. Lanning. *The Netflix Prize.* Proceedings of KDD Cup and Workshop, page 35 (2007).

[14] T.T. Cai, A. Zhang. *Rop : Matrix recovery via rank-one projections.* The Annals of Statistics 43(1), 102-138 (2015).

[15] Klopp, O., Lounici, K. and Tsybakov, A. B. *Robust Matrix Completion.* Probability Theory and Related Fields 169, 1-2, 523-564, 2017.

[16] O. Klopp, Y. Lu, A.B. Tsybakov, H.H. Zhou. *Structured matrix estimation and completion.* Bernoulli 25(4B), 3883-3911 (2019).

[17] Koltchinskii, V., Lounici, K. and Tsybakov, A. B. *Nuclear-Norm Penalization and Optimal Rates for Noisy Low-Rank Matrix Completion.* The Annals of Statistics 39, 5, 2302-2329, 2011.

[18] K.i. Moridomi, K. Hatano, E. Takimoto. *Tighter generalization bounds for matrix completion via factorization into constrained matrices.* IEICE TRANSACTIONS on Information and Systems 101(8), 1997-2004 (2018)

[19] T. T. Mai, P. Alquier. *A Bayesian Approach for Noisy Matrix Completion : Optimal Rate Under General Sampling Distribution.* Electronic Journal of Statistics 9, 1, 823-841, (2015).

[20] Mai, T. T. *Bayesian Matrix Completion with a Spectral Scaled Student Prior : Theoretical Guarantee and Efficient Sampling.* ArXiv preprint arxiv :2104.08191.

[21] Mai, T. T. *Numerical Comparisons Between Bayesian and Frequentist Low-Rank Matrix Completion : Estimation Accuracy and Uncertainty Quantification.* ArXiv preprint arxiv :2103.11749.

[22] Y. Koren, R. Bell, C. Volinsky. *Matrix factorization techniques for recommender systems.* Computer, 42(8) :30-37 (2009).

[23] P. Alquier, X. Li, O. Wintenberger. *Prediction of Time Series by Statistical Learning : General Losses and Fast Rates.* Dependence Modeling 1, 65-93 (2013).

[24] P. Alquier, N. Marie. *Matrix factorization for multivariate time series analysis.* Electronic journal of statistics 13(2), 4346-4366 (2019).

[25] C. Vernade, O. Capp´e. *Learning from missing data using selection bias in movie recommendation.* 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pages 1-9. IEEE (2015).

[26] F. Husson, J. Josse, B. Narasimhan, G. Robin. *Imputation of mixed data with multilevel singular value decomposition.* arXiv : 1804.11087 (2018).

[27] J. M. Stuart, E. Segal, D. Koller, S. K. Kim. *A gene-coexpression network for global discovery of conserved genetic modules.* Science, 302(5643), 249-255 (2003).

[28] N. Vaswani, B. D. Rao. *A matrix factorization approach to multiple subspace tracking.* IEEE Transactions on Signal Processing, 54(3), 891-906 (2006).

[29] D. M. Blei, A. Y. Ng, M. I. Jordan. *Latent Dirichlet allocation.* Journal of Machine Learning Research, 3(Jan), 993-1022 (2003).

[30] H. Akaike. *A new look at the statistical model identification.* IEEE Transactions on Automatic Control, 19(6), 716-723 (1974).

[31] G. Schwarz. *Estimating the dimension of a model.* The Annals of Statistics, 6(2), 461-464 (1978).

[32] A.K. Seghouane, A. Cichocki. *Bayesian estimation of the number of principal components.* Signal Processing 87(3), 562-568 (2007).

[33] E. J. Candes, Y. Plan. *Matrix completion with noise.* Proceedings of the IEEE, 98(6), 925-936 (2010).

[34] C. Ding, T. Li, M. I. Jordan. *Convex and semi-nonnegative matrix factorizations.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(1), 45-55 (2006).

[35] C. Zhang, L. Chen, J. Bu. *A new metric for evaluating top-N recommendations in recommendation systems.* Proceedings of the 22nd ACM international conference on Conference on information and knowledge management, 2199-2202 (2013).

[36] C.T. dos S. Dias, W.J. Krzanowski. *Model selection and cross validation in additive main effect and multiplicative interaction models.* Crop Science 43(3), 865-873 (2003).

[37] M. Lavielle. *Using penalized contrasts for the change-point problem.* Signal processing 85(8), 1501-1510 (2005).

[38] S. Arlot, P. Massart. *Data-driven calibration of penalties for least-squares regression.* Journal of Machine learning research 10(2) (2009).

