Université Lille III - Charles-de-Gaulle

École Doctorale SESAM

Thèse de doctorat

Discipline : Mathématiques Appliquées

présentée par

Camille Ternynck

Contributions à la modélisation de données spatiales et fonctionnelles. Applications.

---

dirigée par Sophie Dabo-Niang et Anne-Françoise Yao,
co-dirigée par Fateh Chebana

Soutenue le Vendredi 28 Novembre 2014 devant le jury composé de :

M^me^ Laurence Broze	Université de Lille	Examinatrice
M^r^ Fateh Chebana	INRS (Québec)	Co-directeur de thèse
M^me^ Sophie Dabo-Niang	Université de Lille	Directrice de thèse
M^r^ Mohamed El Machkouri	Université de Rouen	Examinateur
M^r^ Frédéric Ferraty	Université de Toulouse	Rapporteur
M^me^ Raquel Menezes	Minho University (Portugal)	Rapportrice
M^r^ Cristian Preda	Université de Lille	Examinateur
M^me^ Anne-Françoise Yao	Université de Clermont-Ferrand	Directrice de thèse

Résumé français

Dans ce mémoire de thèse, nous nous intéressons à la modélisation non paramétrique de données spatiales et/ou fonctionnelles, plus particulièrement basée sur la méthode à noyau. En général, les échantillons que nous avons considérés pour établir les propriétés asymptotiques des estimateurs proposés sont constitués de variables dépendantes. La spécificité des méthodes étudiées réside dans le fait que les estimateurs prennent en compte la structure de dépendance des données considérées.

Dans une première partie, nous appréhendons l'étude de variables réelles spatialement dépendantes. Nous proposons une nouvelle approche à noyau pour estimer les fonctions de densité de probabilité et de régression spatiales ainsi que le mode. La particularité de cette approche est qu'elle permet de tenir compte à la fois de la proximité entre les observations et de celle entre les sites. Nous étudions les comportements asymptotiques des estimateurs proposés ainsi que leurs applications à des données simulées et réelles.

Dans une seconde partie, nous nous intéressons à la modélisation de données à valeurs dans un espace de dimension infinie ou dites "données fonctionnelles". Dans un premier temps, nous adaptons le modèle de régression non paramétrique introduit en première partie au cadre de données fonctionnelles spatialement dépendantes. Nous donnons des résultats asymptotiques ainsi que numériques. Puis, dans un second temps, nous étudions un modèle de régression de séries temporelles dont les variables explicatives sont fonctionnelles et le processus des innovations est autorégressif. Nous proposons une procédure permettant de tenir compte de l'information contenue dans le processus des erreurs. Après avoir étudié le comportement asymptotique de l'estimateur à noyau proposé, nous analysons ses performances sur des données simulées puis réelles.

La troisième partie est consacrée aux applications. Tout d'abord, nous présentons des résultats de classification non supervisée de données spatiales (multivariées), simulées et réelles. La méthode de classification considérée est basée sur l'estimation du mode spatial, obtenu à partir de l'estimateur de la fonction de densité spatiale introduit dans le cadre de la première partie de cette thèse. Puis, nous appliquons cette méthode de classification basée sur le mode ainsi que d'autres méthodes de classification non supervisée de la littérature sur des données hydrologiques de nature fonctionnelle. Enfin, cette classification des données hydrologiques nous a amené à appliquer des outils de détection de rupture sur ces données fonctionnelles.

Mots-clefs

Estimation non paramétrique, Estimateur à noyau, Densité de probabilité, Mode, Régression, Statistique spatiale, Données fonctionnelles, Séries temporelles, α-mélange, Classification non supervisée, Détection de rupture

English abstract

Contributions to modeling spatial and functional data. Applications.

In this dissertation, we are interested in nonparametric modeling of spatial and/or functional data, more specifically based on kernel method. Generally, the samples we have considered for establishing asymptotic properties of the proposed estimators are constituted of dependent variables. The specificity of the studied methods lies in the fact that the estimators take into account the structure of the dependence of the considered data.

In a first part, we study real variables spatially dependent. We propose a new kernel approach to estimating spatial probability density of the mode and regression functions. The distinctive feature of this approach is that it allows taking into account both the proximity between observations and that between sites. We study the asymptotic behaviors of the proposed estimates as well as their applications to simulated and real data.

In a second part, we are interested in modeling data valued in a space of infinite dimension or so-called "functional data". As a first step, we adapt the nonparametric regression model, introduced in the first part, to spatially functional dependent data framework. We get convergence results as well as numerical results. Then, later, we study time series regression model in which explanatory variables are functional and the innovation process is autoregressive. We propose a procedure which allows us to take into account information contained in the error process. After showing asymptotic behavior of the proposed kernel estimate, we study its performance on simulated and real data.

The third part is devoted to applications. First of all, we present unsupervised classification results of simulated and real spatial data (multivariate). The considered classification method is based on the estimation of spatial mode, obtained from the spatial density function introduced in the first part of this thesis. Then, we apply this classification method based on the mode as well as other unsupervised classification methods of the literature on hydrological data of functional nature. Lastly, this classification of hydrological data has led us to apply change point detection tools on these functional data.

Keywords

Nonparametric estimation, Kernel estimate, Probability density, Mode, Regression, Spatial statistics, Functional data, Time series, α-mixing, Unsupervised classification, Change point detection

Table des matières

Victor, je te dédie ce mémoire de thèse.

Remerciements

Je tiens tout d'abord à exprimer toute ma reconnaissance à Sophie Dabo-Niang, ma directrice de thèse, qui m'a tellement appris, et ce, depuis ma dernière année de licence, année durant laquelle j'ai réalisé mon premier travail d'étude et de recherche. Toutes ces années, elle a su être à l'écoute et très disponible, malgré ses diverses responsabilités et son emploi du temps chargé. Je la remercie également de m'avoir accordé sa confiance en acceptant d'encadrer mes années de thèse mais aussi de m'avoir donné la chance de vivre toutes ces expériences enrichissantes. "Sophie, tu m'as transmis ta passion pour les statistiques et la recherche et je ne peux que t'en remercier !"

Je remercie vivement Anne-Françoise Yao et Fateh Chebana d'avoir co-encadré ma thèse. Chacun d'eux a largement contribué à mon enthousiasme et épanouissement dans la recherche. Merci à Anne-Françoise pour sa bonne humeur, ses nombreux conseils et pour son accueil lors de ma visite à Clermont-Ferrand. Merci à Fateh pour sa gentillesse, ses remarques très constructives et aussi de m'avoir permis de travailler à l'INRS. C'est grâce à lui que j'ai découvert le monde des "Applications".

Je remercie vivement Raquel Menezes et Frédéric Ferraty d'avoir accepté de rapporter ce travail de thèse. Je suis très honorée qu'ils aient pris le temps de s'intéresser à mon travail. Je souhaite aussi témoigner ma reconnaissance à Laurence Broze, Mohammed El Machkouri et Cristian Preda d'être présents dans mon jury de soutenance.

Bien sûr, je n'oublie pas les personnes qui m'ont permis de m'orienter vers la recherche. Baba, Laurence, Sophie et Thomas, je me souviendrai toujours de notre discussion lors de la porte ouverte de l'université, en Janvier 2010. C'est à cette date que vous m'avez gentiment proposé d'ouvrir un groupe de travail en Statistiques Spatiales afin de me faire découvrir le monde de la recherche. Je remercie également l'ensemble des enseignants de la licence et du master MIASHS pour la qualité de l'enseignement que j'ai reçu mais aussi pour leur disponibilité lorsque j'avais des questions. Une petite pensée aussi à tous les membres du laboratoire EQUIPPE : certains m'ont aidée lorsque j'avais des problèmes avec *Latex* ou *MacBook*, d'autres m'ont donné des conseils lorsque je me posais des questions sur mes choix d'orientation, d'autres encore ont préparé avec moi certains travaux dirigés, et d'autres sont tout simplement devenus des ami(e)s... Merci également à mes collègues de bureau, de Lille 3 comme de l'INRS, avec qui nous avons pu discuter *Mathématiques* mais aussi de nos différentes cultures et de nos petites histoires du quotidien. Il me faut également remercier l'ensemble des personnes présentes à l'UFR MIME pour leur gentillesse et le travail fourni pour permettre le bon déroulement de chaque année universitaire et l'organisation des missions.

Je remercie chaleureusement Laurence Broze qui me soutient depuis toutes ces années. Mais surtout je la remercie, ainsi que tous les membres de l'association *femmes & mathématiques*, pour toute l'énergie déployée dans les nombreuses actions menées dans le but, entre autres, de permettre aux jeunes filles de dépasser les préjugés et de s'orienter vers

des carrières mathématiques. Je suis très heureuse d'avoir pu participer à leurs côtés à certaines de ces actions.

Mes débuts dans la recherche ont également été synonymes de nombreuses expériences et collaborations. Je remercie tout d'abord Sandrine Vaz et Yves Vérin de m'avoir proposé un stage à l'Ifremer. Grâce à eux, j'ai appris que la recherche en statistiques est nécessaire pour des choses très concrètes. J'ai également eu la chance, durant ce stage, d'embarquer quinze jours sur un navire scientifique afin de participer à la collecte des données : ce fût une expérience très enrichissante et inoubliable. Merci aussi à Leila Hamdad avec qui j'ai partagé le stress des révisions mais aussi la joie de voir notre article publié ! J'ai également eu la chance de travailler sur le projet de collaboration Franco-Québécoise entre le laboratoire EQUIPPE et l'INRS et ainsi de travailler avec Fateh, Mohamed, Pierre, Sophie et Taha. Merci à vous pour cette belle collaboration en espérant qu'elle continuera encore longtemps. Cela m'a aussi permis de découvrir cette très belle ville de Québec. Bien sûr, je suis sincèrement reconnaissante envers Taha de m'accueillir dans son équipe à Masdar Institute (Abu Dhabi) pour ma première année de post-doctorat. Enfin, je remercie Serge Guillas d'avoir accepté de travailler avec moi, de m'avoir accueillie à l'University College of London. Merci aussi pour ses précieux conseils et surtout pour sa sympathie.

Je souhaiterais également remercier les réviseurs anonymes des articles publiés ou en révision ainsi que les personnes présentes lors de mes exposés. Grâce à leurs remarques et suggestions, j'ai pu me poser de nouvelles questions et ainsi améliorer mon travail que je peux présenter dans ce manuscrit.

Merci aussi à Aurélie Fischer et Benjamin Auder d'avoir partagé les codes R qu'ils ont développés et que j'ai utilisés dans les applications présentées au Chapitre 7.

J'en profite également pour remercier toute ma famille et mes ami(e)s pour ce que nous partageons ensemble : discussions, rires, sorties, expériences culinaires, danse, repas, et bien d'autres. Merci à ma presque cousine, Mathilde, et à Claire de m'avoir gentiment aidée à relire certains fichiers. Merci aussi à Guillaume pour sa patience, son soutien et ses encouragements lors de mes moments de doute.

Pour terminer, je remercie très sincèrement mes parents et mon petit frère qui m'ont toujours encouragée et soutenue. Je n'aurais jamais pu réaliser tout ce parcours sans leur présence ! Et surtout, "Merci Maman d'avoir tenu tête lorsque je ne voulais pas aller à l'école durant mes premières années de maternelle !"

Lexique

Notations (Français)

$\mathbb{N}$	ensemble des entiers naturels : 0, 1, 2, ...
$\mathbb{N}^*$	ensemble des entiers naturels non nuls : 1, 2, ...
$\mathbb{Z}$	ensemble des entiers relatifs : ..., -2, -1, 0, 1, 2, ...
$\mathbb{Q}$	ensemble des nombres rationnels : $\frac{m}{n}$
$\mathbb{R}$	ensemble des nombres réels : $]-\infty, +\infty[$
$\mathbb{R}^d$	espace euclidien réel de dimension d
$\{a_i : i \in I\}$	ensemble des points a_i indexés par un ensemble I
$\overline{A}$ (or A^c)	complémentaire de A
$A \cup B$	union de A et B
$A \cap B$	intersection de A et B
$\mathrm{Card}(A)$	cardinal de A
$\|\cdot\|$	norme

Soit un vecteur $x = (x_1, x_2, ..., x_n)$

x'	transposée de x
$\|x\|_2$	norme euclidienne : $\|x\|_2 = \sqrt{x_1^2 + x_2^2 + ... + x_n^2}$
$\|x\|_1$	norme 1 : $\|x\|_1 = \sqrt{\|x_1\| + \|x_2\| + ... + \|x_n\|}$
$\|x\|_\infty$	norme infinie : $\|x\|_\infty = \lim_{p\to+\infty} \|(x_1, x_2, ..., x_n)\|_p = \max(\|x_1\|, \|x_2\|, ..., \|x_n\|)$
	remarque : $\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1$
$\langle \cdot, \cdot \rangle$	produit scalaire
$\lfloor \cdot \rfloor$	partie entière
■	fin d'une preuve
$X_{\mathbf{i}}$	variable X observée au site **i**
i.i.d.	indépendantes et identiquement distribuées
$\sigma(\cdots)$	tribu engendrée par $(\cdots)$
$(\Omega, \mathcal{A}, \mathbb{P})$	espace de probabilité
	Ω : ensemble non vide
	$\mathcal{A}$: σ-Algèbre de sous-ensemble de Ω
	$\mathbb{P}$: mesure de probabilité sur $\mathcal{A}$

$\mathcal{B}(X,h)$	boule ouverte de centre X et de rayon h
$d(\cdot,\cdot)$	semi-métrique sur un espace fonctionnel E
(E,d)	espace fonctionnel générique et sa semi-métrique
$L^q(E,\mathcal{B},\mu)$	(ou $L^q(E)$, ou $L^q(\mathcal{B})$, ou $L^q(\mu)$)
	espace de classes des fonctions mesurables f telles que
	$\|f\|_q = (\int_E \|f\|^q d\mu)^{1/q} < +\infty \quad (1 \leq p < +\infty)$
	$\|f\|_\infty = \inf\{a : \mu\{f > a\} = 0\} < +\infty \quad (q = +\infty)$
$u_n = O(v_n)$	Il existe une constante c telle que $u_n \leq cv_n$
$u_n = o(v_n)$	$\frac{u_n}{v_n} \to 0$

Notations (English)

$\mathbb{N}$	set of natural numbers : 0, 1, 2, ...
$\mathbb{N}^*$	set of non-zero natural numbers : 1, 2, ...
$\mathbb{Z}$	set of integers : ..., -2, -1, 0, 1, 2, ...
$\mathbb{Q}$	set of rational numbers : $\frac{m}{n}$
$\mathbb{R}$	set of real numbers : $]-\infty, +\infty[$
$\mathbb{R}^d$	Euclidian space of dimension d
$\{a_i : i \in I\}$	set of points a_i indexed by a set I
$\overline{A}$ (or A^c)	complement of A
$A \cup B$	union of A and B
$A \cap B$	intersection of A and B
$\mathrm{Card}(A)$	cardinality of A
$\|\cdot\|$	norm
Let a vector $x = (x_1, x_2, ..., x_n)$	
x'	transpose of x
$\|x\|_2$	euclidean norm : $\|x\|_2 = \sqrt{x_1^2 + x_2^2 + ... + x_n^2}$
$\|x\|_1$	1-norm : $\|x\|_1 = \sqrt{\|x_1\| + \|x_2\| + ... + \|x_n\|}$
$\|x\|_\infty$	infinity norm : $\begin{aligned}\|x\|_\infty &= \lim_{p\to+\infty} \|(x_1, x_2, ..., x_n)\|_p \\ &= \max(\|x_1\|, \|x_2\|, ..., \|x_n\|)\end{aligned}$
	remark : $\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1$
$\langle\cdot,\cdot\rangle$	scalar product
$\lfloor\cdot\rfloor$	integer part
■	end of a proof
$X_{\mathbf{i}}$	variable X observed at site $\mathbf{i}$
i.i.d.	independent and identically distributed
$\sigma(\cdots)$	σ-algebra generated by $(\cdots)$
$(\Omega, \mathcal{A}, \mathbb{P})$	probability space
	Ω : nonempty set
	$\mathcal{A}$: σ-algebra of subset of Ω
	$\mathbb{P}$: probability mesure on $\mathcal{A}$
$\mathcal{B}(X, h)$	open ball of centre X and of radius h
$d(\cdot,\cdot)$	semi-metric on a functional space E
(E, d)	functional space and its semi-metric
$L^q(E, \mathcal{B}, \mu)$	(or $L^q(E)$, or $L^q(\mathcal{B})$, or $L^q(\mu)$)
	space of classes of mesurable functions f such that
	$\|f\|_q = \left(\int_E \|f\|^q d\mu\right)^{1/q} < +\infty \quad (1 \leq q < +\infty)$
	$\|f\|_\infty = \inf\{a : \mu\{f > a\} = 0\} < +\infty \quad (q = +\infty)$
$u_n = O(v_n)$	a constant c exists such that $u_n \leq c v_n$
$u_n = o(v_n)$	$\frac{u_n}{v_n} \to 0$

Introduction générale

Sommaire

Cette thèse a été financée par un contrat doctoral de l'Université Lille 3 - Charles de Gaulle, du 1er Octobre 2011 au 30 Septembre 2014. Elle a été réalisée au sein du laboratoire EQUIPPE (Economie QUantitative Intégration Politiques Publiques Econométrie) de Lille.

Présentation de la thèse

Ce mémoire de doctorat concerne notamment l'inférence statistique dans un cadre non paramétrique, c'est à dire lorsque la loi des observations est inconnue. Il s'agit de modéliser les caractéristiques inconnues d'une population à partir d'un échantillon de celle-ci. L'estimation non paramétrique diffère de l'estimation paramétrique par le fait que la structure du modèle n'est pas spécifiée *a priori* mais est directement déterminée à partir des données. On dit également de ces méthodes qu'elles permettent de "laisser parler les données" (traduction de "letting the data speak for themselves"). Bien qu'en pratique, ces méthodes requièrent des jeux de données de taille importante, elles présentent l'avantage de ne nécessiter que d'hypothèses de régularité (continuité, dérivabilité, . . .) sur la fonction de lien entre les variables. Notons qu'en estimation non paramétrique, on suppose souvent que la fonction de lien appartient à une certaine classe $\mathcal{P}$, choisie parmi les classes non paramétriques de fonctions (Tsybakov (2009)). Par exemple, si on s'intéresse à l'estimation de la fonction de densité, $\mathcal{P}$ peut être l'ensemble de toutes les densités de probabilités continues sur $\mathbb{R}$ ou l'ensemble de toutes les densités de probabilités de Lipschitz continues sur $\mathbb{R}$.

Le terme "non paramétrique" a été utilisé pour la première fois en 1942 par Wolfowitz (1942). Depuis, de nombreux auteurs travaillent sur le sujet comme en témoigne l'abondante littérature. Parmi les multiples problèmes non paramétriques rencontrés dans la littérature, nous nous focaliserons sur l'estimation des fonctions de densité et de régression ainsi que quelques unes de leurs applications. Pour traiter ces deux points, la plupart des outils développés impliquent des techniques d'interpolation, de lissage ou d'approximation. Parmi elles, on retrouve, par exemple, les méthodes d'estimation sur des bases de splines, par projection ou encore les méthodes à noyau (dites de Parzen-Rosenblatt).

Dans ce travail, la méthode à noyau sera principalement utilisée, celle-ci a été introduite par Rosenblatt (1956b) puis généralisée par Parzen (1962).

La fonction de densité de probabilité, que l'on notera f tout au long de ce manuscrit, est un concept fondamental en statistique. Elle décrit la distribution de la variable d'intérêt, notée X, et permet d'obtenir les probabilités associées à cette variable. L'estimation de la densité est la construction d'un estimateur de la fonction de densité à partir des observations. Une première approche, dite paramétrique, repose sur l'hypothèse que les données sont issues d'une famille de distribution connue, par exemple, la distribution normale de moyenne μ et de variance σ^2. L'estimation de f revient alors à estimer les paramètres μ et σ^2 à partir des données puis à substituer ces estimateurs dans la formule de la densité normale. Cependant les données ne permettent pas toujours de déterminer *a priori* la famille de distribution dont elles sont issues. Les méthodes non paramétriques, étudiées au cours de ce travail, ne requièrent pas d'hypothèses *a priori* sur l'appartenance de f à une famille de lois connues. La façon la plus simple d'estimer la fonction de densité f à partir des données est l'estimation par histogramme. Ce dernier est à l'origine de l'estimateur à noyau que nous considérons dans la suite. Plus précisément, on dispose d'un échantillon de n observations $X_1, \ldots, X_n$, indépendantes et identiquement distribuées (i.i.d.) à valeurs réelles. L'estimateur à noyau de la densité f en un point x est donné par

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

où h est un paramètre de lissage appelé "fenêtre" et K est une fonction de poids appelée "noyau", qui satisfait, par exemple, les conditions suivantes :

$$\int_{-\infty}^{\infty} K(x)dx = 1 \qquad \text{et} \qquad \int_{-\infty}^{\infty} K^2(x)dx < \infty.$$

Cette condition implique que K est aussi une fonction de densité de probabilité.

Dans un second temps, on se place dans le cadre où l'on observe n couples $(X_1, Y_1), \ldots, (X_n, Y_n)$, de même loi que le couple de variables aléatoires (X, Y), obéissant au modèle de régression $Y_i = r(X_i) + \epsilon_i$ pour $i = 1, \ldots, n$, et pour lequel la variable ϵ, représente un terme d'erreur. Ainsi, le problème de la régression consiste à rechercher une relation pouvant éventuellement exister entre les variables X_i et Y_i. La variable X constitue la variable dite d'entrée, explicative, de contrôle, endogène ou régresseur et la variable Y est la variable dite de sortie, réponse, exogène ou expliquée. Les variables ϵ_i, représentant un terme d'erreur, sont généralement supposées indépendantes des X_i et centrées. Le plus souvent, la régression des variables aléatoires Y sur X désigne l'espérance conditionnelle de Y sachant X. Mais, le terme de régression désigne en fait tout élément de la distribution conditionnelle de Y sachant X, considéré comme une fonction de X. Par exemple, on peut également s'intéresser à la médiane conditionnelle, au mode conditionnel, aux quantiles conditionnels, etc. Le modèle de régression le plus connu est l'estimateur linéaire des moindres carrés qui consiste à expliquer Y par une combinaison linéaire des composantes de X. En général, le modèle de régression linéaire désigne un modèle dans lequel l'espérance conditionnelle de Y sachant X est une transformation affine de X. On peut également considérer des modèles dans lesquels c'est la médiane conditionnelle ou n'importe quel quantile de la distribution de Y sachant X qui est une transformation affine de X. Lorsque l'on ne peut pas supposer que la fonction de régression appartienne à une famille paramétrique, on a recours à des méthodes d'estimation non paramétriques. L'approche non paramétrique suppose seulement que $r(\cdot)$ appartienne à une classe non paramétrique donnée. Dans ce travail, on s'intéresse notamment au cas où la fonction de lien $r(\cdot)$ repose sur l'espérance

conditionnelle. En effet, on appellera fonction de lien ou de régression, la fonction $r(x)$ qui à toute réalisation x de la variable explicative X associe la quantité $\mathbb{E}[Y|X = x]$. A partir d'un noyau K et d'une fenêtre h, on peut construire des estimateurs non paramétriques de la fonction de régression $r(\cdot)$ similaires aux estimateurs à noyau de la densité. Le plus célèbre, que nous étudierons au cours de ce travail, est l'estimateur de Nadaraya-Watson, introduit par Nadaraya (1964) et Watson (1964), et qui est défini par

$$r_n(x) = \frac{\frac{1}{nh}\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h}\right)}{f_n(x)} \qquad \text{si} \qquad f_n(x) \neq 0$$

et $r_n(x) = \frac{1}{n}\sum_{i=1}^{n} Y_i$ sinon. Cet estimateur peut être vu comme une moyenne pondérée dont les poids dépendent de la distance entre les variables X_i et x. En particulier, ces poids vont permettre de donner plus d'importance aux valeurs Y_i pour lesquelles X_i est "proche" de x. Qu'il s'agisse de l'estimation des fonctions de densité ou de régression, la qualité de l'estimation est très fortement liée à la fenêtre h, dont le choix s'avère être très important et fait l'objet de nombreux travaux (e.g. Hall (1982), Delaigle et Gijbels (2004)). Notons que dans une moindre mesure, les résultats varient également en fonction du noyau K choisi.

Les estimateurs à noyau ont été beaucoup étudiés dans la littérature, d'une part, pour des données indépendantes et réelles (univariées puis multivariées) et d'autre part, en considérant des séries chronologiques (dépendantes et réelles). Plus récemment, l'estimateur de Parzen-Rosenblatt a été adapté à d'autres types de données comme, par exemple, aux données spatiales. En effet, depuis le milieu du $20^{\text{ème}}$ siècle, un nouvel axe de recherche s'articule autour des données spatialement dépendantes. Ces données ont été largement traitées avec des méthodes paramétriques dont les plus connues sont les méthodes de Krigeage, du nom de l'ingénieur minier Danie Gerhardus Krige. Ces méthodes ont été formalisées par Georges Matheron dans les années 1960 (voir, par exemple, Matheron (1962)). Mais pour pallier à certaines difficultés, certains auteurs ont développé des outils issus de la statistique non paramétrique dont certains reposent sur l'estimateur à noyau, introduit pour l'estimation de la densité spatiale en 1990 par Tran (1990). La première partie de cette thèse est consacrée à la modélisation non paramétrique de données spatiales par la méthode du noyau. Plus particulièrement, les chapitres 2 et 3 concernent une nouvelle approche non paramétrique pour estimer les fonctions de densité de probabilité et de régression ainsi que le mode en présence de variables spatialement dépendantes à valeurs réelles.

Par ailleurs, depuis une trentaine d'années, les avancées technologiques ont notamment permis l'enregistrement massif de données, de manière plus régulière voir parfois continue. Ces données générées, dites “en grande dimension”, “en dimension infinie” ou encore “fonctionnelles”, ne peuvent pas être traitées avec des méthodes classiques de statistiques. Ainsi, une nouvelle dynamique de recherche est apparue favorisant l'essor des méthodes adaptées à l'étude de variables aléatoires fonctionnelles. Parmi elles, des techniques basées sur l'estimateur à noyau ont été développées. La seconde partie de cette thèse y est consacrée puisque le chapitre 4 adapte l'approche proposée au chapitre 3 aux données spatialement dépendantes de nature fonctionnelle. Jusqu'à présent, dans notre procédure d'estimation, nous avons tenu compte de la dépendance présente dans les données mais dans un cadre assez généraliste de variables mélangeantes, sans spécifier la structure de dépendance. Ainsi, nous proposons dans le chapitre 5 une manière de tenir compte d'une structure spécifique de dépendance en présence de données fonctionnelles (non spatialisées). En effet, nous proposons une nouvelle approche à noyau pour estimer la fonction de régression lorsque

les variables explicatives sont fonctionnelles et le processus des innovations est autocorrélé.

La dernière partie de ce manuscrit est consacrée aux applications. Nous avons, au chapitre 6, présenté des résultats de classification non supervisée de données réelles spatiales, simulées et réelles. La méthode considérée est basée sur l'estimation du mode spatial obtenu à partir de l'estimateur de la fonction de densité spatiale introduit dans le cadre de la première partie de cette thèse. Puis, au cours d'un projet de collaboration Franco-Québécoise, nous avons appliqué cette méthode de classification basée sur l'estimation du mode, ainsi que d'autres méthodes de classification non supervisée de la littérature, sur des données hydrologiques de nature fonctionnelle (non spatialisée). Ce travail est présenté au chapitre 7. Les résultats de classification obtenus sur les données hydrologiques nous ont ensuite amené, au chapitre 8, à appliquer des outils de détection de rupture sur ces données fonctionnelles.

Notons que le premier chapitre de ce manuscrit est consacré à l'explication de certains fondements de statistiques spatiales et fonctionnelles, nécessaires à la compréhension de ce travail pour un lecteur peu familiarisé avec ces outils. Ce chapitre 1 permet de mieux appréhender notre contribution aux statistiques spatiales et fonctionnelles, présentée aux parties I et II. En effet, ce chapitre 1 a pour objectif d'introduire les principales caractéristiques des statistiques spatiales et des statistiques pour données fonctionnelles. Nous présentons, de manière plus précise, le cadre statistique qui nous préoccupe. Nous établissons également un état de l'art des travaux sur le sujet. Pour terminer, nous annonçons les motivations qui nous ont conduit à développer les travaux présentés dans ce mémoire de doctorat. Lors de la réalisation des travaux présentés tout au long de ce manuscrit, des perspectives d'étude sont apparues et sont présentées en même temps que la conclusion. La fin de ce mémoire de thèse est composée d'une partie Annexe dans laquelle certains rappels sont formulés.

Dans la suite du document, la dépendance sera exprimée en terme de α-mélange, rencontré parfois sous le nom de mélange fort, dont les principes sont rappelés dans la partie Annexe.

Première partie : Statistique spatiale

La **partie I** de ce mémoire est consacrée à la modélisation de données, à valeurs réelles, géoréférencées, dites aussi données localisées ou spatiales. Nous nous intéressons d'une part à l'estimation de la fonction de densité spatiale (Chapitre 2) et d'autre part à la fonction de régression spatiale (Chapitre 3) en présence de variables α-mélangeantes.

Le **chapitre 2** concerne l'introduction et l'étude d'un nouvel estimateur à noyau de la fonction de densité de probabilité spatiale. Plus précisément, il s'agit d'une adaptation de l'estimateur classique mais dont la particularité est de tenir compte aussi de la proximité spatiale des données en intégrant l'information locationnelle. Pour cela, nous proposons de combiner deux noyaux, l'un sur les valeurs observées et l'autre sur les localisations spatiales des observations. On considère que la variable dépendante Y est réelle et que la variable explicative X est multivariée (réelle). Nous supposons également que les données sont α-mélangeantes. Après avoir étudié le comportement asymptotique de cet estimateur, et plus particulièrement la convergence uniforme presque sûre avec vitesse, nous l'utilisons pour estimer le(s) mode(s) d'une distribution spatiale dont les résultats de convergence sont également donnés. Enfin, une procédure de classification est présentée et sera appliquée dans la troisième partie de ce manuscrit. Notons que le travail présenté dans ce chapitre fait l'objet d'une publication (Dabo-Niang et al. (2014a)) dans *Stochastic Environmental Research and Risk Assessment*. Ce travail a été réalisé en collaboration avec Sophie Dabo-

Niang, Leila Hamdad et Anne-Françoise Yao.

Le **chapitre 3** fait référence à l'estimation non paramétrique de la fonction de régression spatiale dans un contexte similaire à celui considéré au chapitre 2, c'est à dire lorsque la variable réponse est scalaire et la variable explicative est multivariée (réelle), ceci dans le cadre des données spatialement dépendantes. L'estimateur est construit de la même manière que l'estimateur de la fonction de densité spatiale, étudié au chapitre 2, dont la particularité est de combiner deux noyaux permettant de contrôler à la fois la distance entre les observations et celle entre les sites. La convergence presque complète ainsi que la convergence en moyenne d'ordre q (norme L^q) ($q \in \mathbb{N}^*$) sont obtenues en considérant des processus α-mélangeants. Nous présentons également un prédicteur spatial issu de l'estimation de la régression. On termine ce chapitre par des illustrations sur des données issues de simulations ainsi que sur des données environnementales. Ce chapitre est issu d'un travail réalisé en collaboration avec Sophie Dabo-Niang et Anne-Françoise Yao.