[39] S. Arlot. *Minimal penalties and the slope heuristics : a survey.* Journal de la société française de statistique 160(3), 1-106 (2019).

[40] L. Birgé, P. Massart. *Gaussian model selection.* Journal of the European Mathematical Society 3(3), 203-268 (2001).

[41] J.P. Baudry, C. Maugis, B.Michel. *Slope heuristics : overview and implementation.* Statistics and Computing, 22(2) :455-470 (2012).

[42] E. Devijver, M. Gallopin, E. Perthame. *Nonlinear network-based quantitative trait prediction from transcriptomic data.* arXiv preprint : arXiv :1701.07899 (2017).

[43] J. Jacques, C. Preda. *Functional data clustering : a survey.* Advances in Data Analysis and Classification 8(3), 231-255 (2014).

[44] E. Candes, B. Recht. *Matrix completion via regularized least squares.* Proceedings of the 46th Annual Allerton Conference on Communication, Control, and Computing, 222-229, (2009).

[45] D. Liang, J. Altosaar, L. Charlin, D.M. Blei. *Collaborative filtering with recurrent neural networks.* Proceedings of the 26th International Conference on World Wide Web, 411-419, (2016).

[46] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T. S. Chua. *Neural collaborative filtering.* Proceedings of the 26th International Conference on World Wide Web, 173-182 (2017).

[47] P. Doukhan. *Mixing : Properties and Examples.* (Vol. 85). Springer Science and Business Media (1994).

[48] S. Boucheron, G. Lugosi, P. Massart. *Concentration Inequalities.* Oxford University Press (2013).

[49] Samson, P.-M. *Concentration of Measure Inequalities for Markov Chains and ϕ-Mixing Processes.* The Annals of Probability 28, 1, 416-461, 2000.

[50] T. Hastie, R. Mazumder. R Package softImpute, (2013).

[51] D. D. Lee, H. S. Seung. *Learning the parts of objects by non-negative matrix factorization.* Nature, 401(6755), 788-791, (1999).

[52] F. Comte. *Estimation non-paramétrique. 2e édition.* Spartacus IDH, 2017.

[53] F. Comte, N. Marie. *Nonparametric Drift Estimation from Diffusions with Correlated Brownian Motions.* Preprint (arXiv :2210.13173), (2022).

[54] E. A. Nadaraya. *On estimating regression.* Theory of Probability and Its Applications, 9(1), 141-142, (1964).

[55] G. S. Watson. *Smooth regression analysis.* Sankhya : The Indian Journal of Statistics, Series A, 26(4), 359-372, (1964).

[56] A. Goldenshluger, O. Lepski. *Bandwidth Selection in Kernel Density Estimation : Oracle Inequalities and Adaptivity.* Journal of the Royal Statistical Society : Series B (Statistical Methodology), 73(3), 371-396, (2011).

[57] Alquier, P., Bertin, K., Doukhan P. and Garnier, R.. *High Dimensional VAR with Low Rank Transition.* Statistics and Computing 30, 1139-1153, 2020.

[58] Alquier, P. and Ridgway, J. *Concentration of Tempered Posteriors and of their Variational Approximations.* Annals of Statistics, 48, 3, 1475-1497, 2020.

[59] Athey, S., Bayati, M., Doudchenko, N., Imbens, G. and Khosravi, K. *Matrix Completion Methods for Causal Panel Data Models.* (No. w25132). National Bureau of Economic Research, 2018.

[60] Arlot, S. *Minimal Penalties and the Slope Heuristics : a Survey.* Journal de la SFdS 160, 3, 2019.

[61] Bai, J. and Ng, S. *Matrix Completion, Counterfactuals, and Factor Analysis of Missing Data.* ArXiv preprint arXiv :1910.06677.

[62] Basu, S., Li, X. and Michailidis, G. *Low Rank and Structured Modeling of High-Dimensional Vector Autoregressions.* IEEE Transactions on Signal Processing, 67, 5, 1207-1222, 2019.

[63] Bennett, J. and Lanning, S. *The Netflix Prize.* In *Proceedings of KDD Cup and Workshop*, page 35, 2007.

[64] Boucheron, S., Lugosi, G. and Massart, P. *Concentration Inequalities.* Oxford University Press, 2013.

[65] Candès, E.J. and Plan, Y. *Matrix Completion with Noise.* Proceedings of the IEEE, 98, 6, 925-936, 2010.

[66] Candès, E. J. and Plan, Y. *Tight Oracle Inequalities for Low-Rank Matrix Recovery from a Minimal Number of Noisy Random Measurements.* IEEE Trans. Inf. Theory 57, 4, 2342-2359, 2011.