Deuxième partie : Analyse de données fonctionnelles

Dans la **partie II** de ce mémoire, nous nous intéressons à la modélisation de données fonctionnelles. Tout d'abord, l'estimateur de la fonction de régression spatiale, présenté au chapitre 3, est étendu au cadre de données fonctionnelles spatialement dépendantes. Puis, nous proposons une méthode permettant de tenir compte de la structure de dépendance pour les séries chronologiques lorsque les variables explicatives sont fonctionnelles.

Le **chapitre 4** présente un travail étudiant à la fois les données spatialement dépendantes et fonctionnelles. Il s'agit d'une extension au chapitre 3 où les variables $X_{\mathbf{i}}$ ne sont plus à valeurs dans $\mathbb{R}^d$ mais dans un espace de dimension éventuellement infinie. Plus particulièrement, nous proposons un estimateur non paramétrique de la fonction de régression d'une variable spatiale scalaire conditionnellement à une variable fonctionnelle. La convergence en moyenne quadratique de cet estimateur est obtenue quand l'échantillon considéré est une séquence α-mélangeante. Pour terminer, des résultats numériques illustrent le comportement de notre estimateur. Le travail présenté dans ce chapitre fait l'objet d'une publication (Ternynck (2014)) dans *le Journal de la Société Française de Statistique*.

Dans le **chapitre 5**, nous nous intéressons à l'étude de séries chronologiques (temporelles) lorsque les erreurs sont autocorrélées. Nous considérons un modèle de régression pour lequel la variable explicative est fonctionnelle et la variable dépendante est réelle. Nous suggérons une méthode, basée sur l'estimateur à noyau, permettant de tenir compte de l'information contenue dans le terme d'erreur. Nous proposons d'utiliser une procédure préliminaire de décorrélation de la variable dépendante. L'idée principale est de transformer le modèle de régression original, de sorte que le terme d'erreur du modèle de régression transformé devienne non corrélé. La normalité asymptotique de l'estimateur proposé est établie en considérant la variable explicative α-mélangeante, le cas le plus général de variables faiblement dépendantes. Il est montré dans ce chapitre que la fonction d'autocorrélation du processus des erreurs apporte de l'information permettant d'améliorer l'estimateur de la fonction de régression. Les compétences de cette technique sont illustrées par une étude de simulations ainsi que par une application à des données de concentration en ozone. Ce chapitre est issu d'un travail réalisé en collaboration avec Sophie Dabo-Niang et Serge Guillas.

Troisième partie : Applications

La **partie III** est consacrée aux applications réalisées en parallèle ou en complément des chapitres précédents.

Le **chapitre 6** présente l'application à des données simulées puis réelles de la méthodologie développée au chapitre 2. L'objectif est d'appliquer l'estimation de la densité spatiale à la classification non supervisée basée sur le mode. Il s'agit de la méthode de classification descendante hiérarchique proposée dans Ferraty et Vieu (2006), Dabo-Niang et al. (2006, 2007) dans un cadre non-spatial puis adaptée dans Dabo-Niang et al. (2010) au cadre des données spatiales. Cette méthode repose sur l'utilisation d'un indice d'hétérogénéité basé sur l'estimation du mode. Les données réelles sur lesquelles nous avons travaillé portent sur les moussons en Asie.

Puis, dans le cadre d'une collaboration Franco-Québécoise, nous nous sommes intéressés à la modélisation de données hydrologiques avec des méthodes issues de l'analyse de données fonctionnelles. Notre objectif de classification des hydrogrammes des débits de rivières Québécoises, nous a amené, dans le **chapitre 7**, à utiliser la méthode de classification descendante hiérarchique, proposée dans Dabo-Niang et al. (2006, 2007) et présentée au chapitre 2, basée sur l'estimation du mode. Nous avons également considéré d'autres méthodes de classification non supervisée rencontrées dans la littérature pour données fonctionnelles. Ce chapitre établit tout d'abord un état de l'art des travaux sur la classification des hydrogrammes puis sur les méthodes de classification pour données fonctionnelles. Ensuite, après avoir rappelé les principes des méthodes utilisées pour traiter les données, nous les utilisons pour classifier les hydrogrammes de diverses stations hydrologiques du Québec. Nous comparons les résultats avec ceux obtenus par une méthode classique qui ne tient pas compte de la nature fonctionnelle de ces données hydrologiques. Les résultats sont analysés et accompagnés d'interprétations liées à l'hydrologie et à l'environnement. Mohamed Ali Ben Alaya, Fateh Chebana, Sophie Dabo-Niang et Taha Ouarda ont également contribué à la réalisation du travail présenté dans ce chapitre.

Certains résultats issus des applications présentées au chapitre 7, ont fait apparaître des classes dont la majorité des données d'une même classe proviennent d'une période de temps commune. Ainsi, nous nous sommes intéressés, dans le **chapitre 8**, à l'application de méthodes de détection de ruptures pour données fonctionnelles sur ces données hydrologiques.

Les trois parties de ce manuscrit sont suivies d'une conclusion dans laquelle des perspectives d'étude sont données. En complément de ce travail, une annexe est proposée à la fin de ce mémoire où des rappels sont énoncés ainsi que l'explication plus détaillée de certaines notions considérées au cours de cette thèse.

Communications écrites et orales

Travaux et Publications

- Spatial regression estimation for functional data with spatial dependency, *Journal de la Société Française de Statistique*, Numéro Spécial sur l'Analyse de Données Fonctionnelles. Vol. 155, No. 2, pages 138-160.
- A kernel spatial density estimation with applications to spatial clustering and Monsoon Asia Drought Atlas analysis (en collaboration avec S. Dabo-Niang, L. Hamdad et A.-F. Yao), *Stochastic Environmental Research and Risk Assessment*. Accepté, disponible en ligne.
- Streamflow hydrograph classification using functional data analysis (en collaboration avec M. Ali Ben Alaya, F. Chebana, S. Dabo-Niang et T.B.M.J. Ouarda), *Water Resources Research*, en révision.
- A new spatial regression estimator in the multivariate context (en collaboration avec S. Dabo-Niang et A.-F. Yao). *Compte Rendu Mathématique de l'Académie des sciences*, en révision.
- Spatial regression for multivariate data taking spatial characteristics into consideration and applications (en collaboration avec S. Dabo-Niang et A.-F. Yao). *Soumis.*
- Efficiency in functional nonparametric models with autoregressive errors (en collaboration avec S. Dabo-Niang et S. Guillas). *Soumis.*
- Change point detection of flood events using a functional data framework (en collaboration avec M. Ali Ben Alaya, F. Chebana, S. Dabo-Niang et T.B.M.J. Ouarda), *A soumettre*
- *Prévision spatiale non paramétrique*, Mémoire de Master 2, 2011

Séminaires et Conférences

- Séminaire "Modèles aléatoires et applications", laboratoire de Mathématiques Nicolas Oresme (LMNO), Caen, "Une nouvelle approche non paramétrique pour estimer la fonction de régression spatiale". Septembre 2014
- 2nd Conference of the International Society of NonParametric Statistics (ISNPS), Cadiz (Spain), "Efficiency in functional nonparametric models with autoregressive errors". Juin 2014
- 46*èmes* Journées de Statistique de la SFdS, Rennes, "Une nouvelle approche non paramétrique pour estimer la fonction de régression spatiale". Juin 2014
- Séminaire Inter Universitaire de Théorie Economique, laboratoire EQUIPPE, Lille, "Une nouvelle approche non paramétrique pour la régression et la prédiction spatiales". Mars 2014
- Groupe de travail "Systèmes d'information Géographique et Analyse Spatiale", laboratoires EQUIPPE et HALMA-PEL, Lille, "Introduction à la statistique spatiale". Juin 2013
- Séminaire Interne du laboratoire EQUIPPE, Lille, "Analyse et étude de phénomènes hydro-climatiques avec des approches de la statistique fonctionnelle". Décembre 2012

Participation à des conférences, séminaires et journées d'études

- Les séminaires du "Jeudi" d'Econométrie et de Statistique, laboratoire EQUIPPE, Lille, 2011-2014
- Workshop "Kernel methods for big data", Université de Lille 1, Avril 2014
- Journée YSP (Young Statisticians and Probabilists) organisée par la SFdS, Paris, Janvier 2014
- Journées Internationales Analyse Statistique : Théorie et Applications, Oujda (Maroc), Juin 2012
- Stat Learning, Université de Lille 1, Avril 2012
- Journées de Statistique Spatiale organisée par Liliane Bel, Agroparistech, Paris, Mars 2012
- International Workshop on Migrations and Development, laboratoire EQUIPPE, Lille, Novembre 2011
- European Statistical Meeting : Advances in the Treatment of Missing Data, Bruxelles, Novembre 2011

General introduction

Contents

This thesis was supported by a PhD contract from the University Lille 3 - Charles de Gaulle from October 1, 2011 to September 30, 2014. It was carried out within the laboratory EQUIPPE (Quantitative Economics, Integration, Public Policies and Econometrics) in Lille.

Presentation of the thesis

This doctoral dissertation mainly concerns statistical inference in a nonparametric framework, that is when the distribution of the observations is unknown. It matters with modeling the unknown characteristics of a population from a sample of it. Nonparametric estimation differs from parametric estimation by the fact that the structure of the model is not specified *a priori* but is directly determined from the data. It is said that these methods are "letting the data speak for themselves". Although in practice these methods require large data sets, they have the advantage of using only regularity assumptions (continuity, derivability, ...) on the link function between variables. We notice that in nonparametric estimation, it is often assumed that the link function belongs to some class $\mathcal{P}$, chosen amongst nonparametric classes of functions (Tsybakov (2009)). For example, if we are interested in the density function estimation, $\mathcal{P}$ can be the set of all the continuous probability densities on $\mathbb{R}$ or the set of all the Lipschitz continuous probability densities on $\mathbb{R}$.

The term "nonparametric" was first used in 1942 by Wolfowitz (1942). Since then, many contributors are working on the subject which is reflected in its vast literature. Among the many nonparametric issues encountered in the literature, we will focus on the estimation of density and regression functions as well as some of their applications. In order to deal with these two points, most developed tools involve interpolation, smoothing or approximation techniques. Among them, we can find, for instance, estimation methods based on spline basis, projection or even kernel method (so-called Parzen-Rosenblatt). In this work, the kernel method is mainly considered, it has been introduced by Rosenblatt (1956b) then generalized by Parzen (1962).

The probability density function, denoted by f all along this thesis, is a fundamental concept in statistic. It describes the distribution of the variable of interest, denoted X,

and allows to obtain probabilities associated with this variable. Density estimation is the construction of an estimator of the probability density function from the data. One initial approach, so-called parametric, is based upon the assumption that the data come from a known distribution family, for instance, the normal distribution with mean μ and variance σ^2. Thus the estimation of f involves estimating μ and σ parameters from the data and then substituting these estimators in the normal density formula. However the data does not always allow to determine *a priori* the distribution family it came from. The nonparametric methods, studied in the course of this work, do not require any *a priori* assumption on the fact that f belongs to a known family of distribution. The simplest way to estimate the density f from the data is the histogram. It brings out the kernel estimator considered in the following.

Specifically, a sample of n observations $X_1, \ldots, X_n$, independent and identically distributed (i.i.d.), with real values, is available. The kernel estimation of the density function f at point x is given by

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

where h is a smoothing parameter called "bandwidth" and K is a weight function called "kernel" which satisfies, for example, the following conditions

$$\int_{-\infty}^{\infty} K(x)dx = 1 \qquad \text{and} \qquad \int_{-\infty}^{\infty} K^2(x)dx < \infty.$$

This condition implies that K is also a probability density function.

Secondly, we consider the framework where we observe n pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, having same distribution than the pair of random variables (X, Y), that follow the regression model $Y_i = r(X_i) + \epsilon_i$ for $i = 1, \ldots, n$, and for which ϵ represents an error term. Thus, the regression problem consists in seeking a relation that can potentially exist between the variables X_i and Y_i. The variable X is the explanatory variable and the variable Y is the dependent variable. The variables ϵ_i, standing for the error term, are generally supposed independent from the X_i's and centered. Most often, the regression of random variables Y on X refers to the conditional expectation of Y given X. But in fact, the term "regression" refers to any element of the conditional distribution of Y given X considered as a function of X. It might be relevant, for example, to study the conditional median, the conditional mode, the conditional quantiles, etc. The most well-know (regression) model is the least square linear estimate which consists in explaining Y by a linear combination of the X components. Generally, the linear regression model refers to a model in which the conditional expectation of Y given X is an affine transformation of X. In fact, one can also consider models in which it is the conditional median or any other quantile of the distribution of Y given X that is an affine transformation of X. When we can not suppose that the regression function belongs to a parametric family, nonparametric methods of estimation are considered. The nonparametric approach only assumes that $r(\cdot)$ belongs to a given nonparametric class. In this work, we are particularly interested in the case where the link function is based on the conditional expectation. Indeed, the link or regression function will be the function $r(x)$ for which at any realization x of the explanatory variable X associates the quantity $\mathbb{E}[Y|X = x]$. From a kernel K and a bandwidth h, we can construct nonparametric estimates of the regression function $r(\cdot)$ similar to those of the density. The most famous, which we will study in this work, is the Nadaraya-Watson

estimator, introduced by Nadaraya (1964) and Watson (1964), which is defined by

$$r_n(x) = \frac{\frac{1}{nh}\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h}\right)}{f_n(x)} \qquad \text{if} \qquad f_n(x) \neq 0$$

and $r_n(x) = \frac{1}{n}\sum_{i=1}^{n} Y_i$ otherwise. This estimate can be seen as a weighted mean whose the weights depend on the distance between the variables X_i and x. In particular, these weights will allow to give more significance on the values of Y_i for which X_i is "close" to x. Whether it is the density or regression functions estimation, the quality of the estimate is strongly linked to the bandwidth h, for which the choice turns out to be very important and has been the subject of many works (e.g. Hall (1982), Delaigle and Gijbels (2004)). It is worth noting that, to some extent, results also vary according to the chosen kernel K.

The kernel estimates have been extensively studied in the literature, on one hand, for independent and real (univariate then multivariate) data, and on the other hand, for time series (dependent and real). More recently, the Parzen-Rosenblatt estimate has been adapted to other kinds of data as, for instance, spatial data. Indeed, since the mid 1950's, a new research focus is based on spatially dependent data. These data have been widely used with parametric methods, whose the most famous are the kriging methods, from the name of the mining engineer Danie Gerhardus Krige. These methods have been formalized by Georges Matheron in the 1960s (see, for instance, Matheron (1962)). But, in order to overcome some difficulties, some authors have developed some tools from the nonparametric statistics, some of which rely on the kernel estimate, introduced for the spatial density estimate in 1990 by Tran (1990). The first part of this dissertation will be devoted to the nonparametric modeling of spatial data by the kernel method. More particularly, chapters 2 and 3 concern a new nonparametric approach to estimate probability density and regression functions as well as the mode in presence of spatially real valued dependent data.

In addition, for about 30 years, technical advances have achieved to record more and more data, on a more regular way, sometimes continuous. The resulted data, called "high-dimensional", "infinite dimensional" or "functional", can not be dealt with classical statistical methods. Thus a new research dynamic has appeared, encouraging the expansion of methods adapted to the study of functional random variables. Among them, kernel estimate based techniques have also been developed. The second part of this thesis is devoted to these as chapter 4 adapts the proposed approach in chapter 3 to spatially dependent data of functional nature. Until now, in our estimation procedure, we have taken into consideration the dependence presents in the data but in a general framework of mixing variables, without specifying the dependence structure. Thus, we propose in chapter 5 a way enabling to take into account specific structure of dependence in presence of functional data (non-spatialized). Indeed, we propose a new kernel approach to estimate the regression function when explanatory variables are functional and the innovation process is autocorrelated.

The latter part of this manuscript is devoted to applications. Chapter 6 presents unsupervised classification results of real spatial data, simulated and real. The considered method is based on the spatial mode estimation obtained from the spatial density function estimate introduced in the first part of this thesis. Then, during a French-Quebec collaborative project, we applied this classification method based on the mode estimation, as well as other unsupervised classification methods of the literature, on hydrological data of functional nature (non-spatialized). This work is presented in chapter 7. The obtained

classification results on the hydrological data have led us, in chapter 8, to apply some change point detection tools on these functional data.

We note that the first chapter of this dissertation is devoted to the explanation of some spatial and functional statistical fundamental tools, necessary for an understanding of this work for a reader not very familiar with these tools. This chapter 1 gives a better understanding of our contribution to spatial and functional statistics, presented in part I and part II. Indeed, this chapter 1 aims to introduce main features of spatial statistics and functional data analysis. We present, in a more precise manner, the statistical framework that concerns us. We also establish a report on existing methods on the topic. Finally, we announce the motivations leading us to develop the works presented in this dissertation. When carrying out these presented works, some research perspectives appeared and are presented at the same time as the conclusion. The end of this manuscript is composed of an Appendix part in which some recalls are formulated.

Throughout the rest of the document, the dependence will be expressed in terms of α-mixing, sometimes labelled strong mixing, whose principles are recalled in the Appendix part.

First part: Spatial statistic

Part I of this dissertation is devoted to real valued georeferenced data modeling, also called spatially referenced data or spatial data. On one hand, we are interested in the spatial density function estimation (Chapter 2) and, on the other hand, in the spatial regression function (Chapter 3) in presence of α-mixing variables.

Chapter 2 is the introduction and study of a new kernel estimate of the spatial probability density function. Precisely, it is an adjustment of the classical estimate which specifically takes into account the spatial dependency of the data by incorporating locational information. For this purpose, we propose to combine two kernels, one based on the observed values and one based on the spatial locations of the observations. We consider that the dependent variable Y is real and the explanatory variable X is multivariate (real). We also assume that the data are α-mixing. Having studied the asymptotic behavior of this estimator, and more particularly the almost sure uniform convergence with rates, we use it in order to evaluate the mode(s) of a spatial distribution, its asymptotic results are also given. Finally, a clustering procedure is presented and will be applied in the third part of this manuscript. Note that the work presented in this chapter has been published (Dabo-Niang et al. (2014a)) in *Stochastic Environmental Research and Risk Assessment*. Sophie Dabo-Niang, Leila Hamdad and Anne-Françoise Yao collaborated on this work.

Chapter 3 looks at the nonparametric estimation of the spatial regression function in a similar context of chapter 2, that is when the response variable is scalar and the explanatory variable is multivariate, within the framework of spatially dependent data. The estimate follows the pattern of the one of the spatial density, studied in chapter 2, whose peculiarity is the combination of two kernels allowing to control both the distance between observation and that between locations. The almost complete consistency as well as the convergence in mean of order q (L^q norm) ($q \in \mathbb{N}^*$) are obtained considering α-mixing processes. Additionally, we propose a spatial predictor from the estimate of the spatial regression function. We conclude this chapter with illustrations from simulations and environmental data. This chapter is based on a collaborative work with Sophie Dabo-Niang and Anne-Françoise Yao.

Second part: Functional data analysis

In **part II** of this thesis, we are interested in functional data modeling. First, the spatial regression estimate, presented in chapter 3, is extended to the spatially dependent data framework. Then, we propose a method enabling to take into account the dependence structure for time series when explanatory variables are functional.

Chapter 4 presents a work combining both spatially dependent data and functional data. It is an extension of chapter 3 where variables $X_{\mathbf{i}}$ are no longer valued in $\mathbb{R}^d$ but are in an eventually infinite dimensional space. More particularly, we propose a nonparametric estimate of the regression function of a scalar spatial variable given a functional variable. The quadratic mean convergence of this estimate is obtained when the considered sample is an α-mixing sequence. Finally, numerical results illustrate the behavior of our estimate. The work presented in this chapter has been published (Ternynck (2014)) in *Journal de la Société Française de Statistique*.

In **chapter 5**, we are interested in the study of chronological series (time series) when the errors are autocorrelated. We consider a regression model in which the explanatory variable is functional and the response variable is scalar. We suggest a method enabling to take into account the information contained in the error term. A preliminary procedure of decorrelation of dependent variable is proposed. The main idea is to transform the initial regression model, so that the error term of the transformed model becomes uncorrelated. The asymptotic distribution of the proposed estimate is established considering that the explanatory variable is α-mixing, the more general case of weakly dependent data. It is shown in this chapter that the autocorrelation function of the error process brings information allowing to improve regression function estimate. The skills of this technique are illustrated by a study of simulations as well as ozone concentration data application. This chapter comes from a collaborative work with Sophie Dabo-Niang and Serge Guillas.

Third part: Applications

Part III is devoted to applications achieved in parallel or in complement to previous chapters.

Chapter 6 presents application to simulated and real data of the methodology developed in chapter 2. The purpose is to apply spatial density estimation to unsupervised classification based on the mode. It is the hierarchical descendant classification method proposed in Ferraty and Vieu (2006), Dabo-Niang et al. (2006, 2007) in the non-spatial case then adapted in Dabo-Niang et al. (2010) to the spatial framework. This method is based upon an heterogeneity index built on the mode estimation. We have worked on real data from the monsoon season in Asia.

Then, within the framework of a French-Quebec collaboration, we are interested in hydrological data modeling with methods from functional data analysis. Our purpose of classifying flood hydrographs from Quebec rivers, has led us, in **chapter 7**, to use hierarchical descendent classification method, proposed in Dabo-Niang et al. (2006, 2007) and presented in chapter 2, based on the mode estimation. We have also considered other unsupervised classification (clustering) methods for functional data encountered in the literature. We begin by doing a state of the art of hydrograph clustering and then of clustering methods for functional data. Afterwards, we recall the principles of methods that are used to process data. Finally, we classify hydrographs from various hydrologic stations of Quebec by presented methods. We compare the results with those obtained with a classical method that does not take into account the functional nature of the

hydrological data. The results are analyzed and accompanied by interpretations linked to hydrology and environment. Mohammed Ali Ben Alaya, Fateh Chebana, Sophie Dabo-Niang and Taha Ouarda have also contributed to the realization of the work presented in this chapter.

Some of the results from applications presented in chapter 7, have brought classes in which most of data of a same class come from a same period of time. Thus, we are interested, in **chapter 8**, in the application of change point detection methods for functional data on these hydrological data.

The three parts of this manuscript are ended with a conclusion giving several research perspectives. This work is supplementing by an Appendix proposed at the end of the manuscript where some recalls are given as well as a more detailed explanation of some considered notions throughout this thesis.

Written and oral communications

Works and Publications

- Spatial regression estimation for functional data with spatial dependency, *Journal de la Société Française de Statistique*, Special Issue on Functional Data Analysis. Vol. 155, No. 2, pages 138-160.
- A kernel spatial density estimation with applications to spatial clustering and Monsoon Asia Drought Atlas analysis (in collaboration with S. Dabo-Niang, L. Hamdad and A.-F. Yao), *Stochastic Environmental Research and Risk Assessment*. Accepted, available on-line.
- Streamflow hydrograph classification using functional data analysis (in collaboration with M. Ali Ben Alaya, F. Chebana, S. Dabo-Niang and T.B.M.J. Ouarda), *Water Resources Research*, in revision.
- A new spatial regression estimator in the multivariate context (in collaboration with S. Dabo-Niang and A.-F. Yao). *Compte Rendu Mathématique de l'Académie des sciences*, in revision.
- Spatial regression for multivariate data taking spatial characteristics into consideration and applications (in collaboration with S. Dabo-Niang and A.-F. Yao). *Submitted.*
- Efficiency in functional nonparametric models with autoregressive errors (in collaboration with S. Dabo-Niang and S. Guillas). *Submitted.*
- Change point detection of flood events using a functional data framework (in collaboration with M. Ali Ben Alaya, F. Chebana, S. Dabo-Niang and T.B.M.J. Ouarda), *To submit*
- *Prévision spatiale non paramétrique* (in French) [Nonparametric spatial prediction], Master thesis, 2011

Seminars and Conferences

- Seminar "Modèles aléatoires et applications", laboratory LMNO, *Laboratory of Mathematics Nicolas Oresme*, Caen (France), "Une nouvelle approche non paramétrique pour estimer la fonction de régression spatiale"(in French) [A new nonparametric approach to estimate the spatial regression function]. September 2014
- 2nd Conference of the International Society of NonParametric Statistics (ISNPS), Cadiz (Spain), "Efficiency in functional nonparametric models with autoregressive errors". June 2014
- 46*èmes* Journées de Statistique de la SFdS, Rennes (France), "Une nouvelle approche non paramétrique pour estimer la fonction de régression spatiale" (in French) [A new nonparametric approach to estimate the spatial regression function]. June 2014
- Seminar SIUTE, *Séminaire Inter Universitaire de Théorie Economique*, laboratory EQUIPPE, Lille (France), "Une nouvelle approche non paramétrique pour la régression et la prédiction spatiales" (in French) [A new nonparametric approach for the spatial regression and prediction functions]. March 2014
- Working group "Systèmes d'information Géographique et Analyse Spatiale", laboratories EQUIPPE and HALMA-PEL, Lille (France), "Introduction à la statistique spatiale" (in French) [Introduction to spatial statistics]. June 2013

- Internal Seminar of the laboratory EQUIPPE, Lille (France), “Analyse et étude de phénomènes hydro-climatiques avec des approches de la statistique fonctionnelle” (in French) [The analysis and study of hydro-climatic phenomenas with functional statistics approaches]. December 2012

Participation in conferences, seminars and workshops

- "Thursday's" Econometrics and Statistical seminars, laboratory EQUIPPE, Lille (France), 2011-2014
- Workshop "Kernel methods for big data", University Lille 1, Lille (France), April 2014
- Journée YSP (Young Statisticians and Probabilists) organized by the SFdS, Paris (France), January 2014
- Journées Internationales Analyse Statistique: Théorie et Applications, Oujda (Morocco), June 2012
- Stat Learning, University Lille 1, Lille (France), April 2012
- Journée de Statistique Spatiale organized by Liliane Bel, Agroparistech, Paris (France), March 2012
- International Workshop on Migrations and Development, laboratory EQUIPPE, Lille (France), November 2011
- European Statistical Meeting : Advances in the Treatment of Missing Data, Brussels (Belgium), November 2011

Chapitre 1

Concepts fondamentaux

Sommaire

1.1 Introduction à la statistique spatiale

"Everything is related to everything else,
but near things are more related than distant things."
Tobler (1970)

Les méthodes d'analyse spatiale sont utilisées dans de nombreuses disciplines où les données sont collectées en différentes localisations, c'est le cas, entre autres, en géologie, sciences du sol, traitement d'images, épidémiologie, agronomie, écologie, foresterie, astronomie, sciences atmosphériques, etc. Nous présentons dans ce paragraphe plusieurs situations réelles illustrant les enjeux liés à la modélisation spatiale des données. Dans le domaine océanographique, un des intérêts concerne la répartition des espèces de poissons en mer. Des campagnes de collecte de données, telle que la campagne annuelle IBTS, *International Bottom Trawl Survey*, permettent d'observer les abondances, ou encore les biomasses, d'espèces halieutiques en certains sites de la mer. A partir de ces observations, la prédiction des stocks de poissons sur les sites qui n'ont pas été observés est primordiale : détermination des quotas de pêche afin de préserver la ressource halieutique menacée par la pêche intensive. Une autre application possible de la statistique spatiale concerne la pollution des sols par les métaux lourds qui représente un risque important pour la santé. En effet, les sols sont à l'origine du transfert des métaux de l'environnement vers l'organisme : inhalation ou ingestion de particules (poussières) mais aussi d'aliments contaminés. Il est essentiel de pouvoir évaluer la teneur en métaux lourds dans les sols. Cependant, les techniques pour mesurer ces concentrations peuvent être lourdes à mettre en oeuvre (temps et coût). Ainsi des méthodes de prédiction spatiale peuvent aider à déterminer la concentration en métaux lourds dans le sol. Il peut également s'agir de modéliser la pollution atmosphérique. Par exemple, en France, des associations régionales comme "Atmo

Nord-Pas-de-Calais" surveillent la qualité de l'air et utilisent, entre autres, des outils issus de la statistique spatiale pour cartographier les polluants.

Selon les cas, les observations appartiennent à un certain type de données spatiales, à savoir, les données géostatistiques, les données latticielles et les processus ponctuels. Les méthodes à utiliser pour traiter ces données diffèrent selon leur type. Lorsque les données peuvent être mesurées en tout point d'un domaine continu, on se place dans le cadre de la géostatistique. Si les données sont liées à un réseau, on parle de données latticielles, que l'on va plutôt traiter avec la théorie des champs de Markov. Le dernier type de données spatiales, les processus ponctuels, survient lorsque c'est l'ensemble des sites où ont lieu les observations qui est étudié. Il n'est pas toujours aisé de déterminer le type de données à traiter. Cependant, le point commun entre ces catégories est la présence de dépendance dans une ou plusieurs directions mais qui s'affaiblit lorsque les sites d'observations sont plus éloignés. Les méthodes de statistique spatiale vont permettre, entre autres, l'analyse exploratoire des données, l'étude de la corrélation spatiale, leur modélisation jusqu'à la prédiction d'un phénomène en des sites non-observés. En effet, l'un des besoins de nombreuses disciplines scientifiques est la prédiction d'une variable en un site non-observé, par l'utilisation des observations de cette variable en d'autres sites. La prédiction spatiale dépend des observations sur les sites voisins de l'endroit où le champ est à prédire. D'où la nécessité de mesurer la dépendance entre localisations voisines. Cette caractéristique existe également pour les séries temporelles puisque les valeurs observées de variables proches dans le temps ont tendance à être similaires mais à se différencier lorsqu'elles sont observées sur une plus grande période. Ce qui distingue les données spatiales des séries temporelles est l'absence de relation d'ordre comme les notions de passé, présent et futur : l'axe du temps est unidirectionnel. En effet, les évènements passés peuvent avoir une influence sur le futur alors que l'inverse n'est pas vrai. Ainsi les modèles développés pour les séries chronologiques ne peuvent pas être directement appliqués aux données spatiales, ils se doivent d'être plus flexibles. Des différences ainsi que des similarités existant entre les séries spatiales et les séries temporelles sont soulignées dans Tjøstheim (1987).

De manière générale, la statistique spatiale étudie des phénomènes observés sur un ensemble spatial $\mathcal{S}$. Ces observations sont les réalisations d'un champ aléatoire Z sur $\mathcal{S}$, c'est à dire la donnée d'une collection $Z = \{Z_{\mathbf{i}}, \mathbf{i} \in \mathcal{S}\}$ de variables aléatoires indexées par l'ensemble spatial $\mathcal{S}$. La figure 1.1 illustre de façon simpliste un ensemble spatial $\mathcal{S}$ sur lequel on observe la variable aléatoire Z aux sites $\mathbf{i_1}, \ldots, \mathbf{i_n}$. Alors que la figure 1.2 représente des observations réelles, à savoir les cumuls pluviométriques dont l'ensemble spatial est le réseau météorologique Suisse. La nature de l'ensemble spatial $\mathcal{S}$ permet au statisticien de définir le type de données (géostatistiques, latticielles, processus ponctuels) dont il dispose. Dans le cas des données géostatistiques, principalement considérées dans ce travail, $\mathcal{S}$ est un sous-ensemble fixé de $\mathbb{R}^N$ avec $N > 1$. On dénote par $\mathbf{i} = (i_1, \ldots, i_N)$ un site localisé dans un espace Euclidien de dimension N.

Les propriétés asymptotiques des estimateurs peuvent être obtenues en étudiant leurs comportements lorsque $\mathbf{n} \to \infty$ selon deux situations à savoir, les divergences isotropiques ou non-isotropiques. La première correspond au cas le plus restrictif pour lequel on dit que la région $\mathcal{S}$ s'étend à l'infini à la même vitesse dans toutes les directions. Dans ce cas, les conditions suivantes sont requises : $\mathbf{n} \to \infty$ si $\min_{1 \leq k \leq N} n_k \to \infty$ et $n_j/n_k < C$ pour une constante C telle que $0 < C < \infty$ et $1 \leq j, k \leq N$. Dans l'autre situation, la région $\mathcal{S}$ ne s'étend pas à l'infini à la même vitesse dans toutes les directions et $\mathbf{n} \to \infty$ seulement si $\min_{1 \leq k \leq N} n_k \to \infty$.

De plus, notons qu'il existe généralement deux structures possibles pour l'étude asymp-

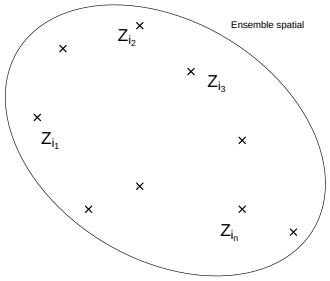


Figure 1.1 – Schéma simplifié d'un ensemble spatial $\mathcal{S}$
Observations de la variable aléatoire Z aux sites $\mathbf{i_1}, \ldots, \mathbf{i_n}$

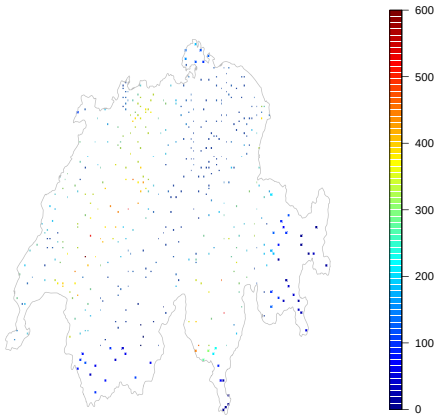


Figure 1.2 – Illustration d'observations spatiales à partir de données réelles
Observation des cumuls pluviométriques sur le réseau météorologique Suisse, le 8 mai 1986, en millimètre (passage du nuage de Tchernobyl). Les données sont disponibles avec le package *geoR* du logiciel R.

totique spatiale (le lecteur intéressé pourra se référer au Chapitre 5 de Gaetan et Guyon (2008)). Tout d'abord, l'asymptotique extensive (*increasing domain asymptotics*) qui concerne la situation où le nombre d'observations croît avec celui du domaine d'observation. Cette situation apparaît lorsque les sites sont séparés. La seconde asymptotique permet de traiter les situations où les observations se densifient dans un domaine $\mathcal{S}$ fixé et borné, on parle alors d'asymptotique intensive (*infill asymptotics*). Notons que la première structure est la plus commune dans la littérature pour données spatiales. Cependant, lorsque les observations sont faiblement dépendantes, la plupart des estimateurs consistants dans le cadre de l'asymptotique extensive ne le sont pas forcément en considérant l'approche intensive (voir Fazekas et Chuprunov (2006)). Par ailleurs, deux situations sont également possibles en ce qui concerne la répartition des sites d'observation. En effet, la position des sites d'observation peut être définie de manière déterministe (*fixed-design*) ou de manière aléatoire (*random design*).

C'est durant la première moitié du 20$^{\text{ème}}$ siècle, que les premières méthodes de statistique spatiale apparaissent. Ce sont des méthodes dites "paramétriques", lesquelles ont largement été étudiées, développées et utilisées, comme en témoigne la vaste littérature sur le sujet. Pour plus de détails sur ces méthodes, on pourra se référer, par exemple, à Ripley (1981), Cressie (1993), Chilès et Delfiner (1999), Gaetan et Guyon (2008). Plus

particulièrement, une méthode de prédiction spatiale est développée, dans le cadre de la géostatistique (on pourra se référer à la monographie de Wackernagel (2003) pour une introduction sur les méthodes géostatistiques), sous le nom de "Krigeage". Ce terme provient du nom de famille de l'ingénieur minier sud-africain Daniel Gerhardus Krige. Le Krigeage a été formalisé, pour la prospection minière, par Georges Matheron, à l'école des Mines de Paris (e.g. Matheron (1962)). Depuis, le domaine de ses applications a largement été étendu, touchant notamment l'océanographie, la météorologie, les sciences de l'environnement,... Cependant ces méthodes ne sont pas toujours adaptées aux données à traiter. Par exemple, le Krigeage suppose que les données suivent une loi normale ce qui, en pratique, n'est pas toujours vérifié. Ainsi, pour pallier les difficultés rencontrées par les méthodes paramétriques, certains auteurs ont développé des méthodes non paramétriques pour données spatiales, qui sont actuellement au coeur d'une dynamique de recherche (e.g. Journel (1983), Tran (1990), Carbon et al. (1997), Biau et Cadre (2004), Carbon et al. (2007), Dabo-Niang et Yao (2013), ...). Une part de ce travail de thèse se veut y contribuer en s'intéressant plus particulièrement à l'estimation fonctionnelle pour données spatiales : estimation des fonctions de densité de probabilité ainsi que son mode, et de régression spatiales.

1.1.1 L'estimation non paramétrique spatiale

Dans le cadre de cette thèse, nous nous intéressons d'une part à l'estimation de la fonction de densité de probabilité spatiale et de son mode et, d'autre part, à celle de la fonction de régression spatiale par des méthodes non paramétriques et plus précisément en considérant l'estimateur à noyau. Ces méthodes sont particulièrement utiles dans des situations pour lesquelles on ne peut dire avec confiance que ces fonctions appartiennent à une famille paramétrique. La majorité des travaux consacrés à l'étude non paramétrique de la densité ou de la régression spatiale, supposent l'ensemble spatial $\mathcal{S}$ être une région rectangulaire définie par $\mathcal{I}_{\mathbf{n}} = \{\mathbf{i} = (i_1, \ldots, i_n), 1 \leq i_k \leq n_k, k = 1, \ldots, N\}$. Notons que les résultats obtenus restent valides en considérant des régions de forme plus générale. Dans la suite, les champs aléatoires considérés sont également définis sur $\mathcal{I}_{\mathbf{n}}$ et on utilise la notation $\widehat{\mathbf{n}} = n_1 \times n_2 \times \cdots \times n_N$ pour faire référence à la taille de l'échantillon. On s'intéresse notamment à la situation où la position des sites est déterministe : d'où, sauf mention du contraire, on se place toujours dans un cadre déterministe. Ci-après, nous rappelons les principaux résultats rencontrés dans la littérature à ce sujet. Nous commençons par établir un état de l'art des travaux qui portent sur l'estimation à noyau de la fonction de densité de probabilité puis celle de la fonction de régression. Enfin, nous citons quelques travaux traitant par la méthode à noyau d'autres problématiques.

Estimation par la méthode à noyau de la fonction de densité de probabilité spatiale

Le concept de fonction de densité de probabilité est fondamental en statistique classique et l'est également en statistique spatiale. En effet, elle permet de décrire la distribution spatiale d'un champ aléatoire d'intérêt $X = \{X_{\mathbf{i}}, \mathbf{i} \in \mathcal{S}\}$ et en donne les probabilités qui lui sont associées. Elle permet également le développement d'autres outils statistiques, par exemple, pour l'estimation du mode spatial. Dans la suite, nous énonçons les travaux qui portent sur l'estimation à noyau de la fonction de densité de probabilité lorsque les données à traiter sont spatialement dépendantes.

Cas où les variables explicatives $X_\mathbf{i}$ sont dans $\mathbb{R}^d$ et les sites i sont dans $\mathbb{Z}^N$ Les premiers résultats sur l'estimation non paramétrique, de la fonction de densité spatiale sont dus à Tran (1990). L'auteur propose une généralisation de l'estimateur à noyau (ou dit de Parzen-Rosenblatt) pour estimer la densité multivariée d'un champ aléatoire stationnaire indexé par $\mathbb{Z}^N$. Il démontre, sous certaines conditions de mélange du processus, la normalité asymptotique de l'estimateur proposé. Des choix de fenêtre appropriés sont également donnés. On se place dans la situation où l'on observe un processus strictement stationnaire $\{X_\mathbf{i}\}$ sur l'ensemble spatial $\mathcal{I}_\mathbf{n}$. De plus, on suppose que $X_\mathbf{i}$ prend ses valeurs dans $\mathbb{R}^d$. L'estimation de la fonction de densité, en un point x à valeurs dans $\mathbb{R}^d$, proposée dans Tran (1990) est définie de la manière suivante :

$$f_\mathbf{n}(x) = \frac{1}{\widehat{\mathbf{n}} h_\mathbf{n}^d} \sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} K\left(\frac{x - X_\mathbf{i}}{h_\mathbf{n}}\right)$$

où $K(\cdot)$ est une fonction noyau et $h_\mathbf{n}$ est une séquence positive de fenêtres qui tendent vers 0 quand $\mathbf{n}$ tend vers l'infini. Tran et Yakowitz (1993) étudient un estimateur de f basé sur les k plus proches voisins (noté k-NN dans la littérature) et obtiennent la normalité asymptotique. Plus tard, Carbon et al. (1996) et Carbon et al. (1997) énoncent respectivement les conditions suffisantes pour que l'estimateur introduit dans Tran (1990) converge en norme L^1 et de manière uniforme en probabilité et presque sûrement. Dans le dernier, les vitesses L^∞ de convergence optimales suivantes sont obtenues :

$$\sup_{x\in\mathcal{D}} |f_\mathbf{n}(x) - f(x)| = O\left(\sqrt{\frac{\log \widehat{\mathbf{n}}}{\widehat{\mathbf{n}} h_\mathbf{n}^d}}\right) \qquad \text{en probabilité et presque sûrement}$$

où $\mathcal{D}$ dénote un ensemble compact arbitraire de $\mathbb{R}^d$. Ces résultats sont applicables à une large classe de processus spatiaux. Hallin et al. (2001) considèrent le problème de l'estimation de la densité d'un processus spatial linéaire, processus non nécessairement fortement mélangeant. Sous des conditions générales, l'estimateur à noyau pour un k-uplet de sites est montré être asymptotiquement normal multivarié. Biau (2002) étudie les moyennes quadratique et quadratique intégrée des erreurs d'estimation obtenues avec l'estimateur de Tran (1990). Il montre, sous de faibles conditions de mélange, que ces deux erreurs ont le même comportement asymptotique que dans le cas i.i.d.. El Machkouri (2011) a également étudié la normalité asymptotique de cet estimateur mais en s'appuyant sur des techniques différentes à celles utilisées dans les autres travaux cités. En effet, la plupart de ces travaux reposent sur la technique de décomposition par blocs avec des hypothèses de mélange alors que El Machkouri (2011) utilise la méthode de Lindeberg.

Cas où les variables explicatives $X_\mathbf{i}$ sont dans $\mathbb{R}^d$ et les sites i sont dans $\mathbb{R}^N$ Ces premières contributions sur l'estimateur à noyau de la fonction de densité spatiale considèrent des champs indexés de manière discrète alors que Biau (2003) étudie cet estimateur sur un ensemble spatial continûment indexé. La moyenne quadratique des erreurs est considérée et sous certaines conditions, des vitesses de convergence optimales et paramétriques sont obtenues. Fazekas et Chuprunov (2006) étudient également l'estimateur à noyau de la fonction de densité spatiale lorsque les sites sont dans $\mathbb{R}^d$. Cependant, ils combinent les deux structures d'asymptotique, intensive et extensive. En effet, ils montrent que l'estimateur est asymptotiquement normal si l'ensemble des sites d'observation devient de plus en plus dense dans une séquence croissante de domaines.

Estimation par la méthode à noyau de la fonction de régression spatiale

Un autre problème qui nous intéresse est celui de l'estimation de la fonction de régression, qui peut être utilisée à des fins de prévision spatiale. On se place dans le cadre où l'on dispose de $\widehat{\mathbf{n}}$ observations $Z_{\mathbf{i}} = (X_{\mathbf{i}}, Y_{\mathbf{i}})_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}$ obéissant au modèle de régression $r(x) = \mathbb{E}[Y_{\mathbf{i}}|X_{\mathbf{i}} = x]$, c'est à dire basé sur l'espérance conditionnelle de Y sachant X. Nous nous intéressons à l'estimation à noyau de la fonction de régression $r(\cdot)$ pour laquelle nous donnons ci-dessous les principaux résultats rencontrés dans la littérature sur le sujet. On notera que l'utilisation de la fonction de densité apparaît dans l'estimation de $r(\cdot)$.

Cas où les variables explicatives $X_{\mathbf{i}}$ sont dans $\mathbb{R}^d$ et les sites i sont dans $\mathbb{Z}^N$
Dans un premier temps, nous évoquons les travaux effectués lorsque le champ aléatoire considéré est indexé par $(\mathbb{N}^*)^N$. Nous pouvons d'abord citer les travaux de Lu et Chen (2002) qui considèrent des données issues d'un processus spatial mélangeant anisotropique. L'anisotropie spatiale repose sur le fait que la dépendance spatiale entre deux sites n'est pas forcément liée à la distance entre ces deux sites mais dépend également de la direction. La dépendance spatiale présente alors des caractéristiques différentes selon les directions. Ils suggèrent un estimateur à noyau de la fonction de régression conditionnelle spatiale qui est défini par

$$r_{\mathbf{n}}(x) = \frac{\dfrac{1}{\widehat{\mathbf{n}}h_{\mathbf{n}}^d}\displaystyle\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} K\left(\frac{x - X_{\mathbf{i}}}{h_{\mathbf{n}}}\right)}{f_{\mathbf{n}}(x)}$$

Des conditions sont données pour assurer la convergence faible ainsi que des vitesses de convergence de l'estimateur qu'ils proposent. Ils proposent une définition des coefficients de mélange adaptée à l'anisotropie. Dans Lu et Chen (2004), ces mêmes auteurs considèrent cet estimateur à noyau de la fonction de régression conditionnelle spatiale, sous des conditions de mélange du processus spatial, dans un cadre isotropique. Ils étudient des propriétés asymptotiques incluant la convergence faible et des vitesses de convergence de l'estimateur. Hallin et al. (2004b) étudient un estimateur linéaire local de la fonction de régression $r(\cdot)$, dont l'estimateur à noyau est un cas particulier. Sous de faibles conditions de régularité, la normalité asymptotique de l'estimateur proposé et celle de ses dérivées sont établies. Des choix appropriés de fenêtre sont donnés. Le processus spatial est supposé satisfaire des conditions très générales de mélange. Biau et Cadre (2004) s'intéressent à la prédiction spatiale de valeurs d'un champ aléatoire. Ils étudient d'abord le problème de régression spatiale basée sur l'estimation de la fonction de régression. Puis ils proposent un prédicteur spatial et montrent la convergence uniforme de ce dernier sur des ensembles compacts ainsi que sa normalité asymptotique. Carbon et al. (2007) étudièrent un modèle autorégressif non paramétrique dans le contexte de la prédiction sur des champs aléatoires. Plus précisément, ils considèrent l'estimation de la régression de $\{X_{\mathbf{n}}\}$ sur les valeurs du champ aléatoire aux sites environnants, disons, $X_{\mathbf{n}_1}, \ldots, X_{\mathbf{n}_l}$. On remarque que $(X_{\mathbf{n}_1}, \ldots, X_{\mathbf{n}_l})$ est un vecteur de dimension ld. Ils supposent l'existence de la fonction de régression $r(x) = \mathbb{E}[\varphi(X_{\mathbf{n}})|(X_{\mathbf{n}_1}, \ldots, X_{\mathbf{n}_l}) = x]$, où x est à valeurs dans $\mathbb{R}^{dl}$ et φ est une fonction continue à valeurs réelles, non nécessairement bornée. Un estimateur à noyau de $r(x)$ est étudié. Ils montrent, sous des conditions générales, que cet estimateur converge uniformément sur des ensembles compacts. Des vitesses de convergence optimales dans L^∞ sont également atteintes. Li et Tran (2009) proposent une méthode alternative pour estimer non paramétriquement la fonction de régression spatiale. Cette méthode est basée sur une pondération par la technique des plus proches voisins. Ils montrent la normalité asymptotique de l'estimateur issu de cette méthodologie. Gheriballah et al. (2010) étu-

dient l'estimation non paramétrique robuste de la fonction de régression spatiale. Plus précisément, ils considèrent une famille d'estimateurs non paramétriques robustes pour la fonction de régression basée sur la méthode à noyau. Sous des conditions générales de mélange, la convergence presque complète et la normalité des estimateurs sont obtenues. Une procédure robuste de sélection des paramètres de lissage adaptée aux données spatiales est discutée.

Robinson (2011) propose de "ranger" les données spatiales sous la forme basique de matrice triangulaire pour estimer la fonction de régression. Ceci permet de tenir compte de diverses formes d'observations spatiales. Il autorise le fait que les observations soient non-identiquement distribuées et présentent de l'hétéroscédasticité conditionnelle. Au lieu des conditions de mélange, un processus linéaire (éventuellement non stationnaire) est supposé pour les perturbations. Des conditions suffisantes sont établies pour la convergence et la normalité asymptotique de l'estimateur à noyau de la régression.

Cas où les variables explicatives $X_{\mathbf{i}}$ sont dans $\mathbb{R}^d$ et les sites i sont dans $\mathbb{R}^N$ Dabo-Niang et Yao (2007) étudient l'estimateur à noyau de la fonction de régression spatiale pour un processus stationnaire spatial multidimensionnel $\{Z_{\mathbf{i}} = (X_{\mathbf{i}}, Y_{\mathbf{i}}), \mathbf{i} \in \mathbb{R}^N\}$. Les convergences faible et forte de l'estimateur à noyau de $r(\cdot)$ sont montrées sous des conditions suffisantes sur les coefficients de mélange et sur les fenêtres. Ils obtiennent des vitesses optimales et sur-optimales des vitesses de convergence fortes. Ils proposent également une première approche de prédiction spatiale pour des processus continûment indexés. Dans le travail de Menezes et al. (2010), la prédiction non paramétrique à noyau est considérée pour des processus stochastiques spatiaux lorsque les sites d'observations sont échantillonnés de manière aléatoire. Sous des hypothèses assez générales, ils montrent que l'erreur quadratique moyenne du prédicteur tend à être négligeable quand la taille de l'échantillon augmente. Ils proposent des approches alternatives de validation croisée pour sélectionner les fenêtres locales et globales. Notons que dans ce travail, les auteurs considèrent le cadre de l'asymptotique intensive, c'est à dire que les données sont collectées dans une région bornée qui ne croît pas avec la taille de l'échantillon. Karácsony et Filzmoser (2010) obtiennent la normalité asymptotique de l'estimateur de la fonction de régression de type Nadaraya-Watson pour des champs aléatoires α-mélangeants. Ils considèrent la structure extensive-intensive d'asymptotique, c'est à dire quand les sites d'observations deviennent denses dans une séquence croissante de domaines. Cette structure comble le fossé entre les modèles continus et discrets.

D'autres problématiques traitées par la méthode à noyau

Le cadre du "*fixed-design setting*" Dans le cadre du "fixed-design setting", El Machkouri (2007) ainsi que El Machkouri et Stoica (2010) considèrent le modèle de régression suivant

$$Y_i = r\left(\frac{i}{n}\right) + \epsilon_i$$

où $(\epsilon_i)_{i\in\mathbb{Z}^d}$ est un champ aléatoire stationnaire réel de moyenne nulle avec $i \in \{1, \ldots, n\}^d$ et $n \in \mathbb{N}^*$. Les erreurs sont issues d'un champ de variables aléatoires dépendantes. Ils proposent un estimateur basé sur la méthode à noyau. Tout d'abord, dans El Machkouri (2007), ils obtiennent des conditions suffisantes pour que l'estimateur converge de manière uniforme. Ils montrent également qu'il est possible d'obtenir des vitesses optimales de convergence et que les résultats s'appliquent à une large classe de champs aléatoires comprenant les champs aléatoires α-mélangeants ainsi que les champs aléatoires de type

différences de Martingale. Puis, dans El Machkouri et Stoica (2010), ils établissent la normalité asymptotique de l'estimateur considéré et montrent également que les résultats s'appliquent à une large classe de champs aléatoires.

Régressions basées sur d'autres caractéristiques conditionnelles Jusqu'ici, nous avons présenté des estimateurs non paramétriques de la fonction de régression spatiale basée sur l'espérance conditionnelle. Cependant, la régression peut désigner tout autre élément de la distribution conditionnelle de Y sachant X, considérée comme une fonction de X. Par exemple, il peut s'agir de la médiane conditionnelle, du mode conditionnel, des quantiles conditionnels, etc. Nous allons présenter brièvement des travaux qui traitent de régression non paramétrique pour données spatialement dépendantes mais qui ne reposent pas sur l'espérance conditionnelle de Y par rapport X.

Les travaux cités dans ce paragraphe concernent le cas où les sites sont indexés de manière discrète. Hallin et al. (2009) proposent un estimateur linéaire local de la fonction de quantile conditionnel spatial. Le travail de Dabo-Niang et Thiam (2010) présente un cadre statistique pour modéliser les quantiles conditionnels des processus spatiaux $(X_{\mathbf{i}}, Y_{\mathbf{i}})$ à valeurs dans $\mathbb{R}^d \times \mathbb{R}$ et supposés fortement mélangeants. Ils définissent également un prédicteur non paramétrique spatial. Ould-Abdi et al. (2010b) et Ould-Abdi et al. (2011) considèrent un processus stationnaire spatial $(X_{\mathbf{i}}, Y_{\mathbf{i}})$ à valeurs dans $\mathbb{R}^d \times \mathbb{R}$. Ils étudient un estimateur à noyau de la fonction de quantile conditionnel spatial, plus général que celui de Dabo-Niang et Thiam (2010), d'une variable réponse $Y_{\mathbf{i}}$ sachant la variable explicative $X_{\mathbf{i}}$. Parfois, le mode conditionnel s'avère être un outil plus adapté pour la prédiction que l'espérance conditionnelle. C'est le cas, par exemple, si la densité conditionnelle est bimodale ou dissymétrique. A notre connaissance, peu de travaux sont consacrés à l'estimation à noyau du mode conditionnel spatial. Les premières contributions sur le sujet sont celles de Ould-Abdi et al. (2010a) et de Dabo-Niang et al. (2014b).

Dans cette section, nous avons essayé de rassembler l'ensemble de travaux rencontrés dans la littérature traitant de l'estimateur à noyau en présence de données spatialement dépendantes. La section suivante énonce les motivations qui nous ont conduit à développer une nouvelle approche à noyau pour estimer les fonctions de densité et de régression spatiale.

1.1.2 Contribution à la modélisation non paramétrique spatiale

Dans la partie I de cette thèse, nous nous sommes attachés à développer une nouvelle approche basée sur l'estimateur à noyau pour estimer les fonctions de densité de probabilité et de régression en présence de données spatialement dépendantes à valeurs réelles. En particulier, nous voulons que les observations proches du site étudié aient plus d'influence que les observations faites en des sites plus éloignés. Notre motivation se résume assez bien dans la phrase suivante "*Everything is related to everything else but near things are more related than distant things*", Tobler (1970). Pour cela, l'approche que nous proposons est basée sur le produit de deux fonctions noyaux, l'une sur les observations et l'autre sur les sites. Ainsi les estimateurs tiennent compte à la fois de la distance entre les observations et celle entre les sites. Cette idée a été introduite dans Hall et al. (2006) dans le cadre des séries temporelles dynamiques. L'estimateur étudié est construit à partir d'un produit de deux noyaux, l'un sur les dates d'observation, et l'autre sur les observations. Ils considèrent l'observation d'un processus multivarié dont la distribution évolue avec le temps. L'estimation à l'instant t tient compte de certaines observations passées : la date la plus ancienne dépend du paramètre de fenêtre du noyau sur le temps. Cette idée est également

présente dans le cadre de données spatio-temporelles avec les travaux de Wang et Wang (2009) et de Wang et al. (2012). Dans le contexte des processus ponctuels spatiaux, Menezes et al. (2010) proposent un prédicteur spatial construit avec un seul noyau sur les sites.

Les chapitres 2 et 3 traitent de cette nouvelle approche lorsque le champ aléatoire étudié est tel que les variables dépendantes X sont multivariées à valeurs dans $\mathbb{R}^d$ et la variable réponse Y est réelle. Comme souvent dans la littérature sur le sujet, la dépendance sera exprimée en terme de α-mélange (que l'on appelle aussi mélange fort) dont des rappels sont formulés dans l'Annexe. Dans un premier temps, dans le chapitre 2, nous proposons un nouvel estimateur de la fonction de densité de probabilité. Sous certaines hypothèses, nous montrons la convergence uniforme presque sûre de l'estimateur, nous donnons également des vitesses de convergence. Ensuite nous en déduisons une estimation du mode spatial (non conditionnel) dont la convergence uniforme est obtenue. Cette estimation est utilisée à des fins de classification non supervisée. Des résultats numériques issus d'une étude de simulation ainsi que de l'application à des données environnementales sont donnés dans la partie III de ce manuscrit consacrée aux applications. Dans un second temps, dans le chapitre 3, nous traitons l'estimation de la fonction de régression, basée sur l'espérance conditionnelle, par cette nouvelle approche. Les variables sont de même nature que celles considérées dans le chapitre 2. Nous étudions le comportement asymptotique de cet estimateur, en particulier, nous montrons les convergences presque complète et en moyenne d'ordre q. Nous donnons également des vitesses de convergence. A partir de l'estimation de la fonction de régression, nous construisons un prédicteur spatial. Enfin, nous appliquons cette méthode à des données simulées puis réelles. Nous comparons les résultats avec ceux obtenus en considérant des méthodes classiques.

Dans cette section, nous avons introduit le cadre statistique qui nous intéresse pour traiter la partie I de ce manuscrit. La section suivante présente les mêmes objectifs mais concerne plutôt les travaux présentés dans la partie II.

1.2 Introduction à l'analyse de données fonctionnelles

Au début des années 1980, une nouvelle branche de la Statistique émerge et connait depuis un important essor notamment avec les monographes de Ramsay et Silverman (1997; 2002; 2005), de Bosq (2000), Bosq et Blanke (2008), de Ferraty et Vieu (2006) et de Horváth et Kokoszka (2012). Il s'agit de la statistique pour "données fonctionnelles" ou "données de dimension infinie" ou encore dite en "grande dimension". Ces dernières apparaissent avec les avancées technologiques (appareil de mesure, capacité de stockage, etc.) qui permettent de récolter des données discrétisées de manière de plus en plus fine. En effet, les méthodes de statistique pour données multivariées rencontrent des difficultés à traiter des données dont la dimension devient trop importante. Par exemple, lorsque la dimension des données dépasse la taille de l'échantillon, la méthode des moindres carrés ordinaires rencontre des problèmes liés à l'inversion des matrices. Une autre difficulté rencontrée par les méthodes statistiques traditionnelles, pour traiter des courbes finement discrétisées, provient de l'existence de fortes corrélations entre les variables. C'est pour contourner ce genre de difficulté qu'est apparue la statistique pour données fonctionnelles. Notons que ces données ne sont pas, par nature, directement fonctionnelles, c'est le praticien qui les rend fonctionnelles (voir Ramsay et Silverman (2005)). Par exemple, des enregistrements

rapprochés des données permettent de les lisser afin d'obtenir des courbes. Les données fonctionnelles surviennent dans de nombreux domaines comme, par exemple, en médecine (courbes de croissance), en hydrologie (débits), en économie (cours boursiers). Les exemples cités ne donnent qu'un infime aperçu de la variété des domaines concernés. Il convient aussi de constater que la notion de variable fonctionnelle couvre un rayon plus large que l'analyse de courbes. En effet, une variable fonctionnelle peut également être une surface aléatoire ou un objet mathématique plus complexe de dimension infinie. On pourra se référer à Cuevas (2014) pour un aperçu partiel mais récent de la théorie des statistiques pour données fonctionnelles.

Les auteurs qui travaillent sur le sujet ont proposé de nombreux outils statistiques pour traiter ce nouveau type de données. Cependant, on distingue plusieurs catégories de techniques liées à la nature supposée des variables fonctionnelles. Plus précisément, on considère qu'il existe d'une part les modèles linéaires (paramétriques, voir e.g. Ramsay et Silverman (2005)) et d'autre part les modèles non paramétriques (voir e.g. Ferraty et Vieu (2006)) pour données fonctionnelles. En particulier, une variable fonctionnelle appartient à un espace fonctionnel et selon la nature de ce dernier, les possibilités de traitement diffèrent. Par exemple, on peut supposer que les données appartiennent à un espace de Banach de fonctions réelles continues, $X : [0,1] \to \mathbb{R}$, muni de la norme sup. On peut également considérer un espace de Hilbert des fonctions réelles de carré intégrable sur $[0,1]$, muni du produit scalaire usuel $\langle f, g\rangle = \int_0^1 f(t)g(t)dt$. Ces deux exemples concernent plutôt une modélisation linéaire, alors que pour une approche non paramétrique on utilisera, par exemple, différentes semi-normes (voir Ferraty et Vieu (2006)). Dans le cadre de cette thèse, nous avons essentiellement considéré la seconde approche (non paramétrique) pour traiter des données fonctionnelles qui sont définies dans Ferraty et Vieu (2006) de la manière suivante

Définition 1.1. Une variable aléatoire X est appelée variable fonctionnelle si elle prend ses valeurs dans un espace de dimension infinie (ou un espace fonctionnel). Une observation de X est appelée une donnée fonctionnelle.

Comme en statistique classique, de nombreux problèmes impliquant des statistiques pour données fonctionnelles existent, la communauté statisticienne a développé de nombreux modèles pour traiter de tels jeux de données. Certains étudient des modèles de régression et proposent des estimateurs de l'espérance conditionnelle (Dabo-Niang et Rhomari (2003), Masry (2005), Berlinet et al. (2011), ...), du mode conditionnel (Ferraty et al. (2005a), Demongeot et al. (2013), ...) ou encore des quantiles conditionnels (Ferraty et al. (2005b), Laksaci et al. (2009), ...). D'autres auteurs s'intéressent plutôt à la classification de courbes (Dabo-Niang et al. (2006), Delaigle et al. (2012), ...), à la détection de rupture (Hörmann et Kokoszka (2010), Horváth et al. (2010), ...) et de données aberrantes (Febrero et al. (2007, 2008), Hyndman and Shang (2010), ...), à l'analyse en composantes principales (Hall and Hosseini-Nasab (2006), Shang (2014), ...). Des travaux portent également sur les tests d'indépendance ou de corrélation des erreurs (Gabrys et Kokoszka (2007), Gabrys et al. (2010), Horváth et al. (2013), ...), etc. Notons que les variables fonctionnelles étudiées peuvent être considérées comme indépendantes mais aussi dépendantes.

Afin de positionner nos travaux parmi ceux rencontrés dans la littérature, nous proposons ci-après un état de l'art des travaux relatifs aux problématiques pour données fonctionnelles que nous avons traitées, à savoir celles liées à l'estimation de la fonction de régression.

1.2.1 Estimation à noyau de la régression pour variables fonctionnelles

On s'intéresse au cas de la régression d'une variable réelle Y sur une variable fonctionnelle X. C'est la forme de régression la plus courante dans la littérature et celle que nous avons considérée dans la partie II de ce manuscrit. Bien que différentes méthodes ont été développées, selon la nature de l'opérateur de régression, pour traiter cette question, on se concentrera sur les travaux basés sur l'estimateur à noyau.