[67] Candès, E.J. and Recht, B. *Exact Matrix Completion via Convex Optimization.* Found. Comput. Math., 9, 6, 717-772, 2009.

[68] Candès, E.J. and Tao, T. *The Power of Convex Relaxation : Near-Optimal Matrix Completion.* IEEE Trans. Inform. Theory, 56, 5, 2053-2080, 2010.

[69] Carel, L. Big data analysis in the field of transportation. Doctoral dissertation, Université Paris-Saclay, 2019.

[70] Carpentier, A., Klopp, O., Löffler, M. and Nickl, R. *Adaptive Confidence Sets for Matrix Completion.* Bernoulli, 24, 4A, 2429-2460, 2018.

[71] Chan, J., Leon-Gonzalez, R. and Strachan, R.W. *Invariant Inference and Efficient Computation in the Static Factor Model.* J. Am. Stat. Assoc. 113, 522, 819-828, 2018.

[72] Cottet, V. and Alquier, P. *1-Bit Matrix Completion : PAC-Bayesian Analysis of a Variational Approximation.* Machine Learning, 107, 3, 579-603, 2018.

[73] Doukhan, P. *Mixing : Properties and Examples (Vol. 85).* Springer Science & Business Media, 1994.

[74] Eshkevari, S. S. and Pakzad, S. N. *Signal Reconstruction from Mobile Sensors Network Using Matrix Completion Approach.* In Topics in Modal Analysis & Testing, Volume 8 (pp. 61-75), Springer, Cham, 2020.

[75] Gillard, J. and Usevich, K. *Structured Low-Rank Matrix Completion for Forecasting in Time Series Analysis.* International Journal of Forecasting 34, 4, 582-597, 2018.

[76] Giordani, P., Pitt, M. and Kohn, R. *Bayesian Inference for Time Series State Space Models.* In : Geweke, J., Koop, G., Van Dijk, H. (eds.) Oxford Handbook of Bayesian Econometrics. Oxford University Press, Oxford, 2011.

[77] Gross, D. *Recovering Low-Rank Matrices from Few Coefficients in any Basis.* IEEE Transactions on Information Theory 57, 3, 1548-1566, 2011.

[78] Hallin, M. and Lippi, M. *Factor Models in High-Dimensional Time Series - A Time-Domain Approach.* Stoch. Process. Appl. 123, 7, 2678-2695, 2013.

[79] Hastie, T., Mazumder, R. and Hastie, M. T. R Package `softImpute`, 2013.

[80] Keshavan, R. H., Montanari, A. and Oh, S. *Matrix Completion from a Few Entries.* IEEE Transactions on Information Theory 56, 6, 2980-2998, 2010.

[81] Keshavan, R. H., Montanari, A. and Oh, S. *Matrix Completion from Noisy Entries.* The Journal of Machine Learning Research 11, 2057-2078, 2010.

[82] Klopp, O. *Noisy Low-Rank Matrix Completion with General Sampling Distribution.* Bernoulli 20, 1, 282-303, 2014.

[83] Koop, G. and Potter, S. *Forecasting in Dynamic Factor Models Using Bayesian Model Averaging.* Econom. J. 7, 2, 550-565, 2004.

[84] Lafond, J., Klopp, O., Moulines, E. and Salmon, J. *Probabilistic Low-Rank Matrix Completion on Finite Alphabets.* Advances in Neural Information Processing Systems 27, 1727-1735, 2014.

[85] Lam, C. and Yao, Q. *Factor Modeling for High-Dimensional Time Series : Inference for The Number of Factors.* Ann. Stat. 40, 2, 694-726, 2012.

[86] Lam, C., Yao, Q. and Bathia, N. *Estimation of Latent Factors for High-Dimensional Time Series.* Biometrika 98, 4, 901-918, 2011.

[87] Lee, D. D. and Seung, H. S. Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788-791, 1999.

[88] Mei, J., De Castro, Y., Goude, Y., Azais, J. M. and Hébrail, G. *Nonnegative Matrix Factorization with Side Information for Time Series Recovery and Prediction.* IEEE Transactions on Knowledge and Data Engineering 31, 3, 493-506, 2018.

[89] Mei, J., De Castro, Y., Goude, Y., and Hébrail, G. *Nonnegative Matrix Factorization for Time Series Recovery from a Few Temporal Aggregates.* Proceedings of the 34th International Conference on Machine Learning, PMLR 70 :2382-2390, 2017.