Cas des variables indépendantes

Soient $(X_i, Y_i)_{i=1,\dots,n}$, n paires de variables indépendantes et identiquement distribuées de même loi que (X, Y) et à valeurs dans $\mathcal{E}\times\mathbb{R}$ où $(\mathcal{E}, d)$ est un espace semi-métrique, autrement dit, X est une variable aléatoire fonctionnelle et d est une semi-métrique. On notera x un élément fixé de $\mathcal{E}$ et y un élément fixé de $\mathbb{R}$. A partir des observations $(X_i, Y_i)_{i=1,\dots,n}$, on souhaite estimer la valeur de y sachant celle de x. Pour cela, on s'intéresse à l'estimation de la fonction de régression r de Y sachant X définie par $r(x) = \mathbb{E}[Y|X = x]$. L'estimateur à noyau de r rencontré dans la littérature (introduit dans Ferraty et Vieu (2000)) est défini par

$$\widehat{r}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{d(x,X_i)}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{d(x,X_i)}{h_n}\right)} \tag{1.1}$$

où K est un noyau asymétrique et h_n est un nombre positif réel qui dépend de n. Il s'agit de l'extension au cadre des données fonctionnelles de l'estimateur de Nadaraya-Watson (Nadaraya (1964), Watson (1964)) introduit pour des données réelles et rappelé dans l'introduction générale de ce manuscrit. La différence principale apparait dans l'objet utilisé pour mesurer la proximité entre les variables explicatives. En effet, l'une des spécificités des données fonctionnelles est la grande dimension des données. De plus, en dimension finie toutes les métriques sont équivalentes, ce qui n'est plus le cas en dimension infinie dont le choix devient crucial. Par conséquent, les normes classiques utilisées pour mesurer la proximité entre les variables ne sont plus adaptées aux données fonctionnelles. Certains auteurs proposent d'utiliser une semi-métrique dont le choix dépend de la nature des données et du problème considéré. Les semi-métriques permettent également de contourner le problème du fléau de la dimension. Des explications plus détaillées sur les semi-métriques sont données en Annexe. Notons que dans le cas où K est un noyau positif à support $[0, 1]$, plus la valeur de $d(x, X_i)$ est petite, plus grande sera la valeur de $K(h^{-1}d(x, X_i))$. Ainsi l'estimateur (1.1) de y ne tient compte que des Y_i pour lesquels le X_i correspondant est distant de x d'au plus h.

Des résultats de convergence de cet estimateur ont été obtenus dans la littérature. Ferraty et Vieu (2000) puis Ferraty et Vieu (2002) introduisent cet estimateur en considérant un modèle de régression dans lequel la variable explicative est à valeurs dans un espace vectoriel semi-normé et la variable réponse est scalaire. Leur démarche repose sur une hypothèse de la loi de probabilité de la variable explicative interprétable en termes de dimension fractale. La convergence presque sûre de leur estimateur est établie et des vitesses de convergence sont obtenues. Sous des conditions générales, Dabo-Niang et Rhomari (2003) étudient l'estimateur à noyau de la régression quand la variable explicative prend ses valeurs dans un espace semi-métrique et la variable réponse est réelle. Les convergences en moyenne d'ordre q et presque sûre sont données avec des bornes supérieures des erreurs d'estimation. Dans Ferraty et al. (2007), les auteurs considèrent que la variable réponse est réelle et la variable explicative est à valeurs dans un espace fonctionnel $\mathcal{E}$, supposé être un espace séparable de Banach muni d'une norme (voir aussi le travail de

Dabo-Niang et Rhomari (2009)). La convergence en moyenne quadratique de l'estimateur à noyau est obtenue avec vitesse et expressions explicites des constantes. La normalité asymptotique, dont la particularité est l'expression explicite de la loi asymptotique, c'est à dire des termes dominants de biais et de variance, est également montrée.

Nous avons fait état de la situation où l'estimateur à noyau de la régression pour données de dimension infinie s'applique aux échantillons indépendants. Dans la suite, nous allons citer les travaux non paramétriques qui considèrent cet estimateur pour des échantillons dépendants et en particulier α-mélangeants. On s'intéressera d'abord au cas des séries temporelles puis à celui des séries spatiales.

Cas des variables chronologiquement dépendantes

Nous allons d'abord rappeler le contexte des séries chronologiques et le lien que l'on peut faire avec les données fonctionnelles puis nous citerons les travaux existants sur le sujet. Une série temporelle (ou chronologique) est issue de l'enregistrement d'un phénomène au cours du temps. L'observation d'une série temporelle peut être considérée comme la réalisation d'un processus stochastique à modéliser. Ces séries apparaissent dans une très grande variété de domaines comme, par exemple, en biologie, en économie, en finance, en hydrologie, en médecine, et bien d'autres. Notons que le pas de temps diffère selon les disciplines et leurs problématiques. En effet, les enregistrements peuvent être annuels, mensuels, hebdomadaires, journaliers, horaires, ... La modélisation de ces séries permet, entre autres, de répondre au besoin de prédire l'intensité d'un évènement futur à partir des observations passées de cet évènement, en tenant compte de l'ordre chronologique des observations. On notera $\{Y_i\}$, $i = 1, \ldots, n$, une série temporelle de longueur n. Ces séries ont été beaucoup étudiées dans la littérature pour données réelles et le sont encore aujourd'hui. Le lecteur pourra se référer à l'ouvrage de Brockwell et Davis (2002) pour une introduction aux séries temporelles et à la prévision ainsi qu'à Brockwell et Davis (2009) pour plus de détails théoriques sur l'analyse des séries temporelles. De nombreuses techniques d'abord issues de la statistique paramétrique, puis non- et semi- paramétriques ont été développées pour traiter ces données.

Plus récemment des auteurs se sont intéressés au traitement des séries chronologiques par des techniques issues de la statistique pour données fonctionnelles. Notons que dans ce travail, on se focalise sur les techniques non paramétriques. Pour bien comprendre le contexte, on se place dans la situation où nous disposons d'observations discrétisées (jusqu'à une certaine date T) $\{Z_1, \ldots, Z_T\}$ du processus $\{Z_i, i \in \mathbb{R}\}$ et nous souhaitons prédire une valeur Z_{T+s} du processus. Comme mentionné dans Ferraty et Vieu (2006), la manière de procéder la plus simple pour prédire la valeur Z_{T+s} est de tenir compte d'une seule observation passée. Pour cela, on construit un échantillon statistique (X_i, Y_i) de taille $n = T - s$, de la façon suivante

$$X_i = Z_i \quad \text{et} \quad Y_i = Z_{i+s}, \qquad i = 1, \ldots, T - s.$$

Il s'agit d'un problème de prédiction standard d'une variable réponse réelle Y à partir d'une variable explicative réelle X mais cette modélisation pourrait ne pas tenir compte de suffisamment d'informations passées. On peut penser à construire un échantillon statistique de dimension $p + 1$ et de taille $n = T - s - p + 1$ et considérer

$$X_i = (Z_{i-p+1}, \ldots, Z_i) \quad \text{et} \quad Y_i = Z_{i+s}, \qquad i = p, \ldots, T - s$$

de façon à obtenir un problème standard de prévision d'une variable réponse réelle Y à partir d'une variable explicative de dimension p. Mais, un problème récurrent avec l'ap-

proche non paramétrique lorsque les données sont multivariées, est ce qu'on appelle dans la littérature le "fléau de la dimension" (curse of dimensionality).

Dans ce travail, nous avons considéré un cadre purement non paramétrique. Plus précisément, nous avons choisi l'approche fonctionnelle de prévision, comme expliquée, par exemple, dans Ferraty et Vieu (2006). Il s'agit de considérer les valeurs explicatives passées comme toute une trajectoire continue du processus. Les observations X_i sont alors des courbes, éventuellement obtenues par interpolation d'observations discrètes. Pour simplifier la notation, on suppose que $T = n\tau$ pour $n \in \mathbb{N}^*$ et un certain $\tau > 0$. Nous pouvons construire un nouvel échantillon statistique de taille $n-1$ de la manière suivante :

$$X_i = \{Z(t), (i-1)\tau < t \leq i\tau\} \quad \text{et} \quad Y_i = Z(i\tau + s), \quad i = 1, \ldots, n-1 \tag{1.2}$$

c'est à dire, un problème de prédiction d'une variable réponse scalaire Y sachant une variable fonctionnelle X. L'idée de découper le processus en trajectoires successives de longueur τ a été introduite par Bosq (1991).

Le cadre qui nous intéresse étant posé, à savoir celui des séries chronologiques avec observations fonctionnelles, il nous faut parler des hypothèses à émettre sur la dépendance de l'échantillon. Dans la littérature, on rencontre souvent l'hypothèse de α-mélange pour modéliser cette dépendance et c'est ce type de dépendance qui sera principalement considéré ici. Pour plus de détails sur les notions de mélange, le lecteur pourra se référer à l'Annexe. Dans ce qui suit, nous allons énoncer les travaux non paramétriques qui traitent de données fonctionnelles α-mélangeantes avec applications, parfois, aux séries temporelles qui en sont un cas particulier. En effet, l'hypothèse de α-mélange rend les résultats opérationnels en prédiction de séries chronologiques.

Les premiers résultats dans ce cadre sont ceux de Ferraty et al. (2002b) qui considèrent une suite (X_i, Y_i) fortement mélangeante, avec $Y_i \in \mathbb{R}$ et X_i appartenant à un espace vectoriel semi-normé. Un estimateur à noyau de l'opérateur de régression est proposé, accompagné d'un résultat de convergence presque complète uniforme. Des conditions de régularités et une hypothèse sur la distribution des X_i liée à la dimension fractale sont également faites. Ferraty et al. (2002a) s'intéressent à un cas particulier de données fonctionnelles, celui où les données proviennent d'une série chronologique continue. L'objectif est de prédire les valeurs futures ($Y \in \mathbb{R}$) d'un processus en tenant compte de l'ensemble continu du passé (X à valeurs dans un espace linéaire semi-normé). Pour caractériser le modèle de dépendance entre les paires de variables aléatoires, ils utilisent les conditions de α-mélange. Ils utilisent l'estimateur présenté dans Ferraty et al. (2002b) et en donnent des résultats de convergence presque complète ponctuelle et uniforme. Les vitesses de convergence obtenues sont liées à la dimension fractale du processus fonctionnel. Nous pouvons remarquer que la forme de l'estimateur est la même que dans le cas des données indépendantes. Les changements surviennent du côté des hypothèses, par exemple de mélange, mais aussi dans les vitesses de convergence obtenues. En effet, sous la modélisation de la dépendance, on tient compte de la covariance des données. Ferraty et Vieu (2004) généralisent les travaux qui précèdent permettant de nombreuses applications. Ils supposent des bornes inférieures et supérieures uniformes des distributions marginales des variables explicatives. Parmi les applications possibles, ils traitent de l'estimation de la régression, de la prédiction en séries temporelles et de la discrimination de courbes. Masry (2005) établit la normalité asymptotique de l'estimateur à noyau pour des processus fortement mélangeants lorsque les variables explicatives sont fonctionnelles. Il considère une hypothèse différente à celle de Ferraty et Vieu (2004) sur les distributions marginales. Pour obtenir la normalité asymptotique de l'estimateur de la fonction de régression, la convergence en probabilité avec vitesse est également obtenue. Aspirot et al. (2009) généralisent

les résultats de Masry (2005) au cas où la variable X est un mélange non stationnaire de processus stationnaires. Les travaux de Delsol (2007a, 2009) généralisent les résultats de Ferraty et al. (2007) et de Masry (2005). En effet, la normalité asymptotique lorsque les variables explicatives sont α-mélangeantes y est établit en donnant l'expression des termes asymptotiquement dominants du biais et de la variance. Delsol (2007a,b) donnent les expressions explicites des termes dominants des moments centrés de l'estimateur ainsi que celles des erreurs en norme L^q dans le cas d'un échantillon de variables α-mélangeantes. Ferraty et al. (2010) étudient l'estimation non paramétrique de certaines fonctions de la distribution conditionnelle d'une variable réponse scalaire sur une variable à valeurs dans un espace semi-métrique. Ces fonctions incluent, entre autres, la fonction de régression, la distribution conditionnelle cumulée, la densité conditionnelle. Ils prouvent la convergence presque complète uniforme avec vitesse des estimateurs à noyau de ces modèles non paramétriques.

Un cas particulier de l'estimateur à noyau de la fonction de régression est une approche locale linéaire. Dans la situation de la régression d'une variable réelle Y sur une variable fonctionnelle X, cette approche a été traitée, par exemple, dans les travaux de Baíllo et Grané (2009), de Barrientos-Marin et al. (2010) et de Aneiros-Pérez et al. (2011).

Autres situations Jusqu'à présent, nous avons considéré le cas de la régression d'une variable réelle Y sur une variable fonctionnelle X. Il existe d'autres modèles de régression où interviennent des variables fonctionnelles, par exemple, Dabo-Niang et Rhomari (2009) étudient un estimateur non paramétrique de la régression lorsque la variable réponse est à valeurs dans un espace séparable de Banach et la variable explicative est à valeurs dans un espace séparable semi-métrique. Sous des conditions générales, des résultats asymptotiques sont établis et les bornes supérieures de l'erreur d'estimation en moyenne d'ordre q et presque sûre sont données. Les auteurs présentent également le cas où la variable explicative est un processus de Wiener. Ferraty et al. (2011) considèrent également un modèle non paramétrique de régression lorsque la variable réponse Y et la variable explicative X sont toutes les deux fonctionnelles. Les vitesses de convergence uniforme presque complète sont établies. Par la suite, Ferraty et al. (2012b) établissent la normalité asymptotique ponctuelle de l'estimateur à noyau de l'opérateur de régression lorsque les variables X et Y sont fonctionnelles. Ferraty et al. (2012a) considèrent l'estimateur à noyau de la fonction de régression lorsque la variable réponse est dans un espace de Banach et la variable explicative est à valeurs dans un espace semi-métrique. Ils établissent la vitesse de convergence presque complète de l'estimateur construit lorsque les observations sont supposées β-mélangeantes.

Nous avons cité les travaux qui traitent de la régression comme l'espérance conditionnelle de Y sachant X. La régression est aussi l'étude de toute autre quantité liée à la distribution conditionnelle de Y sachant X. Dans un contexte non paramétrique, la convergence presque complète de l'estimateur à noyau des quantiles conditionnels est établie dans Ferraty et al. (2006) lorsque les observations sont i.i.d. alors que le cas de dépendance est étudié dans Ferraty et al. (2005b). Dans les cas i.i.d. et α-mélangeant, la normalité asymptotique de cet estimateur a été étudiée par Ezzahrioui et Ould-Saïd (2008). Dans Laksaci et al. (2009), les auteurs estiment non paramétriquement le quantile conditionnel en adaptant la méthode en norme L^1 qui admet des propriétés de robustesse. Ils établissent, sous de faibles conditions, la convergence presque complète et la normalité asymptotique de l'estimateur. Ces résultats complètent ceux de Ferraty et Vieu (2006). Dabo-Niang et Laksaci (2012) considèrent une séquence fortement mélangeante de variables aléatoires $(X_i, Y_i)_{i=1,\ldots,n}$ à valeurs dans $\mathcal{E} \times \mathbb{R}$, où $\mathcal{E}$ est un espace semi-métrique.

Ils montrent la convergence en norme L^q de l'estimateur à noyau de la fonction de régression quantile de Y_i sachant X_i. Une autre quantité liée à la distribution conditionnelle est le mode conditionnel. Ferraty et al. (2005a) proposent un estimateur du mode de la distribution d'une variable réelle Y conditionnée par une variable fonctionnelle X basé sur l'estimation de la densité conditionnelle par des estimateurs à noyau. La convergence presque complète est établie sous la condition de α-mélange. Notons que Ferraty et al. (2006) traitent également de l'estimation conditionnelle du mode et étudient la convergence presque complète. Dabo-Niang et Laksaci (2007) étudient l'estimateur à noyau du mode de la distribution d'une variable réelle Y conditionnellement à une variable explicative X, à valeurs dans un espace semi-métrique. Ils établissent la convergence en norme L^q de cet estimateur. L'étude de toute autre quantité liée à la distribution conditionnelle de Y sachant X a également été traitée avec un estimateur à noyau robuste. Ainsi Azzedine et al. (2008) étudient la généralisation de la fonction de régression classique et proposent un estimateur à noyau robuste en présence de variables indépendantes et identiquement distribuées. Ils obtiennent des vitesses de convergence presque complète de l'estimateur lorsque la variable explicative est fonctionnelle et la variable réponse est réelle. Crambes et al. (2008) proposent un estimateur à noyau robuste pour l'estimation non paramétrique conditionnelle d'une variable réelle avec une covariable fonctionnelle. Ils donnent les expressions asymptotiques exactes des vitesses de convergence en norme L^q ($q < \infty$) des estimateurs robustes de la régression dans le cas où les variables fonctionnelles sont indépendantes mais aussi le cas où elles sont α-mélangeantes. Ces résultats étendent ceux de Delsol (2007b) dans lesquels les expressions asymptotiques sont obtenues pour l'estimation de l'espérance conditionnelle de Y sachant X. Dans le même contexte, Attouch et al. (2009) (respectivement Attouch et al. (2010)) établissent la normalité asymptotique de l'estimateur à noyau robuste pour données indépendantes (respectivement dépendantes).

Cas des variables spatialement dépendantes

Dans la section 1.1.1, nous avons énoncé les travaux qui traitent des estimateurs à noyau des fonctions de densité et de régression pour données spatialement dépendantes à valeurs réelles. La présente sous-section a pour objectif d'étendre cette revue de littérature au cas où les données sont spatialement dépendantes mais à valeurs dans un espace de dimension infinie. Pour ne pas faire écho à la section 1.1.1, nous ne rappelons pas ici les notions de statistique spatiale déjà énoncées. Nous nous contentons de citer les travaux qui traitent de modélisation non paramétrique spatiale pour données fonctionnelles. En particulier, dans la suite, les variables explicatives X sont à valeurs dans un espace de dimension infinie et les variables réponses Y sont à valeurs dans $\mathbb{R}$. Notons que nous citons d'abord les travaux relatifs à l'estimation de la densité spatiale. En effet, bien que l'estimation de la densité spatiale ne soit pas l'objet de cette section, elle intervient dans l'estimation de la fonction de régression.

Estimation par la méthode du noyau de la fonction de densité de probabilité spatiale lorsque les variables explicatives sont dans un espace de dimension infinie et les sites sont dans $\mathbb{Z}^N$ Dans Basse et al. (2008), sous de faibles conditions, la convergence en moyenne quadratique de l'estimateur à noyau de la fonction de densité d'un champ aléatoire stationnaire indexé dans $\mathbb{N}^N$ est obtenue mais lorsque les observations sont à valeurs dans un espace semi-métrique $(\mathcal{E}, d)$, qui peut éventuellement être de dimension infinie. Dabo-Niang et al. (2010) s'intéressent à l'estimation du mode spatial pour des champs aléatoires fonctionnels avec applications aux problèmes de bioturbation.

L'estimation du mode spatial est déduite de l'estimation de la densité spatiale de variables à valeurs dans un espace semi-métrique. L'estimateur à noyau de la densité proposé est montré être uniformément convergent. Dans la continuité, Dabo-Niang et Yao (2013) étudient un estimateur à noyau de la densité spatiale d'un champ aléatoire fonctionnel. En effet, ce dernier est à valeurs dans espace normé de dimension infinie qui admet une densité par rapport à une mesure de référence. Ils étudient les convergences faible et forte de l'estimateur considéré et donnent également ses vitesses de convergence.

Estimation par la méthode à noyau de la fonction de régression spatiale lorsque les variables explicatives sont fonctionnelles et les sites sont dans $\mathbb{Z}^N$ On se place dans le cadre où les observations $(X_{\mathbf{i}}, Y_{\mathbf{i}})_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}$ obéissent au modèle de régression $r(x) = \mathbb{E}[Y_{\mathbf{i}}|X_{\mathbf{i}} = x]$, c'est à dire basé sur l'espérance conditionnelle de Y sachant X. Attouch et al. (2011) s'attachent à étendre l'un des résultats de Gheriballah et al. (2010) mais lorsque les covariables sont de nature fonctionnelles. En effet, cette contribution concerne l'estimation robuste de $r(\cdot)$ quand les variables explicatives sont des champs aléatoires fonctionnels. Pour cela, les auteurs proposent une famille d'estimateurs non paramétriques robustes basés sur la méthode à noyau. Ils montrent la convergence presque complète de ces estimateurs et en donnent les vitesses de convergence. Dabo-Niang et al. (2011b) étudient un estimateur à noyau de l'espérance conditionnelle d'une variable réponse réelle Y sachant un champ aléatoire fonctionnel X à valeurs dans un espace semi-métrique. Les convergences faible et forte de l'estimateur sont montrées et des vitesses de convergence presque sûre sont données.

Régressions basées sur d'autres caractéristiques conditionnelles Dans un premier temps, on s'intéresse aux travaux où la régression est basée sur l'estimation des quantiles conditionnels. Laksaci et Maref (2009) considèrent un champ aléatoire fonctionnel, stationnaire $(Z_{\mathbf{i}} = (X_{\mathbf{i}}, Y_{\mathbf{i}}), \mathbf{i} \in \mathbb{N}^N, N > 0)$ à valeurs dans $\mathcal{F} \times \mathbb{R}$, où $\mathcal{F}$ est un espace semi-métrique, de dimension éventuellement infinie. Ils étudient la covariation spatiale des deux variables $X_{\mathbf{i}}$ et $Y_{\mathbf{i}}$ via l'estimation non paramétrique des quantiles conditionnels de $Y_{\mathbf{i}}$ sachant $X_{\mathbf{i}}$. Un estimateur à noyau est proposé pour ce modèle pour lequel ils établissent une vitesse de convergence presque complète. Dans leur Note, Dabo-Niang et al. (2011a) considèrent un champ aléatoire $(Z_{\mathbf{i}} = (X_{\mathbf{i}}, Y_{\mathbf{i}}), \mathbf{i} \in \mathbb{N}^N, N > 1)$ à valeurs dans $\mathcal{F} \times \mathbb{R}$ de même loi que (X, Y), où $(\mathcal{F}, d)$ est un espace semi-métrique. Ils établissent la convergence en moyenne d'ordre q de l'estimateur à noyau proposé du quantile conditionnel de $Y_{\mathbf{i}}$ sachant $X_{\mathbf{i}}$. Etant donné un processus spatial strictement stationnaire, Dabo-Niang et al. (2012b) prouvent la convergence avec vitesse en norme L^q ainsi que la normalité asymptotique d'un estimateur de la fonction de quantile conditionnel spatial d'une variable réponse univariée $Y_{\mathbf{i}}$ sachant une variable fonctionnelle $X_{\mathbf{i}}$. Dans un second temps, nous présentons les travaux où la régression est basée sur l'estimation du mode conditionnel. Dabo-Niang et al. (2012a) considèrent un processus spatial strictement stationnaire $(Z_{\mathbf{i}} = (X_{\mathbf{i}}, Y_{\mathbf{i}}), \mathbf{i} \in \mathbb{Z}^N, N > 1)$ à valeurs dans $\mathcal{F} \times \mathbb{R}$, où $\mathcal{F}$ est un espace semi-métrique. Ils étudient un estimateur à noyau du mode conditionnel d'une variable réponse réelle $Y_{\mathbf{i}}$ sachant une variable fonctionnelle $X_{\mathbf{i}}$. Le résultat principal de ce travail est la convergence presque complète avec vitesse de l'estimateur à noyau étudié.

Cette section nous a permis d'énoncer les travaux traitant de l'estimateur à noyau en présence de données fonctionnelles. Nous avons d'abord abordé le cas où les variables considérées sont indépendantes puis nous avons parlé du cas des variables dépendantes et plus particulièrement α-mélangeantes. Nous avons distingué le cadre des variables chronologiquement dépendantes des variables spatialement dépendantes. La section suivante énonce

les motivations qui nous ont conduit à apporter notre contribution à la modélisation de données de nature fonctionnelle par les méthodes à noyau.

1.2.2 Contributions à l'analyse de données fonctionnelles

Dans la partie II de ce travail de thèse, nous avons plus particulièrement étudié des modèles de régression pour variables fonctionnelles dépendantes. Nous avons considéré le cas de la régression d'une variable réelle Y sur une variable fonctionnelle X.

Tout d'abord, dans le chapitre 4, nous proposons un estimateur de la fonction de régression (espérance conditionnelle) lorsque les variables étudiées présentent de la dépendance spatiale. L'estimateur proposé dans ce chapitre 4 est une adaptation au cadre des données fonctionnelles du modèle proposé dans le chapitre 3 lorsque les variables explicatives étaient multivariées. La particularité de cet estimateur est de tenir compte à la fois des courbes observées mais aussi de la position des sites où ont lieu les observations. Sous des conditions classiques, nous avons établi la convergence en moyenne quadratique avec vitesse de l'estimateur à noyau étudié. Puis, nous avons proposé un exemple d'application de cet estimateur à la prédiction spatiale. Enfin, le comportement pratique de l'estimateur est étudié au travers d'une étude de simulation.

Après avoir étudié dans le chapitre 4 des données fonctionnelles spatialement dépendantes, nous avons considéré dans le chapitre 5 l'étude non paramétrique de séries temporelles de nature fonctionnelle. On s'intéresse à un modèle de régression pour variables dépendantes dans le temps lorsque la variable réponse est réelle, la variable explicative est fonctionnelle et le terme d'erreur présente de l'autocorrélation. Nous introduisons une procédure basée sur l'estimateur à noyau permettant de tenir compte de l'information contenue dans le terme d'erreur. En effet, une transformation du modèle de régression initial permet d'obtenir un modèle de régression dont le terme d'erreur n'est plus corrélé tout en conservant l'information contenue initialement. Ensuite nous étudions les propriétés asymptotiques de l'estimateur issu de la procédure de transformation. Sous des conditions classiques, la normalité asymptotique est obtenue. Nous proposons ensuite une étude de simulation qui montre l'efficacité de notre procédure lorsque le terme d'erreur présente une corrélation importante. Nous avons également appliqué la procédure sur des données réelles de concentration en ozone. On peut remarquer que les résultats obtenus permettent d'améliorer ceux obtenus avec l'estimateur à noyau classique.

Dans chacun de ces deux chapitres (4 et 5), nous proposons et étudions le comportement asymptotique d'estimateurs de la fonction de régression d'une variable réelle Y sur une variable fonctionnelle X, basés sur la méthode à noyau. Puis, dans la partie III de ce manuscrit, nous avons appliqué des méthodes de classification non-supervisée (Chapitre 7) pour données fonctionnelles sur les débits de rivières Québécoises. Parmi les méthodes considérées, l'une est basée sur l'estimation de la densité et de son mode. Les classifications obtenues sur ces données nous ont conduit à appliquer également des méthodes de détection de rupture (Chapitre 8) sur ces données fonctionnelles.

Dans ce premier chapitre, nous avons introduit les cadres statistiques que nous avons considérés dans les parties I, II et III de ce mémoire de thèse. Les deux premières parties de ce manuscrit concernent plutôt des résultats théoriques alors que la partie III se veut être plus appliquée. La partie I est consacrée à la modélisation non paramétrique de données spatiales pendant que la partie II traite essentiellement de l'estimation non paramétrique de la fonction de régression d'une variable réelle sur une variable fonctionnelle. Ces introductions ont permis d'établir un état de l'art des travaux existants dans chacune des thématiques considérées afin de mieux comprendre et situer l'apport de ce travail de thèse.

Part I

Nonparametric statistics for spatial data

Chapter 2

A kernel spatial density estimation allowing for the analysis of spatial clustering

Contents

Résumé en français

Dans ce chapitre, on s'intéresse à l'estimation à noyau de la fonction de densité f d'un processus spatial $X = \{X_{\mathbf{i}}\}$, strictement stationnaire, à valeurs dans $\mathbb{R}^d$ et indexé sur un ensemble spatial discret $\mathcal{I}_{\mathbf{n}} = \{\mathbf{i} = (i_1, \ldots, i_N), 1 \leq i_k \leq n_k,\ k = 1, \ldots, N\}$ où un point $\mathbf{i} = (i_1, \ldots, i_N) \in \mathbb{Z}^N$ fait référence à un site. Nous supposons également que le processus étudié satisfait la condition de mélange fort (α-mélange). L'estimateur à noyau classique de la densité spatiale (e.g. Tran (1990), Carbon et al. (1997), El Machkouri (2011)) est défini par

$$f_{\mathbf{n}}^{(0)}(x) = \left(\widehat{\mathbf{n}} b_{\mathbf{n}}^d\right)^{-1} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K\left(\frac{x - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right), x \in \mathbb{R}^d, \tag{2.1}$$

où K est un noyau, $(b_{\mathbf{n}})$ est une séquence de fenêtres qui tend vers zéro lorsque $\mathbf{n}$ tend vers l'infini et $\widehat{\mathbf{n}} = n_1 \times \ldots \times n_N$ fait référence à la taille de l'échantillon. Notre objectif est de proposer une nouvelle version de l'estimateur (2.1) qui tient compte à la fois de la valeur des observations mais aussi de la position des sites où ont lieu les observations. En effet, on s'attend à ce que les observations proches du site étudié aient plus d'influence que les observations faites en des sites plus éloignés. On admet, par souci de simplicité,

que $n_1 = n_2 = \ldots = n_N = n$ (e.g. El Machkouri (2007, 2011), El Machkouri and Stoica (2010)), mais les résultats suivants peuvent être étendus à un cadre plus général. Pour chaque $x_{\mathbf{j}} \in \mathbb{R}^d$ fixée et localisée en $\mathbf{j} \in \mathcal{I}_{\mathbf{n}}$, l'estimateur de f que nous proposons dans ce chapitre est défini de la façon suivante

$$f_{\mathbf{n}}(x_{\mathbf{j}}) = \frac{1}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i} - \mathbf{j}\|), \tag{2.2}$$

où $K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|) = C_{\mathbf{nj}} K_2\left(\frac{\|t_{\mathbf{i}}-t_{\mathbf{j}}\|}{\rho_{\mathbf{n}}}\right)$ avec $t_{\mathbf{i}} = \frac{\mathbf{i}}{\mathbf{n}} =: \left(\frac{i_1}{n}, \ldots, \frac{i_N}{n}\right)$, $C_{\mathbf{nj}} > 0$ est une constante de normalisation, K_1 et K_2 sont des noyaux respectivement définis sur $\mathbb{R}^d$ et $\mathbb{R}$. Nous supposons que les séquences $(b_{\mathbf{n}})$ et $(\rho_{\mathbf{n}})$ tendent vers zéro de sorte que $\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N \to +\infty$. Pour chaque site $\mathbf{j}$, on pose $k_{\mathbf{n}} = k_{\mathbf{n},\mathbf{j}} = \sum_{\mathbf{i}} 1_{[\|\mathbf{i}-\mathbf{j}\| \leq d_{\mathbf{n}}]}$ pour dénoter le nombre de voisins $\mathbf{i}$ pour lesquels la distance entre $\mathbf{i}$ et $\mathbf{j}$ est inférieure ou égale à la distance $d_{\mathbf{n}} > 0$ telle que $d_{\mathbf{n}} \to \infty$ lorsque $\mathbf{n} \to \infty$. L'estimateur $f_{\mathbf{n}}(x_{\mathbf{j}})$ est une fonction du nombre $k_{\mathbf{n}}$ avec $k_{\mathbf{n}} \to +\infty$.

Plus généralement, soit $X_{\mathbf{i}}$, $\mathbf{i} \in \mathcal{I}_{\mathbf{n}}$, un processus spatial strictement stationnaire et soit $\mathbf{i_0}$ un site n'appartenant pas à $\mathcal{I}_{\mathbf{n}}$, on peut étendre l'estimateur (2.2) de la manière suivante

$$\widehat{f}_{\mathbf{n}}(x_{\mathbf{i_0}}) = \frac{1}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K_1\left(\frac{x_{\mathbf{i_0}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_2\left(\frac{\mathbf{i} - \mathbf{i_0}}{\rho_{\mathbf{n}}}\right)$$

où les sites $\mathbf{i}$ et $\mathbf{i_0}$ ne sont pas normalisés (voir e.g. Wang et al. (2012)), K_1 et K_2 sont deux noyaux sur $\mathbb{R}^d$ et $\mathbb{R}^N$ respectivement.

Dans la suite, après avoir introduit le contexte dans lequel nous travaillons, nous définissons le nouvel estimateur à noyau de la fonction de densité spatiale f qui a été brièvement introduit précédemment. Sous des conditions classiques, nous étudions le comportement asymptotique de l'estimateur proposé, notamment la convergence presque sûre (p.s.) avec vitesse. La convergence uniforme est également montrée. En particulier, on montre que

$$|f_{\mathbf{n}}(x_{\mathbf{i}}) - f(x_{\mathbf{i}})| = O\left(b_{\mathbf{n}} + \sqrt{\frac{\log \widehat{\mathbf{n}}}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}}\right) \qquad \text{p.s.}$$

et

$$\sup_{\substack{x_{\mathbf{i}} \in R \\ \mathbf{i} \in V_R}} |f_{\mathbf{n}}(x_{\mathbf{i}}) - f(x_{\mathbf{i}})| = O\left(b_{\mathbf{n}} + \sqrt{\frac{\log \widehat{\mathbf{n}}}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}}\right) \qquad \text{p.s.}$$

où R est un ensemble compact de $\mathbb{R}^d$ et $V_R \subseteq \mathbb{R}^N$ est l'ensemble fini des sites $\mathbf{j}$ contenus dans $\mathcal{I}_{\mathbf{n}}$ tels que les $x_{\mathbf{j}}$ correspondants soient dans R. Ensuite, nous proposons d'utiliser cet estimateur de la densité pour estimer le(s) mode(s), noté ω, de la distribution spatiale correspondante. Cet estimateur du mode est défini par

$$\widehat{\omega} = \arg \sup_{\substack{x_{\mathbf{i}} \in R \\ \mathbf{i} \in V_R}} f_{\mathbf{n}}(x_{\mathbf{i}}).$$

pour lequel on montre que

$$\|\widehat{\omega} - \omega\| = O\left(b_{\mathbf{n}} + \sqrt{\frac{\log \widehat{\mathbf{n}}}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}}\right) \qquad \text{p.s.}$$

Enfin, une procédure, basée sur l'estimation du mode spatial, pour résoudre un problème de classification non supervisée spatiale est présentée. Il s'agit d'une méthode de classification descendante hiérarchique basée sur le calcul d'indices d'hétérogénéité.

Les résultats présentés dans ce chapitre ont été obtenus avec la collaboration de Sophie Dabo-Niang (Université Charles de Gaulle), Leila Hamdad (Ecole Nationale Supérieure d'Informatique, Algérie) et Anne-Françoise Yao (Université Blaise Pascal) et font l'objet d'une publication (Dabo-Niang et al. (2014a)) dans *Stochastic Environmental Research and Risk Assessment*.

2.1 Introduction

In many fields such as epidemiology, environmental science, image analysis and many others, one often deals with spatially dependent data. The study of the distribution or any characteristic of such data cannot be done without taking into account their respective geographical positions. Therefore, modeling spatial dependency in statistical inferences (estimation of spatial distribution, prediction, etc.) is a significant feature of spatial data analysis. Spatial statistics have provided tools to solve such problems, particularly within the scope of geostatistics. So far, most spatial modeling methods are parametric. For instance, estimation of the density function of a spatial process can be addressed in a parametric way by assuming an analytical expression. However, such an assumption is not always reasonable; therefore a nonparametric approach must be adopted instead. Among existing parametric geostatistical methods, variogram analysis and kriging are respectively useful tools to measure spatial dependence and achieve spatial prediction. For some backgrounds in parametric spatial statistic modeling, we refer to Ripley (1981), Cressie (1993), Anselin and Florax (1995), Guyon (1995), Stein (1999), Wackernagel (2003), Cressie and Wikle (2011) and the references therein. Nonparametric spatial modeling is much less extensive than parametric. Some works have been done to study nonparametric variogram, density or regression problems for spatial data. We refer, for example, to Tran (1990), Tran and Yakowitz (1993), Biau (2003), Carbon et al. (1997), Carbon (2006), Hallin et al. (2004a), Carbon et al. (2007), Menezes et al. (2010), El Machkouri (2011), Dabo-Niang and Yao (2007), Dabo-Niang and Yao (2013), Dabo-Niang et al. (2011b), Wang et al. (2012), García-Soidán and Menezes (2012) and the references therein.

This work deals with a new version of the kernel density estimation in the spatial setting. To the best of our knowledge, the asymptotic and applied results of a nonparametric density estimate, which incorporates the spatial dependency, have not been studied in the literature. Some recent works gave results concerning spatial kernel weight function but in a different context (variogram estimation or prediction) from the one considered in the current work (density estimation). García-Soidán and Menezes (2012) considered a spatial distribution estimator, through the kernel indicator variogram, but for randomly distributed spatial sites. Wang et al. (2012) proposed a local linear spatio-temporal prediction model, using a kernel weight function taking into account the distance between sites. They gave an asymptotic result and studied some applications to simulated and real data. Dabo-Niang et al. (2010) (see also Dabo-Niang and Yao (2013)) studied a kernel density estimator of spatial functional (and multivariate) data. This estimator does not directly take into account the spatial dependency in the form of the estimator, but the authors provided some discussions on how this can be done by introducing a second kernel, based on the distance between sites. They considered an asymptotic estimator and a real data application to detect heterogeneity in some environmental spatial data. This current work is then a natural extension of the works in Dabo-Niang et al. (2010) and Dabo-Niang and

Yao (2013).

In fact, the aim of this chapter is the study of the spatial density estimation of the margins $X_{\mathbf{i}}$, of a discretely indexed spatial process (called *random field*), $(X_{\mathbf{i}}, \mathbf{i} \in \mathbb{Z}^N)$ where $\mathbb{Z}^N$ is the integer lattice points in the space $\mathbb{R}^N$, $N \geq 1$. We recall that, in spatial statistics, the indices are called *sites* and they often represent geographical positions. They will be written in bold as $\mathbf{i} = (i_1, \ldots, i_N)$. Here, we are interested in the case where the process $(X_{\mathbf{i}})$ is assumed to be a *strictly stationary random field* with values in $\mathbb{R}^d$, $d \geq 1$ and defined over some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We suppose that the $X_{\mathbf{i}}$'s have the same distribution as X admitting an unknown density f. Strict stationarity is a common assumption in nonparametric density or regression estimation for spatial or non-spatial data (see the references below). This assumption can be unreasonable, for instance, when the distribution of $X_{\mathbf{i}}$ is changing sufficiently slowly as a function of $\mathbf{i}$. It can be relaxed by assuming that the $X_{\mathbf{i}}$'s are non-identically distributed. Suppose that there exists a collection of density functions $\{f_{\mathbf{i}}, \mathbf{i} \in \mathbb{Z}^N\}$ of $\{X_{\mathbf{i}}, \mathbf{i} \in \mathbb{Z}^N\}$ and we want to estimate $f_{\mathbf{i}_0}$ at a fixed $\mathbf{i}_0$. If $f_{\mathbf{i}}$ is close to $f_{\mathbf{i}_0}$ when $\mathbf{i}$ is close to $\mathbf{i}_0$ and if there are enough such sites named $\{\mathbf{i}_1, \ldots, \mathbf{i}_M\}$ that are close to $\mathbf{i}_0$, then the variables $\{X_{\mathbf{i}_1}, \ldots, X_{\mathbf{i}_M}\}$ can be used to estimate $f_{\mathbf{i}_0}$. More generally, this can be formulated as *locally identically distributed* or *locally stationary* assuming that there exists a distribution F such that a sufficient number of random variables, in the sequence $X_{\mathbf{i}}$, have a distribution close to F (see, for example, Klemelä (2008)). This means that for neighbors sites $\mathbf{j}$ and $\mathbf{k}$, $f_{\mathbf{j}}$ and $f_{\mathbf{k}}$ are close together.

As usual, the spatial density estimation is based on some observations within a set of sites. In the following, as it is classically done in nonparametric modeling (see, for example, Tran (1990); Biau (2003); Dabo-Niang and Yao (2007); Carbon et al. (1997); Wang et al. (2012), ...), we will assume, without loss of generality, that the set of sites of observations is a rectangular region, defined by $\mathcal{I}_{\mathbf{n}} := \{\mathbf{i} \in (\mathbb{N}^*)^N, 1 \leq i_k \leq n_k, k = 1, \ldots, N\}$. Let us recall that, as in any nonparametric spatial density model, the methods proposed here remain valid if $\mathcal{I}_{\mathbf{n}}$ has a general form (see, for example, El Machkouri (2011)). Let $\widehat{\mathbf{n}} := n_1 \ldots n_N$ be the sample size. Naturally, one can also enumerate, in a particular order, the sites and identify $\mathcal{I}_{\mathbf{n}}$ with the set $\{s_i \in \mathbb{N}^N, i = 1, \ldots, \widehat{\mathbf{n}}\}$. In the case of a rectangular region, considered here, one can also use, for instance, a triangular array structure to obtain $\{s_i \in \mathbb{N}^N, i = 1, \ldots, \widehat{\mathbf{n}}\}$ using a specific order; see Robinson (2011). Before going further, we recall that classically the *spatial kernel density* estimator for f considered in the literature (see Tran (1990); Carbon et al. (1997)) is

$$f_{\mathbf{n}}^{(0)}(x) = \left(\widehat{\mathbf{n}} b_{\mathbf{n}}^d\right)^{-1} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K\left(\frac{x - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right), x \in \mathbb{R}^d, \tag{2.3}$$

where K is a kernel and $(b_{\mathbf{n}})$ is a sequence of bandwidths that tends to zero as $\mathbf{n}$ goes to infinity. The asymptotic behavior of this estimator has been studied in both discrete and continuous domains by several authors such as Tran (1990), Biau (2003), Carbon et al. (1997), Hallin et al. (2004a), Dabo-Niang et al. (2010) whereas the practical use has been tackled only by a few authors (see e.g. Jiang et al. (2007); Davies et al. (2011); Dabo-Niang and Yao (2013)). It is easy to see that the estimator (2.3) follows the strict stationarity condition since $f_{\mathbf{n}}^{(0)}(x_{\mathbf{k}}) = f_{\mathbf{n}}^{(0)}(x_{\mathbf{j}})$ for all $x_{\mathbf{k}} = x_{\mathbf{j}}$ even if $\mathbf{k} \neq \mathbf{j}$. However, it does not take into account the spatial dependence. The aim of this work is to propose a modified version of the kernel estimator (2.3) that considers the spatial dependence in its expression. Theoretically, we will measure the spatial dependency in strong mixing meaning. We recall this notion in the next section.

Clearly, the kernel density estimator (2.3) has the same form either with independent or dependent observations, and asymptotically it behaves differently according to the case.

However, a previous work of Dabo-Niang et al. (2010) shows that even if theoretically the estimate (2.3) perfectly verifies the stationarity condition, it is less efficient in practice than a kernel density estimator incorporating the spatial dependence. It is a common problem in spatial kernel estimation (density or regression). We refer to the works of Dabo-Niang and Yao (2013) and Dabo-Niang et al. (2010, 2011b) for more details. In the works quoted above, they solved the problem by presenting a practical version of the kernel estimator in question, which takes into account the spatial dependence. So in practice, one has to deduce a practical version of (2.3), which includes the spatial dependence. To avoid this double level of estimation (theoretical and practical versions), we propose here an estimator that is able to both consider the spatial dependence and have similar asymptotic behavior as (2.3). This approach raises some questions in particular: how does such an estimator behave under the stationarity condition? These questions will be discussed afterward.

To understand the motivation of our approach, let us recall that $(X_{\mathbf{i}}, \mathbf{i} \in \mathbb{Z}^N)$ is supposed to be strongly mixing, which means that the farther a site $\mathbf{i}$ is from $\mathbf{j}$, the lower the dependence between $X_{\mathbf{i}}$ and $X_{\mathbf{j}}$. Thus, $X_{\mathbf{i}}$ and $X_{\mathbf{j}}$ are nearly independent when $\|\mathbf{i}-\mathbf{j}\|$ is high. In other words, one can consider that any $X_{\mathbf{i}}$ depends only on $X_{\mathbf{j}}$s located at some neighborhood of the site $\mathbf{i}$. So, for any $X_{\mathbf{i}}$ there exists a neighborhood denoted $\mathcal{N}_{\mathbf{i}}$, such that all other observations located in $\mathcal{N}_{\mathbf{i}}$ bring enough information about the distribution of $X_{\mathbf{i}}$. In particular, these observations located in $\mathcal{N}_{\mathbf{i}}$ should suffice to estimate the density. Let us notice that a similar approach based on a kernel that controls the distance between sites has been developed, for example, in spatial prediction by Menezes et al. (2010) or Wang et al. (2012). More precisely, in Menezes et al. (2010), a nonparametric kernel predictor, based on a kernel that controls the distance between sites, is considered for spatial stochastic processes when a stochastic sampling design is assumed for selection of random locations. In the work of Wang et al. (2012), a local linear spatio-temporal prediction model, using a kernel weight function taking into account the distance between sites, is proposed. The specificity of the prediction procedure of Wang et al. (2012) is to be based on the assumption that the error term of the model is autocorrelated.

This chapter is organized as follows. In Section 2.2, we tackle the theoretical framework of the problem where we describe the new kernel density estimator and its application in the spatial modes estimation. We also focus on the strong rate of convergence under stationarity condition. In Section 2.3, we present some algorithms that permit, on one hand, to use the estimator in practice and, on the other hand, to perform a clustering method. This can be useful to compare the spatial structures detected by the density estimation and clustering method. The clustering method proposed is a particular case of the top-down hierarchical method first introduced by Dabo-Niang et al. (2004) (see also Ferraty and Vieu (2006)) for functional independent data. Dabo-Niang et al. (2010) propose a functional spatial version of this method. Let us notice that this clustering method, described in Section 2.3.2, is based on local spatial modes detection. The proofs of the theoretical asymptotic behavior of the estimator are given in the Appendix (see Section 2.5).

2.2 The theoretical framework

In the following, $\|\cdot\|$ will denote any norm in $\mathbb{R}^d$ or $\mathbb{R}^N$ (there will be no ambiguity since the vectors of $\mathbb{R}^N$ are in bold), $C > 0$ will indicate a constant, and, for each real u, $\lfloor u \rfloor$ will indicate the integer part of u.

2.2.1 Notations and assumptions

To take into account the spatial dependency, we assume that the stationary process $(X_{\mathbf{i}})$ satisfies a mixing condition defined in Carbon et al. (1997) as follows: there exists a function $\varphi(x) \searrow 0$ as $x \to \infty$, such that

$$\begin{aligned}\alpha\left(\sigma\left(S\right),\sigma\left(S'\right)\right) &= \sup\left\{\left|\mathbb{P}(A\cap B)-\mathbb{P}\left(A\right)\mathbb{P}\left(B\right)\right|,\quad A\in\sigma\left(S\right),B\in\sigma\left(S'\right)\right\}\\ &\leq \psi\left(\mathrm{Card}(S),\mathrm{Card}(S')\right)\varphi\left(\mathrm{dist}(S,S')\right),\end{aligned}$$

where S and S' are two finite sets of sites, $\sigma\left(S\right)=\{X_{\mathbf{i}},\mathbf{i}\in S\}$ and $\sigma\left(S'\right)=\{X_{\mathbf{i}},\mathbf{i}\in S'\}$ are σ-fields generated by $X_{\mathbf{i}}$, $\mathrm{dist}(S,S')$ is the Euclidean distance between S and S', and $\psi(\cdot)$ is a positive symmetric function nondecreasing in each variable. We recall that the process $(X_{\mathbf{i}})$ is said to be strongly mixing if $\psi\equiv 1$. As usual, we will assume that one of both conditions on $\varphi(i)$ is verified:

$$\varphi\left(i\right)\leq Ci^{-\theta},\text{for some }\theta>0,$$

i.e. that $\varphi\left(i\right)$ tends to zero at a polynomial rate, or

$$\varphi\left(i\right)\leq C\exp(-si),\text{for some }s>0,$$

i.e. that $\varphi\left(i\right)$ tends to zero at an exponential rate.

To keep the analogy with the case $N=1$, in nonparametric spatial modeling, one often supposes, without loss of generality, that the sample size is the rectangular region $\mathcal{I}_{\mathbf{n}}$. But the results remain valid if $\mathcal{I}_{\mathbf{n}}$ is replaced by any subset of a large family of lattices of $\mathbb{R}^N$. From now on, we will write $\mathbf{n}\to\infty$ if $\min_{k=1,\ldots,N} n_k\to\infty$ and for all $1\leq j,k\leq N$, $|\frac{n_j}{n_k}|<C$, for some constant $0<C<\infty$. This means that the number of observations on the rectangular region expands to infinity at the same rate along all directions.

2.2.2 The new spatial kernel density estimator

From now on, we assume for simplicity that $n_1=n_2=\ldots=n_N=n$ (e.g. El Machkouri (2007, 2011), El Machkouri and Stoica (2010)), but the following results can be extended to a more general framework. For each site $\mathbf{j}$, let $k_{\mathbf{n}}=k_{\mathbf{n},\mathbf{j}}=\sum_{\mathbf{i}}1_{[\|\mathbf{i}-\mathbf{j}\|\leq d_{\mathbf{n}}]}$ denote the number of neighbors $\mathbf{i}$ for which the distance between $\mathbf{i}$ and $\mathbf{j}$ is less than or equal to distance $d_{\mathbf{n}}>0$ such that $d_{\mathbf{n}}\to\infty$ as $\mathbf{n}\to\infty$. Taking the Euclidean distance and if $N=2$ (square grid), we have $k_{\mathbf{n}}\leq 4d_{\mathbf{n}}^2-4d_{\mathbf{n}}+4$ which leads to $k_{\mathbf{n}}=O(d_{\mathbf{n}}^2)$ and $k_{\mathbf{n}}=o(d_{\mathbf{n}}^{\eta}),\eta>2$. Let $\rho_{\mathbf{n}}>0$ and $d_{\mathbf{n}}=n\rho_{\mathbf{n}}$; consequently, we have $d_{\mathbf{n}}^2=\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N$ and $k_{\mathbf{n}}=O(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N)$ as well as $k_{\mathbf{n}}=o((\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N)^{\eta/2})$, $\eta>2$. Using this, we construct a spatial kernel estimator with weight on the sites. The weights are assumed to decline as a measure of distance between corresponding sites (that are normalized) increases. Let $\{X_{\mathbf{i}},\mathbf{i}\in\mathbb{Z}^N\}$ be a spatial process observed at any $\mathbf{i}\in\mathcal{I}_{\mathbf{n}}$, $X_{\mathbf{i}}\in\mathbb{R}^d$. For each fixed observation $x_{\mathbf{j}}\in\mathbb{R}^d$ located at $\mathbf{j}\in\mathcal{I}_{\mathbf{n}}$, the new kernel density estimator of f is defined by

$$f_{\mathbf{n}}(x_{\mathbf{j}})=\frac{1}{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}K_1\left(\frac{x_{\mathbf{j}}-X_{\mathbf{i}}}{b_{\mathbf{n}}}\right)K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|),\tag{2.4}$$

where $K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|)=C_{\mathbf{nj}}K_2\left(\frac{\|t_{\mathbf{i}}-t_{\mathbf{j}}\|}{\rho_{\mathbf{n}}}\right)$ where $t_{\mathbf{i}}=\frac{\mathbf{i}}{\mathbf{n}}=:\left(\frac{i_1}{n},\ldots,\frac{i_N}{n}\right)$, $C_{\mathbf{nj}}>0$ is a normalized constant eventually equal to one, K_1 and K_2 are kernels respectively defined on $\mathbb{R}^d$ and $\mathbb{R}$. In the following, we will suppose that the sequences $(b_{\mathbf{n}})$ and $(\rho_{\mathbf{n}})$ tend to zero such that $\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N\to+\infty$. The estimator $f_{\mathbf{n}}(x_{\mathbf{j}})$ is a function of the number $k_{\mathbf{n}}$ for which

the distance $d_{\mathbf{n}}$ is chosen hereafter to be $n\rho_{\mathbf{n}}$, with $k_{\mathbf{n}} \to +\infty$, $k_{\mathbf{n}} = O(d_{\mathbf{n}}^N) = O(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N)$. If one assumes that $d_{\mathbf{n}} = o(\widehat{\mathbf{n}}^{\epsilon})$, $\epsilon \in (0,1)$ then $k_{\mathbf{n}}$ can be expressed in terms of $\widehat{\mathbf{n}}$. More precisely, in what follows, we assume that $k_{\mathbf{n}} = C_N d_{\mathbf{n}}^N + O(d_{\mathbf{n}}^{\beta})$ as $d_{\mathbf{n}} \to +\infty$, $0 < \beta < N$ and C_N is a constant that depends on N. We will formulate by $V_{\mathbf{j}} = \{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}; \|t_{\mathbf{i}} - t_{\mathbf{j}}\| \leq \rho_{\mathbf{n}}\}$ and then $k_{\mathbf{n}} = k_{\mathbf{nj}} = \text{Card}(V_{\mathbf{j}})$.

Let us notice that our approach is based on some methods used in kernel smoothing for time series. In fact, in the time series setting, the locations are often supposed to be: $i = 1, \ldots, n$. Then any $i \neq j$, $|i-j| > 1$ and smoothing along the time axis lead to consider a time modeling version of (2.4) with $K_2\left(\frac{|i-j|}{\varpi_n}\right)$ and bandwidth $\varpi_n \geq 1$ (see, for example, Fan and Yao (2003)). We recall that the role of the bandwidth is the control of the number of observations in the neighborhood of i and should be such as $\frac{\varpi_n}{n} \to 0$ as mentioned in Wang et al. (2012). Other authors such as Hall et al. (2006) defined a time-dynamic version of (2.4) such that $K_2\left(\frac{|i-j|}{\varpi_n}\right) = K_2\left(\frac{|t_i - t_j|}{\rho_n}\right)$ where $t_i = \frac{i}{n}$ and $\rho_n \to 0$. All these approaches motivate our method, which is a generalization of the case $N = 1$. Moreover, we note that the kind of term $K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i} - \mathbf{j}\|)$, which controls the spatial dependency, has also been used, for example, by Wang et al. (2012) in the prediction of spatio-temporal models.

Remark 2.1.

- As mentioned before, the estimator (2.4) takes into account the spatial dependence since only observations on a vicinity of $\mathbf{j}$ will be used to estimate $f(x_{\mathbf{j}})$. However, a main question arises: how can it express this condition in practice? In fact, for distinct sites $\mathbf{k}$ and $\mathbf{j}$, such that $x_{\mathbf{k}} = x_{\mathbf{j}} = a$, since they do not share the same sample of observations, there is a very low probability of observing $f_{\mathbf{n}}(x_{\mathbf{k}}) = f_{\mathbf{n}}(x_{\mathbf{j}})$. In this case, a locally stationary condition (mentioned above) may be assumed. But if the process is really strictly stationary, then $|f_{\mathbf{n}}(x_{\mathbf{k}}) - f_{\mathbf{n}}(x_{\mathbf{j}})|$ should be small and $f(a)$ can be estimated by $f_{\mathbf{n}}(x_{\mathbf{k}})$ as well as by $f_{\mathbf{n}}(x_{\mathbf{j}})$. This is in practice a classical problem in parametric as well as nonparametric estimation even in the independent and identically distributed case: two samples of the same random variable may lead to two different estimations of the same parameter. So as in the independent and identically distributed case, we can construct some confidence intervals to control $|f_{\mathbf{n}}(x_{\mathbf{k}}) - f_{\mathbf{n}}(x_{\mathbf{j}})|$. The latter and the weakening of the strictly stationary condition are the subject of some current investigations in the context of spatial data and are beyond the scope of this work.
- Other kernel functions of the form $K_2\left(\frac{\|\mathbf{i}-\mathbf{j}\|}{\varpi_{\mathbf{n}}}\right)$ where $\mathbf{i}, \mathbf{j} \in \mathbb{N}^N$, $\varpi_{\mathbf{n}}$ depending on $\rho_{\mathbf{n}}$ and $\mathbf{n}$ may be considered. For example, $K_2\left(\frac{\|\mathbf{i}-\mathbf{j}\|}{\varpi_{\mathbf{n}}}\right) = K_2\left(\frac{\|\frac{\mathbf{i}-\mathbf{j}}{m}\|}{\rho_{\mathbf{n}}}\right) = K_2\left(\frac{\|\mathbf{i}-\mathbf{j}\|}{m\rho_{\mathbf{n}}}\right)$ with $\left(\frac{\mathbf{i}}{m}\right) = \left(\frac{i_1}{m}, \ldots, \frac{i_n}{m}\right)$ and $\|\mathbf{i}\| \leq m$; as $m = \sqrt{N}\max\{n_1, \ldots, n_N\}$ where one may assume that $m\rho_{\mathbf{n}} \to +\infty$.
- Some optimal bandwidths for $b_{\mathbf{n}}$ and $\rho_{\mathbf{n}}$ can be obtained using a cross-validation method or optimizing an entropy.
- We notice that the kernel K_2 is used to handle the nearness between locations. Then, to take into account the mixing condition, we can impose that for u large, K_2 is asymptotically:
  - ☐ *exponential*: $\|K_2(u)\| \leq C\exp(-a\,u)$, where $C, a > 0$ (such as the Gaussian kernel)
  - ☐ or *polynomial*: $\|K_2(u)\| \leq Cu^{-\theta}$, where $C, \theta > 0$.

These conditions are satisfied, for example, by several kernels with compact support such as rectangular kernels. For all kernels with compact support, these conditions are verified at least asymptotically. This is sufficient to ensure the control of the mixing condition.

- To give some examples where the assumption on $k_{\mathbf{n}}$ is reasonable, consider $q_{\mathbf{n}}$ the number of standard lattice (in $\mathbb{Z}^N$) points contained in a closed ball $B(\mathbf{j}, d_{\mathbf{n}})$ that is $q_{\mathbf{n}} = \text{Card}\{\mathbf{i} \in \mathbb{Z}^N,\ \|\mathbf{i}-\mathbf{j}\| \leq d_{\mathbf{n}}\}$ where $\mathbf{j}$ is any vector of $\mathbb{R}^N$. It is well known that

$$q_{\mathbf{n}} = \frac{\pi^{N/2}}{\Gamma(N/2+1)} d_{\mathbf{n}}^N + O(d_{\mathbf{n}}^{N-1}),$$

where $\Gamma(\cdot)$ is the gamma function, see for instance, Mitchell (1966), Chamizo and Iwaniec (1995), Tsang (2000) and Meyer (2011). And notice that $k_{\mathbf{n}} = C_N q_{\mathbf{n}}$. In particular, if $N = 2$, $q_{\mathbf{n}} = \frac{\pi}{\Gamma(2)} d_{\mathbf{n}}^2 + O(d_{\mathbf{n}})$ or $q_{\mathbf{n}} = \frac{\pi}{\Gamma(2)} d_{\mathbf{n}}^2 + o(d_{\mathbf{n}}^{2/3})$.

- More generally, let $X_{\mathbf{i}}$, $\mathbf{i} \in \mathcal{I}_{\mathbf{n}}$, a strictly stationary spatial process and let $\mathbf{i_0}$ a site that does not belong to $\mathcal{I}_{\mathbf{n}}$, one can extend estimate (2.4) in the following way

$$\widehat{f}_{\mathbf{n}}(x_{\mathbf{i_0}}) = \frac{1}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K_1\left(\frac{x_{\mathbf{i_0}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_2\left(\frac{\mathbf{i}-\mathbf{i_0}}{\rho_{\mathbf{n}}}\right)$$

where sites $\mathbf{i}$ and $\mathbf{i_0}$ are not normalized (see e.g. Wang et al. (2012)), K_1 and K_2 are two kernels on $\mathbb{R}^d$ and $\mathbb{R}^N$ respectively.

We will now study, in the following section, the theoretical asymptotic behavior of $f_{\mathbf{n}}$. The pointwise as well as uniform almost sure consistency results are given for $f_{\mathbf{n}}$. We also consider, in the same section, an estimator of the mode derived from the density estimator by maximizing the latter over a compact set. An asymptotic behavior of this estimator is also given.

2.2.3 Asymptotic results

The consistency results of $f_{\mathbf{n}}$ are obtained under the following assumptions on f, the kernels, bandwidths and local dependence condition. Let $u_{\mathbf{n}} = \prod_{i=1}^N (\log n_i)(\log\log n_i)^{1+\epsilon}$, then $\sum_{\mathbf{n}\in\mathbb{N}^N} \frac{1}{\widehat{\mathbf{n}} u_{\mathbf{n}}} < \infty$.

H1: The density f satisfies the *Lipschitz condition*:

$$|f(x) - f(y)| \leq C\|x-y\|, \forall x, y \in \mathbb{R}^d.$$

H2: The functions $K_1(\cdot)$ and $K_2(\cdot)$ are respectively bounded integrable kernels on $\mathbb{R}^d$ and $\mathbb{R}$ such that $\int K_1(t)dt = 1$, $\int \|t\|K_1(t)dt < \infty$, $\int |K_1(t)|dt < \infty$ and satisfy some *Lipschitz conditions*. Moreover, the function $K_1(\cdot)$ is a square integrable kernel on $\mathbb{R}^d$.

In the following, we will suppose that $K_{2,\rho_{\mathbf{n}}}$ is such that for each $\mathbf{j}$,

$$\frac{1}{\widehat{\mathbf{n}} \rho_{\mathbf{n}}^N} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|) = 1.$$

H3: There exists some constants C_1 and C_2 with $0 < C_1 < C_2 < \infty$, such that

$$C_1 \mathbf{1}_{[0,1]}(t) \leq K_2(t) \leq C_2 \mathbf{1}_{[0,1]}(t).$$

H4: **Local dependence condition.**
For $\mathbf{i} \neq \mathbf{j}$, the joint probability density $f_{X_\mathbf{i},X_\mathbf{j}}$ of $(X_\mathbf{i}, X_\mathbf{j})$ exists and is bounded uniformly in $\mathbf{i}$ and $\mathbf{j}$: thus, there exists a positive constant C which verifies

$$\sup_{(u,v)\in\mathbb{R}^d\times\mathbb{R}^d} |f_{X_\mathbf{i},X_\mathbf{j}}(u,v) - f_{X_\mathbf{i}}(u) f_{X_\mathbf{j}}(v)| < C.$$

H5: $\psi(n,m) \leq C\min(n,m)$ and $\widehat{\mathbf{n}} b_\mathbf{n}^{\theta_1} \rho_\mathbf{n}^{\theta_2} \log \widehat{\mathbf{n}}^{\theta_3} u_\mathbf{n}^{\theta_4} \to \infty$ with $\theta > 4N$, and with

$$\theta_1 = \frac{d\theta}{\theta - 4N} \qquad \theta_2 = \frac{N\theta}{\theta - 4N} \qquad \theta_3 = \frac{2N-\theta}{\theta-4N} \qquad \theta_4 = \frac{-2N}{\theta-4N}.$$

H6: $\psi(n,m) \leq C(n+m+1)^{\widetilde{\beta}}$ and $\widehat{\mathbf{n}} b_\mathbf{n}^{\theta_1'} \rho_\mathbf{n}^{\theta_2'} \log \widehat{\mathbf{n}}^{\theta_3'} u_\mathbf{n}^{\theta_4'} \to \infty$ with $\theta > N(2\widetilde{\beta}+3)$, and

$$\theta_1' = \frac{d(\theta+N)}{\theta - N(3+2\widetilde{\beta})} \qquad \theta_2' = \frac{N(N+\theta)}{\theta - N(3+2\widetilde{\beta})}$$
$$\theta_3' = \frac{N-\theta}{\theta - N(3+2\widetilde{\beta})} \qquad \theta_4' = \frac{-2N}{\theta - N(3+2\widetilde{\beta})}.$$

Remark 2.2. These assumptions are classically used in spatial nonparametric modeling.