[90] Massart, P. *Concentration Inequalities and Model Selection.* Volume 1896 of *Lecture Notes in Mathematics*, Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, Edited by Jean Picard.

[91] Negahban, S. and Wainwright, M. J. *Restricted Strong Convexity and Weighted Matrix Completion : Optimal Bounds with Noise.* The Journal of Machine Learning Research 13, 1, 1665-1697, 2012.

[92] Poulos, J. *State-Building through Public Land Disposal ? An Application of Matrix Completion for Counterfactual Prediction.* ArXiv preprint arXiv :1903.08028.

[93] Shi, W., Zhu, Y., Philip, S. Y., Huang, T., Wang, C., Mao, Y. and Chen, Y. *Temporal Dynamic Matrix Factorization for Missing Data Prediction in Large Scale Coevolving Time Series.* IEEE Access 4, 6719-6732, 2016.

[94] Suzuki, T. *Convergence Rate of Bayesian Tensor Estimator and its Minimax Optimality.* The 32nd International Conference on Machine Learning (ICML2015), JMLR Workshop and Conference Proceedings 37, 1273-1282, 2015.

[95] Tonnelier, E., Baskiotis, N., Guigue, V. and Gallinari, P. Anomaly detection in smart card logs and distant evaluation with Twitter : a robust framework. Neurocomputing, 298, 109-121, 2018.

[96] Tsagkatakis, G., Beferull-Lozano, B. and Tsakalides, P. *Singular Spectrum-Based Matrix Completion for Time Series Recovery and Prediction.* EURASIP Journal on Advances in Signal Processing 1, 66, 2016.

[97] Tsybakov, A. *Introduction to Nonparametric Estimation.* Springer, 2009.

[98] Xie, K., Ning, X., Wang, X., Xie, D., Cao, J., Xie, G. and Wen, J. *Recover Corrupted Data in Sensor Networks : A Matrix Completion Solution.* IEEE Transactions on Mobile Computing 16, 5, 1434-1448, 2016.

[99] Yu, H. F., Rao, N. and Dhillon, I. S. *Temporal Regularized Matrix Factorization for High-Dimensional Time Series Prediction.* Advances in Neural Information Processing Systems 29, 847-855, 2016.

[100] A. Cohen, M.A. Leviatan and D. Leviatan. *On the Stability and Accuracy of Least Squares Approximations.* Found. Comput. Math. 13, 819-834, 2013.

[101] F. Comte and V. Genon-Catalot. *Regression Function Estimation on Non Compact Support as a Partly Inverse Problem.* The Annals of the Institute of Statistical Mathematics 72, 4, 1023-1054, 2020.

[102] F. Comte and V. Genon-Catalot. *Nonparametric Drift Estimation for i.i.d. Paths of Stochastic Differential Equations.* The Annals of Statistics 48, 6, 3336-3365, 2020.

[103] F. Comte and V. Genon-Catalot. *Drift Estimation on Non Compact Support for Diffusion Models.* Stoch. Proc. Appl. 134, 174-207, 2021.

[104] F. Comte, V. Genon-Catalot and Y. Rozenholc. *Penalized Nonparametric Mean Square Estimation of the Coefficients of Diffusion Processes.* Bernoulli 12, 2, 514-543, 2007.

[105] F. Comte, V. Genon-Catalot and A. Samson. *Nonparametric Estimation for Stochastic Differential Equations with Random Effects.* Stoch. Proc. Appl. 123, 2522-2551, 2013.

[106] F. Comte and N. Marie. *On a Nadaraya-Watson Estimator with Two Bandwidths.* Electronic Journal of Statistics 15, 1, 2566-2607, 2021.

[107] A. Dalalyan. *Sharp Adaptive Estimation of the Trend Coefficient for Ergodic Diffusion.* The Annals of Statistics 33, 6, 2507-2528, 2005.

[108] M. Delattre, V. Genon-Catalot and C. Larédo. *Parametric Inference for Discrete Observations of Diffusion Processes with Mixed Effects.* Stoch. Proc. Appl. 128, 1929-1957, 2018.

[109] M. Delattre, V. Genon-Catalot and A. Samson. *Maximum Likelihood Estimation for Stochastic Differential Equations with Random Effects.* Scand. J. Stat. 40, 322-343, 2013.

[110] M. Delattre and M. Lavielle. *Coupling the SAEM Algorithm and the Extended Kalman Filter for Maximum Likelihood Estimation in Mixed-Effects Diffusion Models.* Stat. Interface 6, 519-532, 2013.