- The assumptions **H1** and **H2** allow controlling the bias of the estimator.
  - ☐ If the condition on K_1 is a classical one, the condition on $K_{2,\rho_\mathbf{n}}$ is satisfied as soon as $C_\mathbf{nj}$ is such that $C_\mathbf{nj} = \frac{\widehat{\mathbf{n}}\rho_\mathbf{n}^N}{\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} K_2\left(\rho_\mathbf{n}^{-1}\|t_\mathbf{i}-t_\mathbf{j}\|\right)}$. We notice that such $C_\mathbf{nj}$ can always be built for any K_2 verifying assumption **H3**. Moreover, we have the following inequalities, which will be useful later on in the proof part $\frac{1}{(2^N+1)C_2} \leq \frac{\widehat{\mathbf{n}}\rho_\mathbf{n}^N}{C_2 k_\mathbf{n}} \leq C_\mathbf{nj} \leq \frac{\widehat{\mathbf{n}}\rho_\mathbf{n}^N}{C_1 k_\mathbf{n}} \leq \frac{2}{C_1}$.
  - ☐ Instead of normalization, a condition as the following, given in Hall et al. (2006), can be considered to get the result. This condition is: "$b_\mathbf{n}$ and $\rho_\mathbf{n}$ are the bandwidths tending to zero such that $\hat{\mathbf{n}}^{1-\epsilon} b_\mathbf{n}^d \rho_\mathbf{n}^N \to +\infty$ with $0 < \epsilon < 1$ and $k_\mathbf{n} = o(\hat{\mathbf{n}}^\epsilon)$".
- We will use assumption **H3** to control both the bias and distances between sites. This condition is verified, for example, if K_2 is defined by $K_2(t) = \mathbf{1}_{[0,1]}(t)$ or any function defined as $K_2(t) = u(t)\mathbf{1}_{[0,1]}(t)$ where u is a non-increasing function such that $u(1) > 0$ as for instance $K_2(t) = C\exp(-a\,t)\mathbf{1}_{[0,1]}(t)$ (with $C > 0$ and $a > 0$), $K_2(t) = (\frac{3}{2} - t^2)\mathbf{1}_{[0,1]}(t)$ and so on.
- The **local dependence condition (H4)** is a classical condition in kernel estimation based on dependent data (see e.g. Bosq (1998); Carbon et al. (1997)). It is necessary to control the dependence between the marginals of any couple $(X_\mathbf{i}, X_\mathbf{j})$. The difference between this condition and the mixing condition is: Condition **H4** controls the dependency through the distance between $f_{X_\mathbf{i}X_\mathbf{j}}$ and $f_{X_\mathbf{i}} f_{X_\mathbf{j}}$ when the mixing condition controls the dependency through the distance between $\mathbb{P}(A\cap B)$ and $\mathbb{P}(A)\,\mathbb{P}(B)$ (as previously defined). Naturally, both conditions are linked. The link between them can be found, for example, in Dabo-Niang and Yao (2007) or Bosq (1998). Like the mixing condition, condition **H4** is used to control the variance term of the estimation.

- The assumptions **H5** and **H6** are classical technical assumptions, which appear (in the calculations when studying the asymptotic behavior of the estimator) in the particular case where the mixing coefficient is such that $\varphi(i)$ verifies: $\varphi(i) \leq Ci^{-\theta}$, for some $\theta > 0$.

An almost sure consistency result is given in the following theorem.

Theorem 2.3. Under assumptions **H1-H4**, **H5** or **H6**, we have

$$|f_{\mathbf{n}}(x_{\mathbf{i}}) - f(x_{\mathbf{i}})| = O\left(b_{\mathbf{n}} + \sqrt{\frac{\log \widehat{\mathbf{n}}}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}}\right) \text{ a.s.}$$

The proof of this theorem is given in the Appendix.

Since we are looking for observations over a given set and are interested in an application of the density estimation namely, a clustering algorithm using the mode, let us consider a uniform consistency of the density estimate $f_{\mathbf{n}}$ over the set in question, denoted by R and supposed to be a compact. Let $V_R \subseteq \mathbb{R}^N$ be the finite set of sites $\mathbf{j}$ contained in $\mathcal{I}_{\mathbf{n}}$ such that the corresponding $x_{\mathbf{j}}$ are in R. The mode of a distribution over a given set of observations V_R is defined by

$$\omega = \arg\sup_{\substack{x_{\mathbf{i}} \in R \\ \mathbf{i} \in V_R}} f(x_{\mathbf{i}}).$$

The mode is an important centrality parameter that allows detecting some heterogeneity on a data set (see Dabo-Niang et al. (2004); Dabo-Niang et al. (2010)). Hence, an estimation of the mode can be obtained from the density estimate as follows:

$$\widehat{\omega} = \arg\sup_{\substack{x_{\mathbf{i}} \in R \\ \mathbf{i} \in V_R}} f_{\mathbf{n}}(x_{\mathbf{i}}).$$

The following additional assumptions are needed to obtain the uniform consistency of our estimator:

H7: $\psi(n,m) \leq C \min(n,m)$ and $\widehat{\mathbf{n}} b_{\mathbf{n}}^{\theta_5} \rho_{\mathbf{n}}^{\theta_6} \log \widehat{\mathbf{n}}^{\theta_7} u_{\mathbf{n}}^{\theta_8} \to \infty$ with $\theta > N(d+4)$, and with

$$\theta_5 = \frac{d(N(d+2)+\theta)}{\theta - N(d+4)} \qquad \theta_6 = \frac{N(dN+\theta)}{\theta - N(d+4)}$$
$$\theta_7 = \frac{N(d+2)-\theta}{\theta - N(d+4)} \qquad \theta_8 = \frac{-2N}{\theta - N(d+4)}.$$

H8: $\psi(n,m) \leq C(n+m+1)^{\widetilde{\beta}}$ and $\widehat{\mathbf{n}} b_{\mathbf{n}}^{\theta'_5} \rho_{\mathbf{n}}^{\theta'_6} \log \widehat{\mathbf{n}}^{\theta'_7} u_{\mathbf{n}}^{\theta'_8} \to \infty$ with $\theta > N(3+2\widetilde{\beta}+d)$, and

$$\theta'_5 = \frac{d(N(d+3)+\theta)}{\theta - N(3+2\widetilde{\beta}+d)} \qquad \theta'_6 = \frac{N(N(d+1)+\theta)}{\theta - N(3+2\widetilde{\beta}+d}$$
$$\theta'_7 = \frac{N(d+1)-\theta}{\theta - N(3+2\widetilde{\beta}+d)} \qquad \theta'_8 = \frac{-2N}{\theta - N(3+2\widetilde{\beta}+d)}.$$

Moreover, by assumption **H3**, we have $\forall \mathbf{j} \in \mathcal{I}_{\mathbf{n}}$, $C_1 k_{\mathbf{nj}} \leq \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K_2\left(\frac{\|t_{\mathbf{i}} - t_{\mathbf{j}}\|}{\rho_{\mathbf{n}}}\right) \leq C_2 k_{\mathbf{nj}}$ (or $k_{\mathbf{nj}} \approx \widehat{\mathbf{n}} \rho_{\mathbf{n}}^N$). We assume that $\exists C_{N,1}$, $C_{N,2}$ such that

$$C_{N,1} \widehat{\mathbf{n}} \rho_{\mathbf{n}}^N \leq k_{\mathbf{n}1} \leq k_{\mathbf{n}2} \leq C_{N,2} \widehat{\mathbf{n}} \rho_{\mathbf{n}}^N$$

with $k_{\mathbf{n}1} = \min_{\mathbf{j}\in V_R} k_{\mathbf{nj}} \approx C_{N,1}\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N \to \infty$ and $k_{\mathbf{n}2} = \max_{\mathbf{j}\in V_R} k_{\mathbf{nj}} \approx C_{N,2}\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N$, where $C_{N,1}$ and $C_{N,2}$ are two constants depending on N. This assumption is not restrictive since $\forall \mathbf{j}$, $k_{\mathbf{nj}} \approx C_N\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N$ (see above).

Then, in order to establish an asymptotic result concerning the modes estimate, we give the following uniform almost sure consistency of the density estimate.

Theorem 2.4. Under assumptions **H1**-**H4**, **H7** or **H8**, we have

$$\sup_{\substack{x_\mathbf{i}\in R\\ \mathbf{i}\in V_R}} |f_\mathbf{n}(x_\mathbf{i}) - f(x_\mathbf{i})| = O\left(b_\mathbf{n} + \sqrt{\frac{\log \widehat{\mathbf{n}}}{\widehat{\mathbf{n}} b_\mathbf{n}^d \rho_\mathbf{n}^N}}\right) \quad \text{a.s.}$$

The proof of this theorem is given in the Appendix.
The consistency of the modes estimator $\hat{\omega}$ is then deduced from the previous result as follows.

Corollary 2.5. Under the same assumptions as Theorem 2.4, and if f admits a unique mode on R, then:

$$\|\widehat{\omega} - \omega\| = O\left(b_\mathbf{n} + \sqrt{\frac{\log \widehat{\mathbf{n}}}{\widehat{\mathbf{n}} b_\mathbf{n}^d \rho_\mathbf{n}^N}}\right) \qquad \text{a.s.}$$

Proof. To prove this corollary, it suffices to remark that

$$\begin{aligned}
|f(\widehat{\omega}) - f(\omega)| &\leq |f(\widehat{\omega}) - f_\mathbf{n}(\widehat{\omega}) + f_\mathbf{n}(\widehat{\omega}) - f(\omega)| \\
&\leq \sup_{\substack{x_\mathbf{i}\in R\\ \mathbf{i}\in V_R}} |f(x_\mathbf{i}) - f_\mathbf{n}(x_\mathbf{i})| + \left|\sup_{\substack{x_\mathbf{i}\in R\\ \mathbf{i}\in V_R}} f_\mathbf{n}(x_\mathbf{i}) - \sup_{\substack{x_\mathbf{i}\in R\\ \mathbf{i}\in V_R}} f(x_\mathbf{i})\right| \\
&\leq 2\sup_{\substack{x_\mathbf{i}\in R\\ \mathbf{i}\in V_R}} |f(x_\mathbf{i}) - f_\mathbf{n}(x_\mathbf{i})|,
\end{aligned}$$

and considering the fact that since f is uniformly continuous with a unique mode ω, then there is a positive real $\eta(\epsilon)$ such that for each point x, $\|x-\omega\| \geq \epsilon$ implies $|f(x) - f(\omega)| > \eta(\epsilon)$. Some calculations lead to conclude using Borel-Cantelli lemma. ■

Now that we have checked the theoretical behavior of our estimator, we will study its behavior through some applications in Chapter 6. Let us notice that such a study requires the development of an algorithm that translates the practical use of our estimator, which is done in Section 2.3. Then a kernel density-based clustering method, based on this algorithm, is proposed in Section 2.3.2.

2.3 Some algorithms for the use of our estimator in practice

In this section, we give a procedure that will help us to use the density estimate in practice. Moreover, a hierarchical unsupervised clustering algorithm, based on the modes estimate given above, is performed. Contrary to several unsupervised clustering methods, such as k-means, the knowledge of the number of clusters is not necessary. The procedure intrinsically determines the number of clusters. This also allows detecting some spatial heterogeneity.

Let us denote by $\mathcal{O}_\mathbf{n} \subset (\mathbb{N}^*)^N$ the set of indices of observed data (not necessarily rectangular). Then, the estimation of the spatial density, based on observations in $\mathcal{O}_\mathbf{n}$, is obtained as follows.

2.3.1 Procedure of estimation of the spatial density in practice

Let us recall that the calculation of (2.4) requires the choice of the bandwidths $b = b_{\mathbf{n}}$ and $\rho = \rho_{\mathbf{n}}$. This choice will be done through some sets chosen by the user. Let $S(b)$ and $S(\rho)$ be the two sets of bandwidths.

- **Step 1**. For each $b_{\mathbf{n}} \in S(b)$ and $\rho_{\mathbf{n}} \in S(\rho)$, compute the function defined by $f_{\mathbf{n}}(x_{\mathbf{i}})$, for all $\mathbf{i} \in \mathcal{O}_{\mathbf{n}}$.
- **Step 2**. Determine the values $b_{\mathbf{n},opt}$ and $\rho_{\mathbf{n},opt}$ that minimize the entropy (see the definition in Chapter 6) of $f_{\mathbf{n}}$ (or maximize the cross-validation procedure) respectively over $S(b)$ and $S(\rho)$.
- **Step 3**. Compute the function $f_{\mathbf{n}}$ that corresponds to $b_{\mathbf{n},opt}$ and $\rho_{\mathbf{n},opt}$. To ease the reading, this function will be denoted by $f_{\mathbf{n},opt}$.

Remark 2.6. When $N = 2$, a 3D plot of $f_{\mathbf{n}}$ can be obtained even if it is defined on $\mathbb{R}^d$, $d > 3$. It suffices to identify $f_{\mathbf{n}}$ with the function $\overline{f}_{\mathbf{n}}$ defined by: $\bar{f}_{\mathbf{n}}(\mathbf{i}) = f_{\mathbf{n}}(x_{\mathbf{i}})$, $x_{\mathbf{i}} \in \mathbb{R}^d$, $\mathbf{i} \in \mathbb{Z}^N$. In fact, it is a specificity of the spatial modeling (unlike the independent case) that allows visualizing the spatial structure of the process $(X_{\mathbf{i}})$ in a given domain (see examples in Dabo-Niang and Yao (2013)).

As a supplement to this spatial structure visualization, one can proceed to a spatial classification method based on local modes estimation that will bring complementary information about the spatial structure distribution on the domain of interest. This is the idea proposed by Dabo-Niang et al. (2010), which is recalled in the next section. For this topic, there are several algorithms; see e.g., CART algorithm of Bel et al. (2009), for spatial data.

2.3.2 A spatial descending hierarchical method

Hereinafter, we suppose that the distribution of $X_{\mathbf{i}}$ has at least a mode. We will use the heterogeneity index of a sample S based on the modes estimation suggested by Dabo-Niang et al. (2006) (see also Ferraty and Vieu (2006)) defined by

$$HI(S) = \frac{\|\hat{\omega}_S - M_S\|}{\|\hat{\omega}_S\| + \|M_S\|},$$

where $\hat{\omega}_S$ and M_S denote respectively the mode and mean (or median) of the sample S. We will also consider its robust version:

$$SHI(S) = \frac{1}{K}\sum_{k=1}^{K} HI(S_k); S_k \subset S,$$

where the S_k's are randomly generated subsamples of S.

Now, since the local modes estimation is an interesting tool to detect a mixture of a population in the data set, Dabo-Niang et al. (2004) (see also Ferraty and Vieu (2006)) proposed a new clustering method based on the heterogeneity measure computed from the estimated local modes. We notice that this technique, first built for functional *i.i.d.* observations (by these authors), has been adapted to functional spatially dependent data by Dabo-Niang et al. (2010). Here, we deal with an adaptation of this procedure in the finite dimensional setting. Let us describe the procedure used. Suppose that the function $f_{\mathbf{n},opt}$ admits L local modes $\hat{\omega}_1, \ldots, \hat{\omega}_L$. Then, $f_{\mathbf{n},opt}$ admits at least $L-1$ local minima: $m_1, \ldots, m_{L-1}$ corresponding to each local mode. In the following, to ease the reading, we set $m_0 = 0$ and $m_L = 1$.

The procedure consists first of computing a splitting score as follows:

1. Compute $f_{\mathbf{n},opt}$ and detect the L spatial local modes and the corresponding minima, $m_1, \ldots, m_{L-1}$; $m_0 = 0$ and $m_L = 1$.
2. Compute $\{\widehat{p}_{\mathbf{i}}(b_{\mathbf{n},\,opt}, \rho_{\mathbf{n},opt})\}_{\mathbf{i}\in\mathcal{O}_{\mathbf{n}}}$ where $\widehat{p}_{\mathbf{i}}(b_{\mathbf{n},\,opt}, \rho_{\mathbf{n},opt}) = \frac{\hat{\mathbf{n}}_i}{\hat{\mathbf{n}}}$, $\hat{\mathbf{n}}_{\mathbf{i}} = \mathrm{Card}(V_{x_{\mathbf{i}}})$ and $V_{x_{\mathbf{i}}} = \{x_{\mathbf{k}};\ \|x_{\mathbf{i}} - x_{\mathbf{k}}\| < b_{\mathbf{n},\,opt}$ and $\|t_{\mathbf{i}} - t_{\mathbf{k}}\| < \rho_{\mathbf{n},opt}\}$, $x_{\mathbf{i}}$, $x_{\mathbf{k}}$ are some of the observation data.
3. Build a partition of the sample: $\mathcal{O}_{\mathbf{n}} = \cup_{\ell=1}^{L} S_\ell$ such that $S_\ell = \{\mathbf{i},\ m_{\ell-1} < \widehat{p}_{\mathbf{i}}(b_{\mathbf{n},\,opt}, \rho_{\mathbf{n},opt}) \leq m_\ell\}$.
4. Compute the splitting score:

$$Gain(\mathcal{O}_{\mathbf{n}}; S_1, \ldots, S_L) = \frac{SHI\,(S) - PHI(S; S_1, \ldots, S_L)}{SHI\,(S)},$$

where $PHI(\mathcal{O}_{\mathbf{n}}; S_1, \ldots, S_L) = \dfrac{1}{\hat{\mathbf{n}}} \displaystyle\sum_{\ell=1}^{L} \mathrm{Card}(S_\ell) \times SHI(S_\ell)$.

Let us notice that if this quantity is positive, it means that there is a loss of heterogeneity (or gain of homogeneity), then a splitting is accepted. If this quantity is negative, S is not split. To control this property of the splitting score, it is necessary to fix a threshold $\gamma > 0$, as it is done in the final procedure clustering algorithm given in the following.

A spatial clustering algorithm

Set a threshold γ.

- Step 1: Compute $b_{\mathbf{n},opt}$ and $\rho_{\mathbf{n},opt}$ using Procedure 2.3.1
  $\star$ **if** $f_{\mathbf{n},opt}$ has many modes, **then** go to Step 2,
  else go to Step 3.
- Step 2: Split the sample and compute *Gain*
  $\star$ **if** $Gain > \gamma$ **then** perform Step 1 for each subgroup,
  else go to Step 3.
- Step 3: Stop.

Remark 2.7.

- In the previous procedure, when the mean is not well adapted, we suggest (as the former authors above mentioned) to replace it by the median.
- Here, since we are in a spatially dependent framework, unlike the *i.i.d.* case, the subsamples should be generated while keeping the spatial dependence.
- If the distribution of a sample is symmetric, we suggest replacing the mean (or the median) by another quantile other than the median.

2.4 Conclusion

In this work, we proposed a new spatial density estimator. The particularity of the new estimator compared to those that exist in the literature is that it includes the distances between sites. Then, the pointwise estimation of the density at a given location is calculated using all the observations while controlling the vicinity of the concerned site. We have also proposed a clustering method based on the new density estimator. The obtained results show that, although this work covers the case of strictly stationary spatial processes, the presented methodology can be generalized to other types of processes such

as *locally dependent spatial processes*. The approach proposed here to estimate the spatial density is used in the following chapter 3 in order to estimate the regression function of a spatial process. Moreover, we illustrate in chapter 6 the classification procedure with some simulations and an application to the MADA data set.

2.5 Appendix

Preliminary result for the proofs

Lemma 2.8. Under conditions of Theorem 2.3, we have

$$\mathbf{I_n}(x_\mathbf{j}) + \mathbf{R_n}(x_\mathbf{j}) = O\left(\frac{1}{\widehat{\mathbf{n}} b_\mathbf{n}^d \rho_\mathbf{n}^N}\right).$$

Proofs

Proof of Theorem 2.3

Let us first show that for each $x_\mathbf{j}$, $f_\mathbf{n}(x_\mathbf{j})$ converges almost completely to $f(x_\mathbf{j})$, *i.e.* that $\forall \epsilon > 0$, $\sum_{\mathbf{n}\in\mathbb{N}^N} \mathbb{P}(|f_\mathbf{n}(x_\mathbf{j}) - f(x_\mathbf{j})| > \epsilon) < \infty$. For $x_\mathbf{j} \in \mathbb{R}^d$ located at somewhere $\mathbf{j}$, remark that:

$$|f_\mathbf{n}(x_\mathbf{j}) - f(x_\mathbf{j})| \leq |f_\mathbf{n}(x_\mathbf{j}) - \mathbb{E}(f_\mathbf{n}(x_\mathbf{j}))| + |\mathbb{E}(f_\mathbf{n}(x_\mathbf{j})) - f(x_\mathbf{j})|.$$

Then, as usual, Theorem 2.3 is obtained by studying the bias and variance terms separately.

The bias $|\mathbb{E}(f_\mathbf{n}(x_\mathbf{j})) - f(x_\mathbf{j})|$**.** Before going further, let us note that

$$\begin{aligned}
|\mathbb{E}(f_\mathbf{n}(x_\mathbf{j})) - f(x_\mathbf{j})| &= \left| \mathbb{E}\left(\frac{1}{\widehat{\mathbf{n}} b_\mathbf{n}^d \rho_\mathbf{n}^N} \sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} K_1\left(\frac{x_\mathbf{j} - X_\mathbf{i}}{b_\mathbf{n}}\right) K_{2,\rho_\mathbf{n}}(\|\mathbf{i}-\mathbf{j}\|) \right) - f(x_\mathbf{j}) \right| \\
&= \left| \frac{1}{\widehat{\mathbf{n}} b_\mathbf{n}^d \rho_\mathbf{n}^N} \sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} K_{2,\rho_\mathbf{n}}(\|\mathbf{i}-\mathbf{j}\|) \int K_1\left(\frac{x_\mathbf{j} - u}{b_\mathbf{n}}\right) f(u)du - f(x_\mathbf{j}) \right| \\
&= \left| \frac{1}{\widehat{\mathbf{n}} \rho_\mathbf{n}^N} \sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} K_{2,\rho_\mathbf{n}}(\|\mathbf{i}-\mathbf{j}\|) \int K_1(v) f(x_\mathbf{j} - v b_\mathbf{n})dv - f(x_\mathbf{j}) \right|.
\end{aligned}$$

Furthermore, since $\frac{1}{\widehat{\mathbf{n}} \rho_\mathbf{n}^N} \sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} K_{2,\rho_\mathbf{n}}(\|\mathbf{i}-\mathbf{j}\|) = 1$ by assumption **H2**,

$$\begin{aligned}
|\mathbb{E}(f_\mathbf{n}(x_\mathbf{j})) - f(x_\mathbf{j})| &\leq C \int K_1(v) |f(x_\mathbf{j} + v b_\mathbf{n}) - f(x_\mathbf{j})| dv \\
&\leq C \int K_1(v) \|x_\mathbf{j} + v b_\mathbf{n} - x_\mathbf{j}\| dv \qquad \text{by assumption } \mathbf{H1} \\
&\leq C \int K_1(v) \|v b_\mathbf{n}\| dv \leq C b_\mathbf{n}
\end{aligned}$$

because $\int \|v\| K_1(v)\, dv < +\infty$.

The study of the asymptotic behavior of the term $|f_{\mathbf{n}}(x_{\mathbf{j}}) - \mathbb{E}(f_{\mathbf{n}}(x_{\mathbf{j}}))|$**.** Let

$$f_{\mathbf{n}}(x_{\mathbf{j}}) - \mathbb{E}(f_{\mathbf{n}}(x_{\mathbf{j}})) = \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \mathbf{\Lambda}_{\mathbf{i}}(x_{\mathbf{j}}) = S_{\mathbf{n}}(x_{\mathbf{j}}),$$

with $\mathbf{\Lambda}_{\mathbf{i}}(x_{\mathbf{j}}) = \lambda_{\mathbf{i}}(x_{\mathbf{j}}) - \mathbb{E}(\lambda_{\mathbf{i}}(x_{\mathbf{j}}))$ where

$$\lambda_{\mathbf{i}}(x_{\mathbf{j}}) = \frac{1}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|).$$

We will now introduce the spatial blocks decomposition introduced by Tran (1990) which will be useful afterward. Without loss of generality, we suppose that $n_k = 2pq_k$ for $1 \leq k \leq N$. The random variables $\mathbf{\Lambda}_{\mathbf{i}}(x_{\mathbf{j}})$ can then be grouped into $2^N q_1 \dots q_N$ cubic blocks of side p. Let,

$$U(1,\mathbf{n},x_{\mathbf{j}},\mathbf{m}) = \sum_{\substack{i_k=2m_kp+1,\\ k=1,\dots,N}}^{(2m_k+1)p} \mathbf{\Lambda}_{\mathbf{i}}(x_{\mathbf{j}}),$$

$$U(2,\mathbf{n},x_{\mathbf{j}},\mathbf{m}) = \sum_{\substack{i_k=2m_kp+1,\\ k=1,\dots,N-1}}^{(2m_k+1)p} \sum_{i_N=(2m_N+1)p+1}^{2(m_N+1)p} \mathbf{\Lambda}_{\mathbf{i}}(x_{\mathbf{j}}),$$

$$U(3,\mathbf{n},x_{\mathbf{j}},\mathbf{m}) = \sum_{\substack{i_k=2m_kp+1,\\ k=1,\dots,N-2}}^{(2m_k+1)p} \sum_{i_{N-1}=(2m_{N-1}+1)p+1}^{2(m_{N-1}+1)p} \sum_{i_N=2m_Np+1}^{(2m_N+1)p} \mathbf{\Lambda}_{\mathbf{i}}(x_{\mathbf{j}}),$$

$$U(4,\mathbf{n},x_{\mathbf{j}},\mathbf{m}) = \sum_{\substack{i_k=2m_kp+1,\\ k=1,\dots,N-2}}^{(2m_k+1)p} \sum_{i_{N-1}=(2m_{N-1}+1)p+1}^{2(m_{N-1}+1)p} \sum_{i_N=(2m_N+1)p+1}^{(2m_N+1)p} \mathbf{\Lambda}_{\mathbf{i}}(x_{\mathbf{j}}),$$

and so on, noticing that

$$U(2^{N-1},\mathbf{n},x_{\mathbf{j}},\mathbf{m}) = \sum_{\substack{i_k=(2m_k+1)p+1,\\ k=1,\dots,N-1}}^{2(m_k+1)p} \sum_{i_N=2m_Np+1}^{(2m_N+1)p} \mathbf{\Lambda}_{\mathbf{i}}(x_{\mathbf{j}}),$$

$$U(2^N,\mathbf{n},x_{\mathbf{j}},\mathbf{m}) = \sum_{\substack{i_k=(2m_k+1)p+1,\\ k=1,\dots,N}}^{2(m_k+1)p} \mathbf{\Lambda}_{\mathbf{i}}(x_{\mathbf{j}}).$$

For each integer $1 \leq l \leq 2^N$, we define $T(\mathbf{n},l,x_{\mathbf{j}}) = \sum_{\substack{m_k=0,\\ k=1,\dots,N}}^{q_k-1} U(l,\mathbf{n},x_{\mathbf{j}},\mathbf{m})$. We obtain $S_{\mathbf{n}}(x_{\mathbf{j}}) = \sum_{l=1}^{2^N} T(\mathbf{n},l,x_{\mathbf{j}})$. For $\epsilon > 0$,

$$\begin{aligned} P = \mathbb{P}(|f_{\mathbf{n}}(x_{\mathbf{j}}) - \mathbb{E}(f_{\mathbf{n}}(x_{\mathbf{j}}))| > \epsilon) &= \mathbb{P}(|S_{\mathbf{n}}(x_{\mathbf{j}})| > \epsilon) \\ &= \mathbb{P}\left(\left|\sum_{l=1}^{2^N} T(\mathbf{n},l,x_{\mathbf{j}})\right| > \epsilon\right) \\ &\leq 2^N \mathbb{P}\left(|T(\mathbf{n},1,x_{\mathbf{j}})| > \frac{\epsilon}{2^N}\right). \end{aligned}$$

We enumerate in an arbitrary manner the $\widehat{q} = q_1 \times \dots \times q_N$ terms $U(1,\mathbf{n},x_{\mathbf{j}},\mathbf{m})$ of the sum $T(\mathbf{n},1,x_{\mathbf{j}})$ and refer to them as $W_1,\dots,W_{\widehat{q}}$. Note that $U(1,\mathbf{n},x_{\mathbf{j}},\mathbf{m})$ is a measurable

σ-algebra generated by $\lambda_{\mathbf{i}}(x_{\mathbf{j}})$, with $\mathbf{i}$ such that $2j_kp+1 \leq i_k \leq (2j_k+1)p$, $k=1,\dots,N$. For all $l=1,\dots,\widehat{q}$, the sets of the sites in W_l are separated by a distance of at least equal to p. In addition, since K_2 and K_1 are bounded and $C_{\mathbf{nj}} \leq 2/C_1$, we can write $|W_l| \leq (\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N)^{-1}Cp^N$, with $C = 2\|K_1\|_\infty\|K_2\|_\infty/C_1$. (where $\|\cdot\|_\infty$ is the sup norm). Lemma A.2 insures the existence of some random variables $W_1^*, W_2^*, \dots, W_{\widehat{q}}^*$ such that

$$\begin{aligned}\sum_{l=1}^{\widehat{q}}\mathbb{E}|W_l - W_l^*| &\leq 2\widehat{q}\frac{1}{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}p^NC\psi((\widehat{q}-1)p^N,p^N)\varphi(p)\\ &\leq 2\widehat{q}\frac{1}{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}p^NC\psi(\widehat{\mathbf{n}},p^N)\varphi(p).\end{aligned}$$

Markov inequality allows us to write

$$\mathbb{P}\left(\sum_{l=1}^{\widehat{q}}|W_l - W_l^*| > \frac{\epsilon}{2^{N+1}}\right) \leq 2\widehat{q}\frac{1}{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}p^NC\psi(\widehat{\mathbf{n}},p^N)\varphi(p)\frac{2^{N+1}}{\epsilon},$$

by Bernstein inequality, we have

$$\mathbb{P}\left(\sum_{l=1}^{\widehat{q}}|W_l^*| > \frac{\epsilon}{2^{N+1}}\right) \leq 2\exp\left\{\frac{-\epsilon^2/(2^{N+1})^2}{4\sum_{l=1}^{\widehat{q}}\mathbb{E}(W_l^{*2}) + \frac{2\epsilon}{2^{N+1}}\frac{1}{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}p^NC}\right\}$$

which leads to

$$\begin{aligned}P &\leq 2^N\mathbb{P}\left(\sum_{l=1}^{\widehat{q}}|W_l^*| > \frac{\epsilon}{2^{N+1}}\right) + 2^N\mathbb{P}\left(\sum_{l=1}^{\widehat{q}}|W_l - W_l^*| > \frac{\epsilon}{2^{N+1}}\right)\\ &\leq 2^{N+1}\exp\left\{\frac{-\epsilon^2/(2^{N+1})^2}{4\sum_{l=1}^{\widehat{q}}\mathbb{E}(W_l^{*2}) + \frac{2\epsilon}{2^{N+1}}\frac{C}{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}p^N}\right\} + 2^{N+1}\widehat{q}\frac{1}{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}p^NC\psi(\widehat{\mathbf{n}},p^N)\varphi(p)\frac{2^{N+1}}{\epsilon}.\end{aligned}$$

Let $\delta > 0$ and

$$\epsilon = \epsilon_{\mathbf{n}} = \delta\left(\frac{\log\widehat{\mathbf{n}}}{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\right)^{\frac{1}{2}} \qquad \text{and} \qquad p = \left\lfloor\left(\frac{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}{\log\widehat{\mathbf{n}}}\right)^{\frac{1}{2N}}\right\rfloor.$$

Since the variables W_l and W_l^* have the same distributions, we have

$$\begin{aligned}\sum_{l=1}^{\widehat{q}}\mathbb{E}W_l^{*2} = \sum_{l=1}^{\widehat{q}}Var(W_l^*) &= \sum_{l=1}^{\widehat{q}}Var(W_l)\\ &\leq \mathbf{I}_{\mathbf{n}}(x_{\mathbf{j}}) + \mathbf{R}_{\mathbf{n}}(x_{\mathbf{j}}).\end{aligned}$$

Lemma 2.8 leads then to $\sum_{l=1}^{\widehat{q}}\mathbb{E}(W_l^{*2}) = O\left(\frac{1}{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\right)$. Hence,

$$\begin{aligned}P &\leq 2^{N+1}\exp\left\{\frac{-\epsilon^2/(2^{N+1})^2}{4\frac{C}{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N} + 2C\frac{1}{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}p^N\frac{\epsilon}{2^{N+1}}}\right\} + C2^{N+1}\psi(\widehat{\mathbf{n}},p^N)\frac{1}{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\widehat{q}p^N\frac{2^{N+1}}{\epsilon}\varphi(p)\\ &\leq 2^{N+1}\exp\left\{\frac{-\epsilon^2}{2^{2N+4}\frac{C}{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N} + 2^{N+2}\frac{C}{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\delta}\right\} + C2^{N+1}\psi(\widehat{\mathbf{n}},p^N)\frac{1}{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\frac{\widehat{\mathbf{n}}}{2^Np^N}p^N\frac{2^{N+1}}{\epsilon}\varphi(p)\end{aligned}$$

$$
\begin{aligned}
P &\leq 2^{N+1}\exp\left\{\frac{-\delta^2\log\widehat{\mathbf{n}}}{2^{2N+4}C+2^{N+2}C\delta}\right\}+C2^{N+2}\psi(\widehat{\mathbf{n}},p^N)\frac{1}{b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\epsilon^{-1}p^{-\theta}\\
&\leq 2^{N+1}\exp\left\{\log\widehat{\mathbf{n}}^{-a}\right\}+C2^{N+2}\psi(\widehat{\mathbf{n}},p^N)\frac{1}{b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\epsilon^{-1}p^{-\theta}\\
&\leq C\widehat{\mathbf{n}}^{-a}+C2^{N+2}\psi(\widehat{\mathbf{n}},p^N)\frac{1}{b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\delta^{-1}\left(\frac{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}{\log\widehat{\mathbf{n}}}\right)^{\frac{N-\theta}{2N}}
\end{aligned}
$$

with $a=\frac{\delta^2}{2^{2N+4}C+2^{N+2}C\delta}$. We first show that the series $\sum_{\mathbf{n}\in\mathbb{N}^N}C\widehat{\mathbf{n}}^{-a}$ converges if and only if $a>1$, that is $\delta^2-2^{N+2}C\delta-2^{2N+4}C>0$. Solving the second order equation and using the fact that $\delta>0$, one easily shows that the solution is $\delta>2^{N+1}C(1+\sqrt{C4})>2^{N+1}C$. Now, we treat the second term for which two cases from assumptions on $\psi(n,m)$. In the situation of assumption **H5**, we have $\psi(n,m)\leq C\min(n,m)$ and we set

$$
\begin{aligned}
g_{\mathbf{n}} &= C2^{N+2}\psi(\widehat{\mathbf{n}},p^N)\frac{1}{b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\delta^{-1}\left(\frac{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}{\log\widehat{\mathbf{n}}}\right)^{\frac{N-\theta}{2N}}\\
&\leq C2^{N+2}p^N\frac{1}{b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\delta^{-1}\left(\frac{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}{\log\widehat{\mathbf{n}}}\right)^{\frac{N-\theta}{2N}}\\
&\leq C2^{N+2}\frac{1}{b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\delta^{-1}\left(\frac{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}{\log\widehat{\mathbf{n}}}\right)^{\frac{2N-\theta}{2N}}
\end{aligned}
$$

To show that $\sum_{\mathbf{n}\in\mathbb{N}^N}g_{\mathbf{n}}<\infty$, let the following sequence of calculations

$$
\begin{aligned}
g_{\mathbf{n}}\widehat{\mathbf{n}}u_{\mathbf{n}} &\leq C2^{N+2}\frac{1}{b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\delta^{-1}\left(\frac{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}{\log\widehat{\mathbf{n}}}\right)^{\frac{2N-\theta}{2N}}\widehat{\mathbf{n}}u_{\mathbf{n}}\\
&\leq C\left(\widehat{\mathbf{n}}b_{\mathbf{n}}^{\frac{-d\theta}{4N-\theta}}\rho_{\mathbf{n}}^{\frac{-N\theta}{4N-\theta}}\log\widehat{\mathbf{n}}^{\frac{\theta-2N}{4N-\theta}}u_{\mathbf{n}}^{\frac{2N}{4N-\theta}}\right)^{\frac{4N-\theta}{2N}}\leq C\mathcal{A}_{\mathbf{n}}^{\frac{4N-\theta}{2N}}.
\end{aligned}
$$

Note that in the case where $\theta>4N$, one has $\frac{4N-\theta}{2N}<0$. Furthermore, assumption **H5** leads to $\mathcal{A}_{\mathbf{n}}\to+\infty$, as $\mathbf{n}\to\infty$. So, $\mathcal{A}_{\mathbf{n}}^{\frac{4N-\theta}{2N}}\to 0$ and $g_{\mathbf{n}}\widehat{\mathbf{n}}u_{\mathbf{n}}\to 0$; consequently, $g_{\mathbf{n}}<\frac{1}{\widehat{\mathbf{n}}u_{\mathbf{n}}}$ leads to $\sum_{\mathbf{n}\in\mathbb{N}^N}g_{\mathbf{n}}<\sum_{\mathbf{n}\in\mathbb{N}^N}\frac{1}{\widehat{\mathbf{n}}u_{\mathbf{n}}}<+\infty$.
In the second situation, which corresponds to assumption **H6**, we have $\psi(n,m)\leq C(n+m+1)^{\widetilde{\beta}}$. Let us note that $\psi(\widehat{\mathbf{n}},p^N)\leq C(\widehat{\mathbf{n}}+p^N+1)^{\widetilde{\beta}}\leq C\widehat{\mathbf{n}}^{\widetilde{\beta}}$ and set

$$
\begin{aligned}
h_{\mathbf{n}} &= C2^{N+2}\psi(\widehat{\mathbf{n}},p^N)\frac{1}{b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\delta^{-1}\left(\frac{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}{\log\widehat{\mathbf{n}}}\right)^{\frac{N-\theta}{2N}}\\
&\leq C2^{N+2}\widehat{\mathbf{n}}^{\widetilde{\beta}}\frac{1}{b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\delta^{-1}\left(\frac{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}{\log\widehat{\mathbf{n}}}\right)^{\frac{N-\theta}{2N}}
\end{aligned}
$$

To show that $\sum_{\mathbf{n}\in\mathbb{N}^N}h_{\mathbf{n}}<\infty$, let the following sequence of calculations

$$
\begin{aligned}
h_{\mathbf{n}}\widehat{\mathbf{n}}u_{\mathbf{n}} &\leq C2^{N+2}\widehat{\mathbf{n}}^{1+\widetilde{\beta}}\frac{1}{b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\delta^{-1}\left(\frac{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}{\log\widehat{\mathbf{n}}}\right)^{\frac{N-\theta}{2N}}u_{\mathbf{n}}\\
&\leq C\left(\widehat{\mathbf{n}}b_{\mathbf{n}}^{\frac{-d(\theta+N)}{N(3+2\widetilde{\beta})-\theta}}\rho_{\mathbf{n}}^{\frac{-N(N+\theta)}{N(3+2\widetilde{\beta})-\theta}}\log\widehat{\mathbf{n}}^{\frac{\theta-N}{N(3+2\widetilde{\beta})-\theta}}u_{\mathbf{n}}^{\frac{2N}{N(3+2\widetilde{\beta})-\theta}}\right)^{\frac{N(3+2\widetilde{\beta})-\theta}{2N}}\leq C\mathcal{B}_{\mathbf{n}}^{\frac{N(3+2\widetilde{\beta})-\theta}{2N}}.
\end{aligned}
$$

Now, since in this case $\theta > N(2\widetilde{\beta}+3)$, we have $\frac{N(3+2\widetilde{\beta})-\theta}{2N} < 0$. Assumption **H6** allows us to say that $\mathcal{B}_{\mathbf{n}} \to +\infty$, as $\mathbf{n} \to \infty$. Moreover, $\mathcal{B}_{\mathbf{n}}^{\frac{N(3+2\widetilde{\beta})-\theta}{2N}} \to 0$ then $\widehat{\mathbf{n}} u_{\mathbf{n}} h_{\mathbf{n}} \to 0$. Hence, $h_{\mathbf{n}} < \dfrac{1}{\widehat{\mathbf{n}} u_{\mathbf{n}}} \Leftrightarrow \sum\limits_{\mathbf{n}\in\mathbb{N}^N} h_{\mathbf{n}} < \sum\limits_{\mathbf{n}\in\mathbb{N}^N} \dfrac{1}{\widehat{\mathbf{n}} u_{\mathbf{n}}} < +\infty$. Consequently, in the two cases, we have

$$\sum_{\mathbf{n}\in\mathbb{N}^N} C\widehat{\mathbf{n}}^{-a} + C2^{N+2}\psi(\widehat{\mathbf{n}}, p^N)\frac{1}{b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}\delta^{-1}\left(\frac{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}{\log \widehat{\mathbf{n}}}\right)^{\frac{N-\theta}{2N}} < \infty.$$

■

Proof of Lemma 2.8

From the sequel of calculations:

$$\mathbb{E}\left[|f_{\mathbf{n}}(x_{\mathbf{j}}) - \mathbb{E}(f_{\mathbf{n}}(x_{\mathbf{j}}))|^2\right] = \mathbb{E}\left[\left(\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \mathbf{\Lambda}_{\mathbf{i}}(x_{\mathbf{j}})\right)^2\right] \leq \mathbf{I}_{\mathbf{n}}(x_{\mathbf{j}}) + \mathbf{R}_{\mathbf{n}}(x_{\mathbf{j}})$$

where $\mathbf{I}_{\mathbf{n}}(x_{\mathbf{j}}) = \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \mathbb{E}\left[\left(\mathbf{\Lambda}_{\mathbf{i}}(x_{\mathbf{j}})\right)^2\right]$, $\mathbf{R}_{\mathbf{n}}(x_{\mathbf{j}}) = \sum_{\mathbf{i},\mathbf{k}\in\mathcal{I}_{\mathbf{n}}} \sum_{\mathbf{i}\neq\mathbf{k}} \left|\mathbb{E}\left[\mathbf{\Lambda}_{\mathbf{i}}(x_{\mathbf{j}})\mathbf{\Lambda}_{\mathbf{k}}(x_{\mathbf{j}})\right]\right|$. We deduce

$$\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N \mathbf{I}_{\mathbf{n}}(x_{\mathbf{j}}) = \widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \mathbb{E}\left(\lambda_{\mathbf{i}}(x_{\mathbf{j}})\right)^2 - \widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \left(\mathbb{E}\lambda_{\mathbf{i}}(x_{\mathbf{j}})\right)^2.$$

Noticing that

$$\frac{C_1}{(2^N+1)C_2}\mathbf{1}_{[0,1]}\left(\frac{\|t_{\mathbf{i}} - t_{\mathbf{j}}\|}{\rho_{\mathbf{n}}}\right) \leq K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|) \leq \frac{2C_2}{C_1}\mathbf{1}_{[0,1]}\left(\frac{\|t_{\mathbf{i}} - t_{\mathbf{j}}\|}{\rho_{\mathbf{n}}}\right).$$

First, we have

$$\begin{aligned}
\diamondsuit \quad \widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \mathbb{E}\left(\lambda_{\mathbf{i}}(x_{\mathbf{j}})\right)^2 &= \widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \mathbb{E}\left(\frac{1}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|)\right)^2 \\
&= \frac{1}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{2,\rho_{\mathbf{n}}}^2(\|\mathbf{i}-\mathbf{j}\|)\mathbb{E}\left(K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right)\right)^2 \\
&\leq \frac{C}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|) \int K_1^2(t) f(x_{\mathbf{j}} - b_{\mathbf{n}}t)(-b_{\mathbf{n}})^d dt \\
&\leq \frac{C}{\widehat{\mathbf{n}} \rho_{\mathbf{n}}^N} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|) \int K_1^2(t) f(x_{\mathbf{j}} - b_{\mathbf{n}}t) dt \\
&\leq C.
\end{aligned}$$

Thus, $\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \mathbb{E}\left(\lambda_{\mathbf{i}}(x_{\mathbf{j}})\right)^2 = O\left((\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N)^{-1}\right)$. Now, noticing that

$$\begin{aligned}
\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \left(\mathbb{E}\lambda_{\mathbf{i}}(x_{\mathbf{j}})\right)^2 &= \frac{1}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{2,\rho_{\mathbf{n}}}^2(\|\mathbf{i}-\mathbf{j}\|)\left(\mathbb{E}K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right)\right)^2 \\
&\leq \frac{C}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|)\left(\int K_1(u) f(x_{\mathbf{j}} - ub_{\mathbf{n}})(-b_{\mathbf{n}}^d) du\right)^2 \\
&\leq C b_{\mathbf{n}}^d
\end{aligned}$$

we have $\lim_{\widehat{\mathbf{n}}\to\infty} \widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \left(\mathbb{E}\lambda_{\mathbf{i}}(x_{\mathbf{j}})\right)^2 = O(b_{\mathbf{n}}^d)$.
We now treat the term $\mathbf{R_n}(x_{\mathbf{j}}) = \sum_{\mathbf{i},\mathbf{k}\in\mathcal{I}_{\mathbf{n}}} \sum_{\mathbf{i}\neq\mathbf{k}} \left|\mathbb{E}\mathbf{\Lambda_i}(x_{\mathbf{j}})\mathbf{\Lambda_k}(x_{\mathbf{j}})\right|$ and show that there exists a constant C such that $\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N \mathbf{R_n}(x_{\mathbf{j}}) < C$, for $\widehat{\mathbf{n}}$ large enough. Let $D_{\mathbf{n}}$ be a sequence of real numbers tending to ∞ as $\widehat{\mathbf{n}} \to \infty$. Let $S = \{\mathbf{i},\mathbf{k} \in \mathcal{I}_{\mathbf{n}}, \|\mathbf{i}-\mathbf{k}\| \leq D_{\mathbf{n}}\}$ and denote by S^c the complementary of S. Let $R_{\mathbf{n}}^{(1)} = \sum_{\mathbf{i},\mathbf{k}\in S} \left|\mathbb{E}\mathbf{\Lambda_i}(x_{\mathbf{j}})\mathbf{\Lambda_k}(x_{\mathbf{j}})\right|$ and $R_{\mathbf{n}}^{(2)} = \sum_{\mathbf{i},\mathbf{k}\in S^c} \left|\mathbb{E}\mathbf{\Lambda_i}(x_{\mathbf{j}})\mathbf{\Lambda_k}(x_{\mathbf{j}})\right|$. Hence, $\mathbf{R_n}(x_{\mathbf{j}}) \leq R_{\mathbf{n}}^{(1)} + R_{\mathbf{n}}^{(2)}$.

$$\diamondsuit \quad R_{\mathbf{n}}^{(1)} = \sum_{\mathbf{i},\mathbf{k}\in S} \left|\mathbb{E}(\lambda_{\mathbf{i}}(x_{\mathbf{j}})\lambda_{\mathbf{k}}(x_{\mathbf{j}})) - \mathbb{E}\lambda_{\mathbf{i}}(x_{\mathbf{j}})\mathbb{E}\lambda_{\mathbf{k}}(x_{\mathbf{j}})\right| = \sum_{\mathbf{i},\mathbf{k}\in S} |\mathbf{A}-\mathbf{B}|.$$

$$\mathbf{A} = \frac{1}{\widehat{\mathbf{n}}^2 b_{\mathbf{n}}^{2d} \rho_{\mathbf{n}}^{2N}} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{k}-\mathbf{j}\|) \mathbb{E}\left(K_1\left(\frac{x_{\mathbf{j}}-X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_1\left(\frac{x_{\mathbf{j}}-X_{\mathbf{k}}}{b_{\mathbf{n}}}\right)\right)$$
$$= \frac{1}{\widehat{\mathbf{n}}^2 b_{\mathbf{n}}^{2d} \rho_{\mathbf{n}}^{2N}} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{k}-\mathbf{j}\|) \mathbf{A}'$$
$$\mathbf{A}' = \int\int K_1\left(\frac{x_{\mathbf{j}}-u}{b_{\mathbf{n}}}\right) K_1\left(\frac{x_{\mathbf{j}}-v}{b_{\mathbf{n}}}\right) f_{X_{\mathbf{i}}X_{\mathbf{k}}}(u,v)dudv$$
$$\mathbf{B} = \frac{1}{\widehat{\mathbf{n}}^2 b_{\mathbf{n}}^{2d} \rho_{\mathbf{n}}^{2N}} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{k}-\mathbf{j}\|) \mathbb{E}\left(K_1\left(\frac{x_{\mathbf{j}}-X_{\mathbf{i}}}{b_{\mathbf{n}}}\right)\right) \mathbb{E}\left(K_1\left(\frac{x_{\mathbf{j}}-X_{\mathbf{k}}}{b_{\mathbf{n}}}\right)\right)$$
$$= \frac{1}{\widehat{\mathbf{n}}^2 b_{\mathbf{n}}^{2d} \rho_{\mathbf{n}}^{2N}} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{k}-\mathbf{j}\|) \mathbf{B}'$$
$$\mathbf{B}' = \left(\int K_1\left(\frac{x_{\mathbf{j}}-u}{b_{\mathbf{n}}}\right) f_{X_{\mathbf{i}}}(u)du\right)\left(\int K_1\left(\frac{x_{\mathbf{j}}-v}{b_{\mathbf{n}}}\right) f_{X_{\mathbf{k}}}(v)dv\right)$$

By assumption **H4**, we have: $\sup_{u,\,v} \sup_{\mathbf{i},\,\mathbf{k}} |f_{X_{\mathbf{i}}X_{\mathbf{k}}}(u,v) - f_{X_{\mathbf{i}}}(u) f_{X_{\mathbf{k}}}(v)| \leq C$, then

$$|\mathbf{A}'-\mathbf{B}'| \leq C \int\int K_1\left(\frac{x_{\mathbf{j}}-u}{b_{\mathbf{n}}}\right) K_1\left(\frac{x_{\mathbf{j}}-v}{b_{\mathbf{n}}}\right) dudv.$$

Thus,

$$R_{\mathbf{n}}^{(1)} = \sum_{\mathbf{i},\mathbf{k}\in S} \frac{1}{\widehat{\mathbf{n}}^2 b_{\mathbf{n}}^{2d} \rho_{\mathbf{n}}^{2N}} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{k}-\mathbf{j}\|) \left|\mathbf{A}'-\mathbf{B}'\right|$$
$$\leq C \sum_{\mathbf{i},\mathbf{k}\in S} \frac{1}{\widehat{\mathbf{n}}^2 b_{\mathbf{n}}^{2d} \rho_{\mathbf{n}}^{2N}} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{k}-\mathbf{j}\|) \int\int K_1\left(\frac{x_{\mathbf{j}}-u}{b_{\mathbf{n}}}\right) K_1\left(\frac{x_{\mathbf{j}}-v}{b_{\mathbf{n}}}\right) dudv$$

and **H3** leads to

$$\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N R_{\mathbf{n}}^{(1)} \leq C \frac{\widehat{\mathbf{n}} b_{\mathbf{n}}^{3d} \rho_{\mathbf{n}}^N}{\widehat{\mathbf{n}}^2 b_{\mathbf{n}}^{2d} \rho_{\mathbf{n}}^{2N}} \sum_{\mathbf{i},\mathbf{k}\in S} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{k}-\mathbf{j}\|) \left(\int |K_1(u)|\, du\right)^2$$
$$\leq C \frac{\widehat{\mathbf{n}} b_{\mathbf{n}}^{3d} \rho_{\mathbf{n}}^N}{\widehat{\mathbf{n}}^2 b_{\mathbf{n}}^{2d} \rho_{\mathbf{n}}^{2N}} \sum_{\mathbf{i},\mathbf{k}\in S} \mathbf{1}_{[0,1]}\left(\frac{\|t_{\mathbf{i}}-t_{\mathbf{j}}\|}{\rho_{\mathbf{n}}}\right) \mathbf{1}_{[0,1]}\left(\frac{\|t_{\mathbf{k}}-t_{\mathbf{j}}\|}{\rho_{\mathbf{n}}}\right) \left(\int |K_1(u)|\, du\right)^2$$
$$\leq C \frac{b_{\mathbf{n}}^d}{\widehat{\mathbf{n}} \rho_{\mathbf{n}}^N} \sum_{\mathbf{i},\mathbf{k}\in V_{\mathbf{j}}} \mathbf{1}_{[0,1]}\left(\frac{\|\mathbf{k}-\mathbf{i}\|}{D_{\mathbf{n}}}\right) \left(\int |K_1(u)|\, du\right)^2$$
$$\leq 2^N C \frac{b_{\mathbf{n}}^d}{\widehat{\mathbf{n}} \rho_{\mathbf{n}}^N} \sum_{\mathbf{i}\in V_{\mathbf{j}}} \sum_{\mathbf{i}-\mathbf{u}\in V_{\mathbf{j}}} \mathbf{1}_{\{\mathbf{u};\|\mathbf{u}\|\leq D_{\mathbf{n}}\}} \left(\int |K_1(u)|\, du\right)^2$$

where $V_{\mathbf{j}} = \{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}, \|t_{\mathbf{i}}-t_{\mathbf{j}}\| \leq \rho_{\mathbf{n}}\}$ with $\mathrm{Card}(V_{\mathbf{j}}) = k_{\mathbf{n}} \leq C\hat{\mathbf{n}}\rho_{\mathbf{n}}^N$. Then

$$\sum_{\mathbf{i}\in V_{\mathbf{j}}} \sum_{\mathbf{i}-\mathbf{u}\in V_{\mathbf{j}}} \mathbf{1}_{\{\mathbf{u};\|\mathbf{u}\|\leq D_{\mathbf{n}}\}} \leq C_0 \hat{\mathbf{n}} \rho_{\mathbf{n}}^N \lfloor D_{\mathbf{n}} \rfloor^N \leq C_0 \hat{\mathbf{n}} \rho_{\mathbf{n}}^N D_{\mathbf{n}}^N$$

where $\lfloor\cdot\rfloor$ denotes the integer part. So, $\hat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N R_{\mathbf{n}}^{(1)} \leq Cb_{\mathbf{n}}^d D_{\mathbf{n}}^N$.
$\diamondsuit$ Now, since the functions $K_1(\cdot)$ and $K_2(\cdot)$ are bounded, by applying Lemma A.1 we get:

$$|\mathbb{E}\mathbf{\Lambda}_{\mathbf{i}}(x_{\mathbf{j}})\mathbf{\Lambda}_{\mathbf{k}}(x_{\mathbf{j}})| \leq C\frac{K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|)K_{2,\rho_{\mathbf{n}}}(\|\mathbf{k}-\mathbf{j}\|)}{\hat{\mathbf{n}}^2 b_{\mathbf{n}}^{2d}\rho_{\mathbf{n}}^{2N}}\psi(1,1)\varphi(\|\mathbf{i}-\mathbf{k}\|).$$

and we can write

$$\begin{aligned}
\hat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N R_{\mathbf{n}}^{(2)} &\leq \hat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N \sum_{\mathbf{i},\mathbf{k}\in S^c} C\frac{K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|)K_{2,\rho_{\mathbf{n}}}(\|\mathbf{k}-\mathbf{j}\|)}{\hat{\mathbf{n}}^2 b_{\mathbf{n}}^{2d}\rho_{\mathbf{n}}^{2N}}\psi(1,1)\varphi(\|\mathbf{i}-\mathbf{k}\|) \\
&\leq \frac{C}{\hat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\sum_{\mathbf{i},\mathbf{k}\in S^c\cap V_{\mathbf{j}}}\varphi(\|\mathbf{i}-\mathbf{k}\|) \\
&\leq \frac{2^N C}{\hat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\sum_{\mathbf{k}\in V_{\mathbf{j}}}\sum_{\mathbf{k}-\mathbf{u}\in V_{\mathbf{j}},\|\mathbf{u}\|>D_{\mathbf{n}}}\varphi(\|\mathbf{u}\|) \\
&\leq \frac{Ck_{\mathbf{n}}}{\hat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\sum_{\|\mathbf{i}\|>D_{\mathbf{n}}}\varphi(\|\mathbf{i}\|) \\
&\left(\text{because}\,\forall\,\mathbf{k}\in\mathbb{N}^N, 2^N\sum_{\mathbf{k}-\mathbf{u}\in V_{\mathbf{j}},\|\mathbf{u}\|>D_{\mathbf{n}}}\varphi(\|\mathbf{u}\|) = \sum_{\|\mathbf{i}\|>D_{\mathbf{n}}}\varphi(\|\mathbf{i}\|)\right) \\
&\leq \frac{C}{b_{\mathbf{n}}^d}\sum_{\|\mathbf{i}\|>D_{\mathbf{n}}}\varphi(\|\mathbf{i}\|).
\end{aligned}$$

Since $\sum_{\|\mathbf{i}\|>D_{\mathbf{n}}}\varphi(\|\mathbf{i}\|) \leq C\sum_{\|\mathbf{i}\|>D_{\mathbf{n}}}\|\mathbf{i}\|^{-\theta} \leq C\sum_{\|\mathbf{i}\|>D_{\mathbf{n}}}\|\mathbf{i}\|^{-\theta}\|\mathbf{i}\|^{-N}\|\mathbf{i}\|^{N}$ and $\|\mathbf{i}\|^{-N} \leq (D_{\mathbf{n}})^{-N}$, we have $\sum_{\|\mathbf{i}\|>D_{\mathbf{n}}}\varphi(\|\mathbf{i}\|) \leq CD_{\mathbf{n}}^{-N}\sum_{\|\mathbf{i}\|>D_{\mathbf{n}}}\|\mathbf{i}\|^{N-\theta}$.

$$\hat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N R_{\mathbf{n}}^{(2)} \leq \frac{C}{b_{\mathbf{n}}^d}D_{\mathbf{n}}^{-N}\sum_{\|\mathbf{i}\|>D_{\mathbf{n}}}\|\mathbf{i}\|^{N-\theta}$$

The fact that $\theta > 4N$ leads to choose $D_{\mathbf{n}} = (b_{\mathbf{n}}^d)^{-1/N}$which gives the expected result. In fact, we have $\lim_{\mathbf{n}\to\infty}\hat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N R_{\mathbf{n}}^{(2)} = C$ and $\lim_{\mathbf{n}\to\infty}\hat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N R_{\mathbf{n}}^{(1)} = 0$. So, $R_{\mathbf{n}}^{(1)} = O\left((\hat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N)^{-1}\right)$, $R_{\mathbf{n}}^{(2)} = O\left((\hat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N)^{-1}\right)$ and then, $\mathbf{R}_{\mathbf{n}}(x_{\mathbf{j}}) = O\left((\hat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N)^{-1}\right)$. That shows that $\mathbf{I}_{\mathbf{n}}(x_{\mathbf{j}}) + \mathbf{R}_{\mathbf{n}}(x_{\mathbf{j}}) = O\left((\hat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N)^{-1}\right)$ for $\hat{\mathbf{n}}$ enough large. ■

Proof of Theorem 2.4

We note that

$$\sup_{\substack{x_{\mathbf{j}}\in R\\ \mathbf{j}\in V_R}}|f_{\mathbf{n}}(x_{\mathbf{j}}) - f(x_{\mathbf{j}})| \leq \sup_{\substack{x_{\mathbf{j}}\in R\\ \mathbf{j}\in V_R}}|f_{\mathbf{n}}(x_{\mathbf{j}}) - \mathbb{E}(f_{\mathbf{n}}(x_{\mathbf{j}}))| + \sup_{\substack{x_{\mathbf{j}}\in R\\ \mathbf{j}\in V_R}}|\mathbb{E}(f_{\mathbf{n}}(x_{\mathbf{j}})) - f(x_{\mathbf{j}})|$$

First, from the bias term of the proof of Theorem 2.3, we can easily write that

$$\sup_{\substack{x_{\mathbf{j}}\in R\\ \mathbf{j}\in V_R}}|\mathbb{E}(f_{\mathbf{n}}(x_{\mathbf{j}})) - f(x_{\mathbf{j}})| = O(b_{\mathbf{n}}).$$

Second, we deal with $\sup_{\substack{x_{\mathbf{j}}\in R\\ \mathbf{j}\in V_R}}|f_{\mathbf{n}}(x_{\mathbf{j}}) - \mathbb{E}(f_{\mathbf{n}}(x_{\mathbf{j}}))|$. Since R is a compact set in $\mathbb{R}^d$, it can be covered with, say, v cubes I_k having sides of length l and the center at x_k, $k = 1, \ldots, v$.