[111] L. Della Maestra and M. Hoffmann. *Nonparametric Estimation for Interacting Particle Systems : McKean-Vlasov Models.* To appear in Probability Theory and Related Fields, 2021.

[112] C. Denis, C. Dion and M. Martinez. *A Ridge Estimator of the Drift from Discrete Repeated Observations of the Solutions of a Stochastic Differential Equation.* To appear in Bernoulli, 2021.

[113] C. Dion and V. Genon-Catalot. *Bidimensional Random Effect Estimation in Mixed Stochastic Differential Model.* Stat. Inference Stoch. Process. 19, 131-158, 2016.

[114] S. Ditlevsen and A. De Gaetano. *Mixed Effects in Stochastic Differential Equation Models.* REVSTAT 3, 137-153, 2005.

[115] E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models.* Cambridge University Press, 2015.

[116] L. Györfi, M. Kohler, A. Krzyzak and H. Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer Series in Statistics. Springer-Verlag, New York, 2002.

[117] H. Halconruy and N. Marie. *Kernel Selection in Nonparametric Regression.* Mathematical Methods of Statistics 29, 1, 32-56, 2020.

[118] M. Hoffmann. *Adaptive Estimation in Diffusion Processes.* Stoch. Proc. Appl. 79, 135-163, 1999.

[119] C. Houdré and P. Reynaud-Bouret. *Exponential Inequalities, with Constants, for U-statistics of Order Two.* Stochastic Inequalities and Applications, vol. **56** of Progr. Proba, 55-69, Birkhauser, 2003.

[120] M. Kessler. *Simple and Explicit Estimating Functions for a Discretely Observed Diffusion Process.* Scandinavian Journal of Statistics 27, 1, 65-82, 2000.

[121] S. Kusuoka and D. Stroock. *Applications of the Malliavin Calculus, Part II.* J. Fac. Sci. Univ. Tokyo 32, 1-76, 1985.

[122] Y. Kutoyants. *Statistical Inference for Ergodic Diffusion Processes.* Springer, 2004.

[123] C. Lacour, P. Massart and V. Rivoirard. *Estimator Selection : New Method with Applications to Kernel Density Estimation.* Sankhia A 79, 2, 298-335, 2017.

[124] S. Menozzi, A. Pesce and X. Zhang. *Density and Gradient Estimates for Non Degenerate Brownian SDEs with Unbounded Measurable Drift.* Journal of Differential Equations 272, 330-369, 2021.

[125] R. Overgaard, N. Jonsson, C. Tornoe, H. Madsen. *Non-Linear Mixed Effects Models with Stochastic Differential Equations : Implementation of an Estimation Algorithm.* J. Pharmacokinet. Pharmacodyn. 32, 85-107, 2005.

[126] U. Picchini, A. De Gaetano and S. Ditlevsen. *Stochastic Differential Mixed-Effects Models.* Scand. J. Stat. 37, 67-90, 2010.

[127] U. Picchini and S. Ditlevsen. *Practical Estimation of High Dimensional Stochastic Differential Mixed-Effects Models.* Comput. Statist. Data Anal. 55, 1426-1444, 2011.

[128] T. Hastie, R. Mazumder. *SoftImpute : Matrix completion via iterative softthresholded svd.* R package version 1, p1 (2015).

[129] E.J. Candes, C.A. Sing-Long, J.D. Trzasko. *Unbiased risk estimates for singular value thresholding and spectral estimators.* IEEE transactions on signal processing 61(19), 4643-4657 (2013).

[130] M.O. Ulfarsson, V. Solo. *Dimension estimation in noisy pca with sure and random matrix theory.* IEEE transactions on signal processing 56(12), 5804-5816 (2008).

[131] H.F. Lopes, M. West. *Bayesian model assessment in factor analysis.* Statist. Sinica 14(1), 41-68 (2004).

[132] X. Collilieux, E. Lebarbier, S. Robin. *A factor model approach for the joint segmentation with between-series correlation.* Scandinavian Journal of Statistics 46(3), 686-705 (2019).

[133] P. Alquier, N. Marie, A. Rosier. *Tight risk bound for high dimensional time series completion.* Electronic Journal of Statistics 16(1), 3001-3035 (2022).

[134] M.O. Ulfarsson, V. Solo. *Dimension estimation in noisy pca with sure and random matrix theory.* IEEE transactions on signal processing 56(12), 5804-5816 (2008).

[135] C.T. dos S. Dias, W.J. Krzanowski. *Model selection and cross validation in additive main effect and multiplicative interaction models.* Crop Science, 43(3), 865-873 (2003).