For $x_{\mathbf{j}}$ at a fixed site $\mathbf{j}$, let $x_{\mathbf{j}}$ belong to I_k and

$$\widehat{f}(x_k) = \frac{1}{a_{\mathbf{nj}} b_{\mathbf{n}}^d} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_1\left(\frac{x_k - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_2\left(\frac{\|t_{\mathbf{i}} - t_{\mathbf{j}}\|}{\rho_{\mathbf{n}}}\right)$$

with $a_{\mathbf{nj}} = \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_2\left(\frac{\|t_{\mathbf{i}}-t_{\mathbf{j}}\|}{\rho_{\mathbf{n}}}\right)$. According to Remark 2.2, we can rewrite $f_{\mathbf{n}}(x_{\mathbf{j}})$ as

$$f_{\mathbf{n}}(x_{\mathbf{j}}) = \frac{1}{a_{\mathbf{nj}} b_{\mathbf{n}}^d} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_2\left(\frac{\|t_{\mathbf{i}} - t_{\mathbf{j}}\|}{\rho_{\mathbf{n}}}\right)$$

Now, we can use the following decomposition

$$\begin{aligned}
&\sup_{\substack{x_{\mathbf{j}}\in R\\ \mathbf{j}\in V_R}} |f_{\mathbf{n}}(x_{\mathbf{j}}) - \mathbb{E}(f_{\mathbf{n}}(x_{\mathbf{j}}))| \\
&\le \max_{1\le k\le v} \sup_{\substack{x_{\mathbf{j}}\in R\\ \mathbf{j}\in V_R}} |f_{\mathbf{n}}(x_{\mathbf{j}}) - \widehat{f}(x_k) + \widehat{f}(x_k) - \mathbb{E}\widehat{f}(x_k) + \mathbb{E}\widehat{f}(x_k) - \mathbb{E}f_{\mathbf{n}}(x_{\mathbf{j}})| \\
&\le \max_{1\le k\le v} \sup_{\substack{x_{\mathbf{j}}\in R\\ \mathbf{j}\in V_R}} |f_{\mathbf{n}}(x_{\mathbf{j}}) - \widehat{f}(x_k)| + \max_{\substack{1\le k\le v\\ \mathbf{j}\in V_R}} |\widehat{f}(x_k) - \mathbb{E}\widehat{f}(x_k)| + \max_{1\le k\le v} \sup_{\substack{x_{\mathbf{j}}\in R\\ \mathbf{j}\in V_R}} |\mathbb{E}\widehat{f}(x_k) - \mathbb{E}f_{\mathbf{n}}(x_{\mathbf{j}})| \\
&\le A_1 + A_2 + A_3
\end{aligned}$$

Since the proofs of A_1 and A_3 are similar, we only give the proof of A_1.

$$\begin{aligned}
A_1 &= \max_{1\le k\le v} \sup_{\substack{x_{\mathbf{j}}\in R\\ \mathbf{j}\in V_R}} |f_{\mathbf{n}}(x_{\mathbf{j}}) - \widehat{f}(x_k)| \\
&= \max_{1\le k\le v} \sup_{\substack{x_{\mathbf{j}}\in R\\ \mathbf{j}\in V_R}} \left| \frac{1}{a_{\mathbf{nj}} b_{\mathbf{n}}^d} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_2\left(\frac{\|t_{\mathbf{i}} - t_{\mathbf{j}}\|}{\rho_{\mathbf{n}}}\right) \left[K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) - K_1\left(\frac{x_k - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right)\right]\right| \\
&\le \frac{C}{a_{\mathbf{nj}} b_{\mathbf{n}}^d} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_2\left(\frac{\|t_{\mathbf{i}} - t_{\mathbf{j}}\|}{\rho_{\mathbf{n}}}\right) \left\| \frac{x_{\mathbf{j}} - x_k}{b_{\mathbf{n}}}\right\| \qquad \text{by Lipschitz conditions on } K_1(\cdot) \\
&\le C b_{\mathbf{n}}^{-(d+1)} l
\end{aligned}$$

We choose $l = b_{\mathbf{n}}^{d+1}\rho_{\mathbf{n}}^N \epsilon$ and $v \le C l^{-d} \le C(b_{\mathbf{n}}^{d+1}\rho_{\mathbf{n}}^N\epsilon)^{-d}$; thus, we have

$$\begin{aligned}
A_1 &\le C b_{\mathbf{n}}^{-(d+1)} b_{\mathbf{n}}^{d+1} \rho_{\mathbf{n}}^N \epsilon \\
&\le C \rho_{\mathbf{n}}^N \epsilon \\
&\le C\epsilon \\
&= O\left(\sqrt{\frac{\log \widehat{\mathbf{n}}}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}}\right).
\end{aligned}$$

In the same way, we obtain that $A_3 = O\left(\sqrt{\frac{\log \widehat{\mathbf{n}}}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}}\right)$. It remains to show that $A_2 = \max_{\substack{1\le k\le v\\ \mathbf{j}\in V_R}} |\widehat{f}(x_k) - \mathbb{E}\widehat{f}(x_k)| = O\left(\left(\frac{\log \widehat{\mathbf{n}}}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}\right)^{1/2}\right)$. This term will be treated in the same manner as in the proof of Theorem 2.3 but using assumptions on $k_{\mathbf{n}1}$ and $k_{\mathbf{n}2}$. It is equivalent

to show that $\max_{\substack{1\le k\le v\\ \mathbf{j}\in V_R}} |T(\mathbf{n},x_k,i)| = O\left(\sqrt{\frac{\log\widehat{\mathbf{n}}}{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}}\right)$ for each $i=1,\dots,2^N$. Without loss of generality, we will consider the case $i=1$ and obtain as above

$$\mathbb{P}\left[\max_{\substack{1\le k\le v\\ \mathbf{j}\in V_R}} |T(\mathbf{n},x_j,1)|>\epsilon_{\mathbf{n}}\right] \le Cv\left(\widehat{\mathbf{n}}^{-a}+\psi(\widehat{\mathbf{n}},p^N)\frac{1}{b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\left(\frac{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}{\log\widehat{\mathbf{n}}}\right)^{\frac{N-\theta}{2N}}\right)$$

with $a=\frac{\delta^2}{2^{2N+4}C+2^{N+2}C\delta}$. We first study the convergence of the series $\sum_{\mathbf{n}\in\mathbb{N}^N} Cv\widehat{\mathbf{n}}^{-a}$. We have $v\le C(b_{\mathbf{n}}^{d+1}\rho_{\mathbf{n}}^N\epsilon)^{-d}$. The assumption $\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N\to\infty$ implies $\widehat{\mathbf{n}}>Cb_{\mathbf{n}}^{-d}\rho_{\mathbf{n}}^{-N}$ or $\widehat{\mathbf{n}}^{\frac{d}{2}}>b_{\mathbf{n}}^{\frac{-d^2}{2}}\rho_{\mathbf{n}}^{\frac{-dN}{2}}$. It also implies that $\widehat{\mathbf{n}}b_{\mathbf{n}}^d\to\infty$ and then $\widehat{\mathbf{n}}>Cb_{\mathbf{n}}^{-d}$.

$$\begin{aligned}
v\widehat{\mathbf{n}}^{-a}\le C(b_{\mathbf{n}}^{d+1}\rho_{\mathbf{n}}^N\epsilon)^{-d}\widehat{\mathbf{n}}^{-a} &\le Cb_{\mathbf{n}}^{-d}b_{\mathbf{n}}^{\frac{-d^2}{2}}\rho_{\mathbf{n}}^{\frac{-dN}{2}}\log\widehat{\mathbf{n}}^{\frac{-d}{2}}\widehat{\mathbf{n}}^{\frac{d}{2}-a}\\
&\le C\widehat{\mathbf{n}}\widehat{\mathbf{n}}^{\frac{d}{2}}\log\widehat{\mathbf{n}}^{\frac{-d}{2}}\widehat{\mathbf{n}}^{\frac{d}{2}-a}\\
&\le C\widehat{\mathbf{n}}^{d+1-a}\log\widehat{\mathbf{n}}^{\frac{-d}{2}}\\
&\le C\widehat{\mathbf{n}}^{d+1-a}
\end{aligned}$$

Consequently, the series $\sum_{\mathbf{n}\in\mathbb{N}^N} Cv\widehat{\mathbf{n}}^{-a}$ converges if and only if $a>d+1$. Now, we treat the second term for which two cases arise from assumptions on $\psi(n,m)$. In the situation of assumption **H7**, we have $\psi(n,m)\le C\min(n,m)$ and set

$$g_{\mathbf{n}}^\star = Cv\frac{\psi(\widehat{\mathbf{n}},p^N)}{b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\left(\frac{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}{\log\widehat{\mathbf{n}}}\right)^{\frac{N-\theta}{2N}}.$$

To show that $\sum_{\mathbf{n}\in\mathbb{N}^N} g_{\mathbf{n}}^\star<\infty$, let the following sequence of calculations

$$\begin{aligned}
g_{\mathbf{n}}^\star\widehat{\mathbf{n}}u_{\mathbf{n}} &\le Cv\left(\widehat{\mathbf{n}}^{\frac{4N-\theta}{2N}}b_{\mathbf{n}}^{\frac{-d\theta}{2N}}\rho_{\mathbf{n}}^{\frac{-N\theta}{2N}}\log\widehat{\mathbf{n}}^{\frac{\theta-2N}{2N}}u_{\mathbf{n}}\right)\\
&\le Cb_{\mathbf{n}}^{\frac{-d^2}{2}-d}\rho_{\mathbf{n}}^{\frac{-dN}{2}}\log\widehat{\mathbf{n}}^{\frac{-d}{2}}\widehat{\mathbf{n}}^{\frac{d}{2}}\widehat{\mathbf{n}}^{\frac{4N-\theta}{2N}}b_{\mathbf{n}}^{\frac{-d\theta}{2N}}\rho_{\mathbf{n}}^{\frac{-N\theta}{2N}}\log\widehat{\mathbf{n}}^{\frac{\theta-2N}{2N}}u_{\mathbf{n}}\\
&\le C\widehat{\mathbf{n}}^{\frac{N(d+4)-\theta}{2N}}b_{\mathbf{n}}^{\frac{-d(dN+2N+\theta)}{2N}}\rho_{\mathbf{n}}^{\frac{-N(dN+\theta)}{2N}}\log\widehat{\mathbf{n}}^{\frac{\theta-N(2+d)}{2N}}u_{\mathbf{n}}\\
&\le C\left(\widehat{\mathbf{n}}b_{\mathbf{n}}^{\frac{d(N(d+2)+\theta)}{\theta-N(d+4)}}\rho_{\mathbf{n}}^{\frac{N(dN+\theta)}{\theta-N(d+4)}}\log\widehat{\mathbf{n}}^{\frac{N(2+d)-\theta}{\theta-N(d+4)}}u_{\mathbf{n}}^{\frac{-2N}{\theta-N(d+4)}}\right)^{\frac{\theta-N(d+4)}{2N}}\\
&\le C(a_{\mathbf{n}}^\star)^{\frac{N(d+4)-\theta}{2N}}
\end{aligned}$$

Note that in the case where $\theta>N(d+4)$, one has $\frac{N(d+4)-\theta}{2N}<0$ and assumption **H7** leads to $a_{\mathbf{n}}^\star\to+\infty$, as $\mathbf{n}\to\infty$. So, $a_{\mathbf{n}}^{\star\frac{N(d+4)-\theta}{2N}}\to 0$ and $g_{\mathbf{n}}^\star\widehat{\mathbf{n}}u_{\mathbf{n}}\to 0$; consequently, $g_{\mathbf{n}}^\star<\frac{1}{\widehat{\mathbf{n}}u_{\mathbf{n}}}$ leads to $\sum_{\mathbf{n}\in\mathbb{N}^N} g_{\mathbf{n}}^\star<\sum_{\mathbf{n}\in\mathbb{N}^N}\frac{1}{\widehat{\mathbf{n}}u_{\mathbf{n}}}<+\infty$.
In the second situation, which corresponds to assumption **H8**, we have $\psi(n,m)\le C(n+m+1)^{\widetilde{\beta}}$. Let us note that $\psi(\widehat{\mathbf{n}},p^N)\le C(\widehat{\mathbf{n}}+p^N+1)^{\widetilde{\beta}}\le C\widehat{\mathbf{n}}^{\widetilde{\beta}}$ and set

$$h_{\mathbf{n}}^\star = Cv\psi(\widehat{\mathbf{n}},p^N)\frac{1}{b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\left(\frac{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}{\log\widehat{\mathbf{n}}}\right)^{\frac{N-\theta}{2N}}$$

To show that $\sum_{\mathbf{n}\in\mathbb{N}^N} h^\star_{\mathbf{n}} < \infty$, let the following sequence of calculations

$$
\begin{aligned}
h^\star_{\mathbf{n}}\widehat{\mathbf{n}}u_{\mathbf{n}} &\leq Cv\left(\widehat{\mathbf{n}}^{\frac{N(3+2\widetilde{\beta})-\theta}{2N}} b_{\mathbf{n}}^{\frac{-d(\theta+N)}{2N}} \rho_{\mathbf{n}}^{\frac{-N(N+\theta)}{2N}} \log\widehat{\mathbf{n}}^{\frac{\theta-N}{2N}} u_{\mathbf{n}}\right) \\
&\leq Cb_{\mathbf{n}}^{\frac{-d^2}{2}-d} \rho_{\mathbf{n}}^{\frac{-dN}{2}} \log\widehat{\mathbf{n}}^{\frac{-d}{2}} \widehat{\mathbf{n}}^{\frac{d}{2}} \widehat{\mathbf{n}}^{\frac{N(3+2\widetilde{\beta})-\theta}{2N}} b_{\mathbf{n}}^{\frac{-d(\theta+N)}{2N}} \rho_{\mathbf{n}}^{\frac{-N(N+\theta)}{2N}} \log\widehat{\mathbf{n}}^{\frac{\theta-N}{2N}} u_{\mathbf{n}} \\
&\leq Cb_{\mathbf{n}}^{\frac{-d(N(3+d)+\theta)}{2N}} \rho_{\mathbf{n}}^{\frac{-N(N(d+1)+\theta)}{2N}} \log\widehat{\mathbf{n}}^{\frac{\theta-N(d+1)}{2N}} \widehat{\mathbf{n}}^{\frac{N(3+2\widetilde{\beta}+d)-\theta}{2N}} u_{\mathbf{n}} \\
&\leq C\left(\widehat{\mathbf{n}} b_{\mathbf{n}}^{\frac{d(N(3+d)+\theta)}{\theta-N(3+2\widetilde{\beta}+d)}} \rho_{\mathbf{n}}^{\frac{N(N(d+1)+\theta)}{\theta-N(3+2\widetilde{\beta}+d)}} \log\widehat{\mathbf{n}}^{\frac{N(d+1)-\theta}{\theta-N(3+2\widetilde{\beta}+d)}} u_{\mathbf{n}}^{\frac{-2N}{\theta-N(3+2\widetilde{\beta}+d)}}\right)^{\frac{\theta-N(3+2\widetilde{\beta}+d)}{2N}} \\
&\leq C\left(h^\star_{\mathbf{n}}\right)^{\frac{\theta-N(3+2\widetilde{\beta}+d)}{2N}}
\end{aligned}
$$

Now, since in this case $\theta > N(3+2\widetilde{\beta}+d)$, we have $\frac{N(3+2\widetilde{\beta}+d)-\theta}{2N} < 0$ and assumption **H8** allows us to say that $d^\star_{\mathbf{n}} \to +\infty$, as $\mathbf{n}\to\infty$. Moreover, $d^{\star\,\frac{N(3+2\widetilde{\beta}+d)-\theta}{2N}}_{\mathbf{n}} \to 0$ then $\widehat{\mathbf{n}}u_{\mathbf{n}}h^\star_{\mathbf{n}} \to 0$. Hence, $h^\star_{\mathbf{n}} < \dfrac{1}{\widehat{\mathbf{n}}u_{\mathbf{n}}} \Leftrightarrow \displaystyle\sum_{\mathbf{n}\in\mathbb{N}^N} h^\star_{\mathbf{n}} < \sum_{\mathbf{n}\in\mathbb{N}^N} \frac{1}{\widehat{\mathbf{n}}u_{\mathbf{n}}} < +\infty$. Therefore, $A_2 = O\left(\sqrt{\dfrac{\log\widehat{\mathbf{n}}}{\widehat{\mathbf{n}}b^d_{\mathbf{n}}\rho^N_{\mathbf{n}}}}\right)$. To conclude,

$$
\sup_{\substack{x_{\mathbf{j}}\in R\\ \mathbf{j}\in V_R}} \left|f_{\mathbf{n}}(x_{\mathbf{j}}) - f(x_{\mathbf{j}})\right| = O\left(\left(\frac{\log\widehat{\mathbf{n}}}{\widehat{\mathbf{n}}b^d_{\mathbf{n}}\rho^N_{\mathbf{n}}}\right)^{1/2}\right) + O(b_{\mathbf{n}}),\ a.s.
$$

■

Chapter 3

Spatial regression estimation, prediction and illustration

Contents

Résumé en français

Dans le chapitre 2 qui précède, nous nous sommes intéressés à l'estimation non paramétrique de la fonction de densité spatiale. Nous avons proposé une nouvelle approche de l'estimateur à noyau qui permet de tenir compte à la fois de la distance entre les observations et de celle entre les sites. Dans ce chapitre, nous adaptons cette approche pour l'estimation de la fonction de régression spatiale dans l'objectif de faire de la prévision spatiale. Notons que le problème d'estimation à noyau de la fonction de régression pour données spatialisées a déjà été étudié dans la littérature, par exemple, dans Biau et Cadre (2004), Menezes et al. (2010). Cependant, l'approche proposée dans ce chapitre est différente.

Nous considérons un processus spatial $(Z_{\mathbf{i}} = (X_{\mathbf{i}}, Y_{\mathbf{i}}) \in \mathbb{R}^d \times \mathbb{R}, \mathbf{i} \in \mathbb{Z}^N)$ défini sur l'espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ de même distribution que le processus (X, Y) de densité inconnue $f_{X,Y}$ sur $\mathbb{R}^{d+1}$. La fonction de densité de X sur $\mathbb{R}^d$ est $f(\cdot)$. On s'intéresse au modèle de régression défini par $Y_{\mathbf{i}} = r(X_{\mathbf{i}}) + \epsilon_{\mathbf{i}}$ où $r(x) = \mathbb{E}(Y|X = x)$, le bruit $\epsilon_{\mathbf{i}}$ est centré, α-mélangeant et indépendant des $X_{\mathbf{i}}$. Notre objectif est d'estimer la fonction de régression $r(\cdot)$ définie par $r(x) = \varphi(x)/f(x)$ où $\varphi(x) = \int y f_{XY}(x, y)dy$, $x \in \mathbb{R}^d$. On suppose que le processus étudié est observé sur l'ensemble spatial discret

$\mathcal{I}_{\mathbf{n}} = \{\mathbf{i} = (i_1, \ldots, i_N), 1 \leq i_k \leq n_k,\ k = 1, \ldots, N\}$ où un point $\mathbf{i} = (i_1, \ldots, i_N) \in \mathbb{Z}^N$ fait référence à un site. On note $\mathbf{n} = (n_1, \ldots, n_N)$ et on pose $\widehat{\mathbf{n}} := n_1 \times \ldots \times n_N$. Par souci de simplicité, on suppose que $n_1 = n_2 = \ldots = n_N = n$ (e.g. El Machkouri (2007, 2011), El Machkouri and Stoica (2010)), mais les résultats suivants peuvent être étendus à un cadre plus général. On écrit $\mathbf{n} \to \infty$ si $n \to \infty$. La dépendance spatiale implique le besoin de déterminer quelles autres unités de $\mathcal{I}_{\mathbf{n}}$ ont une influence sur le site considéré. Pour cela, pour chaque site $\mathbf{j}$, on pose $k_{\mathbf{n}} = k_{\mathbf{n},\mathbf{j}} = \sum_{\mathbf{i}} 1_{[\|\mathbf{i}-\mathbf{j}\| \leq d_{\mathbf{n}}]}$ le nombre de voisins $\mathbf{i}$ pour lequel la distance entre $\mathbf{i}$ et $\mathbf{j}$ est inférieure ou égale à la distance $d_{\mathbf{n}} > 0$ telle que $d_{\mathbf{n}} \to \infty$ lorsque $\mathbf{n} \to \infty$. Soit $d_{\mathbf{n}} = n\rho_{\mathbf{n}}$ et $k_{\mathbf{n}} = O(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N)$. On construit un nouvel estimateur à noyau avec des poids sur les sites qui ont plus d'importance lorsque la distance entre les sites est petite. Nous nous intéressons à l'estimation de la fonction de régression $r(\cdot)$, et en particulier à la prédiction de $Y_{\mathbf{j}}$ sous la condition que $X_{\mathbf{j}} = x$ (comme dans Wang et al. (2012)), noté $x_{\mathbf{j}}$ dans la suite. En considérant les sites normalisés, l'estimateur à noyau de la fonction de régression $r(\cdot)$ en ce point est défini par

$$r_{\mathbf{n}}(x_{\mathbf{j}}) = \begin{cases} \dfrac{\varphi_{\mathbf{n}}(x_{\mathbf{j}})}{f_{\mathbf{n}}(x_{\mathbf{j}})} & \text{si} \quad f_{\mathbf{n}}(x_{\mathbf{j}}) \neq 0; \\ \overline{Y} & \text{sinon,} \end{cases}$$

où $\overline{Y}$ est la moyenne empirique des $Y_{\mathbf{i}}$, $\varphi_{\mathbf{n}}(x_{\mathbf{j}})$ et $f_{\mathbf{n}}(x_{\mathbf{j}})$ sont définis par

$$\varphi_{\mathbf{n}}(x_{\mathbf{j}}) = \frac{1}{a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j} - \mathbf{i}\|),$$

$$f_{\mathbf{n}}(x_{\mathbf{j}}) = \frac{1}{a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j} - \mathbf{i}\|),$$

avec $a_{\mathbf{n},\mathbf{j}} = \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j}-\mathbf{i}\|)$ et $K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j}-\mathbf{i}\|) = K_2\left(\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{j}-\mathbf{i}}{\mathbf{n}}\right\|\right)$, (où $\frac{\mathbf{i}}{\mathbf{n}} = \left(\frac{i_1}{n}, \frac{i_2}{n}, \ldots, \frac{i_N}{n}\right)$), K_1 et K_2 sont des noyaux respectivement définis sur $\mathbb{R}^d$ et $\mathbb{R}$, $b_{\mathbf{n}}$ et $\rho_{\mathbf{n}}$ sont des paramètres de fenêtres qui tendent vers zéro, telles que $\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N \to \infty$.

Plus généralement, soit $(X_{\mathbf{i}}, Y_{\mathbf{i}})$, $\mathbf{i} \in \mathcal{I}_{\mathbf{n}}$, un processus spatial strictement stationnaire et soit $\mathbf{i_0}$ un site n'appartenant pas à $\mathcal{I}_{\mathbf{n}}$, on peut étendre les estimateurs $\varphi_{\mathbf{n}}(\cdot)$ et $f_{\mathbf{n}}(\cdot)$ de la manière suivante

$$\widehat{\varphi}_{\mathbf{n}}(x_{\mathbf{i_0}}) = \frac{1}{a^{\star}_{\mathbf{n},\mathbf{i_0}} b_{\mathbf{n}}^d} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} K_1\left(\frac{x_{\mathbf{i_0}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_2\left(\frac{\mathbf{i} - \mathbf{i_0}}{\rho_{\mathbf{n}}}\right)$$

$$\widehat{f}_{\mathbf{n}}(x_{\mathbf{i_0}}) = \frac{1}{a^{\star}_{\mathbf{n},\mathbf{i_0}} b_{\mathbf{n}}^d} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K_1\left(\frac{x_{\mathbf{i_0}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_2\left(\frac{\mathbf{i} - \mathbf{i_0}}{\rho_{\mathbf{n}}}\right)$$

où les sites $\mathbf{i}$ et $\mathbf{i_0}$ ne sont pas normalisés (voir e.g. Wang et al. (2012)), $a^{\star}_{\mathbf{n},\mathbf{i_0}} = \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K_2\left(\frac{\mathbf{i}-\mathbf{i_0}}{\rho_{\mathbf{n}}}\right)$ et $K_2(\cdot)$ est un noyau sur $\mathbb{R}^N$.

La suite de ce chapitre est composée d'une introduction au cadre de travail qui nous intéresse. L'estimateur de la régression $r_{\mathbf{n}}(\cdot)$ présenté ci-dessus est étudié plus en détails. Après avoir donné les hypothèses, des résultats de convergence sont obtenus. Plus particulièrement, les convergences presque complète (p.c.) et en moyenne d'ordre q ($q \in \mathbb{R}^*$) sont obtenues et des vitesses de convergence sont atteintes. Les résultats sont les suivants

$$|r_{\mathbf{n}}(x_{\mathbf{j}}) - r(x_{\mathbf{j}})| = O\left(b_{\mathbf{n}} + \sqrt{\frac{\log \widehat{\mathbf{n}}}{\widehat{\mathbf{n}}\, b_{\mathbf{n}}^d\, \rho_{\mathbf{n}}^N}}\right) \qquad \text{p.c.}$$

et

$$\|r_{\mathbf{n}}(x_{\mathbf{j}}) - r(x_{\mathbf{j}})\|_q = O\left(b_{\mathbf{n}} + \sqrt{\frac{1}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}}\right),\ q > 1.$$

Puis un prédicteur spatial est déduit de l'estimateur de la fonction de régression. Avant de conclure, des résultats numériques sont présentés et montrent que notre approche permet d'améliorer les résultats obtenus avec l'estimateur à noyau classique, en particulier lorsque la dépendance spatiale est importante. Il s'agit de résultats issus d'une étude de simulations ainsi que de l'application à des données environnementales liées à la concentration en métaux lourds dans le sol.

Les résultats présentés dans ce chapitre ont été obtenus avec la collaboration de Sophie Dabo-Niang (Université Charles de Gaulle) et Anne-Françoise Yao (Université Blaise Pascal).

3.1 Introduction

Spatial statistics deal with the problem, amongst others, of reconstructing a phenomenon over its domain from a discrete set of observed values. This problem has applications in many subject areas such as in soil sciences, oceanology, epidemiology, climatology and many others where data are available at specific spatial locations. More precisely, the goal is to predict unsampled locations by taking into account spatial dependence. During the first half of the twentieth century, spatial prediction is studied in the scope of geostatistics, commonly known as *kriging*. The latter is a spatial interpolation method, allowing a linear estimate, based on mean and variance of the data. Since their apparition, these parametric methods have been widely studied in the literature, (see Cressie (1993), Wackernagel (2003) for an introduction). However, a preselected parametric model might be too restricted or too low-dimensional to fit unexpected features. In response to that, nonparametric approaches have sometimes been suggested as an alternative. Consequently, nowadays, a dynamic concerns the deployment of nonparametric methods to spatial statistics including prediction methods. The first results in this direction are those of Biau and Cadre (2004) and concerned the kernel prediction of a strictly stationary random field indexed in $(\mathbb{N}^*)^N$. Later, Dabo-Niang and Yao (2007) contribute to Biau and Cadre (2004)'s investigations since they are interested in the kernel regression estimation and prediction of continuously indexed random fields. In Menezes et al. (2010), nonparametric kernel prediction is considered for spatial stochastic processes when a stochastic sampling design is assumed for selection of random locations. The spatial predictors presented in these three works dealt with the kernel method. A main difference is that the last is based on a kernel that controls the distance between sites contrary to the others that deal with a kernel on the values of the field. In this work, our first interest lies in proposing a nonparametric regression estimation approach which then will be used for the purpose of prediction. The originality of the suggested regression estimator is to take advantages of each estimator introduced previously. In fact, our estimator depends on two kernels, one of which controls the distance between observations and the other controls the spatial dependence structure. The advantage of the proposed estimate is to take directly into account the spatial dependency in its form, that is particularly interesting in a prevision context. This idea has been presented in Dabo-Niang et al. (2014a) in the context of density estimation and in Ternynck (2014) to deal with a regression problem for functional data. Note that Wang et al. (2012) proposed a local linear spatio-temporal prediction model, using a kernel weight function taking into account the distance between sites. The specificity of

the prediction procedure of Wang et al. (2012) is to be based on the assumption that the error term of the model is autocorrelated. In the present chapter, our regression method is more general since there are no additional assumption on the error term.

Nonparametric spatial regression estimation has received a great deal of attentional from the scientific community. For instance, the works of Biau and Cadre (2004), Hallin et al. (2004b), Carbon et al. (2007), Dabo-Niang and Yao (2007), Guillas and Lai (2010), Attouch et al. (2011), Dabo-Niang et al. (2011b), Robinson (2011) and Ternynck (2014) are devoted to this matter. Besides, other authors dealt with the spatial quantile regression estimation (e.g. Hallin et al. (2009) and Dabo-Niang et al. (2012b)) or nonparametric conditional mode estimation (e.g. Dabo-Niang et al. (2014b)).

The chapter is structured as follows. In Section 3.2, after providing some notations, a kernel estimate of the regression function is introduced. Section 3.3 describes the assumptions and gives the related asymptotic results. More specifically, almost complete convergence and consistency in L^q norm ($q \in \mathbb{N}^*$) with rates of the kernel estimate are achieved when the considered sample is an α-mixing sequence. A spatial predictor is then deduced. Some numerical investigations are presented in Section 3.4, which are based on simulation studies but also on an environmental data set. The last section is devoted to conclusions whereas proofs and technical results are given in Appendix (see Section 3.6).

3.2 Kernel spatial estimator of the regression function

In the following, $\|\cdot\|$ will denote any norm in $\mathbb{R}^d$ or $\mathbb{R}^N$ (there will be no ambiguity since the vectors of $\mathbb{R}^N$ are in bold), C and C' will indicate some arbitrary positive constants that may vary from line to line, for each real u, $\lfloor u \rfloor$ will indicate the integer part of u. Moreover, we write $u_{\mathbf{n}} = O(v_{\mathbf{n}})$ means that $\exists C$ such that $|u_{\mathbf{n}}|/v_{\mathbf{n}} \leq C$ as $v_{\mathbf{n}} \to \infty$ and $u_{\mathbf{n}} = o(v_{\mathbf{n}})$ means that $|u_{\mathbf{n}}|/v_{\mathbf{n}} \to 0$ as $v_{\mathbf{n}} \to \infty$.

We consider a spatial process $(Z_{\mathbf{i}} = (X_{\mathbf{i}}, Y_{\mathbf{i}}) \in \mathbb{R}^d \times \mathbb{R}, \mathbf{i} \in \mathbb{Z}^N)$ defined over some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with same distribution as (X, Y) having unknown density $f_{X,Y}$ on $\mathbb{R}^{d+1}$. The density function of X on $\mathbb{R}^d$ is $f(\cdot)$. For a seek of simplicity, we will suppose that the variable Y is bounded. In this chapter, we are interested in the following regression model $Y_{\mathbf{i}} = r(X_{\mathbf{i}}) + \varepsilon_{\mathbf{i}}$, where $r(x) = \mathbb{E}(Y|X = x)$ is an unknown function, with real values, defined by $r(x) = \varphi(x)/f(x)$ where $\varphi(x) = \int y f_{XY}(x, y)dy$, $x \in \mathbb{R}^d$, $(\varepsilon_{\mathbf{i}})_{\mathbf{i}\in\mathbb{Z}^N}$ is a centered spatial process independent of $(X_{\mathbf{i}})_{\mathbf{i}\in\mathbb{Z}^N}$. Then, the main goal of this chapter is to estimate the regression function $r(\cdot)$. As it is classically assumed in the literature, the process under study $(Z_{\mathbf{i}})$ is observed over the rectangular domain $\mathcal{I}_{\mathbf{n}} = \{\mathbf{i}, 1 \leq i_k \leq n_k, k = 1, \ldots, N\}$ where a point $\mathbf{i} = (i_1, \ldots, i_N) \in \mathbb{Z}^N$ refers to a site. We denote $\mathbf{n} = (n_1, \ldots, n_N)$ and let $\widehat{\mathbf{n}} := n_1 \times \ldots \times n_N$ be the sample size.

From now on, we assume for simplicity that $n_1 = n_2 = \ldots = n_N = n$ (e.g. El Machkouri (2007, 2011), El Machkouri and Stoica (2010)), but the following results can be extended to a more general framework. We write $\mathbf{n} \to \infty$ if $n \to \infty$. The spatial dependence implies the need to determine which other units in $\mathcal{I}_{\mathbf{n}}$ have an influence on a considered location. Thus, for each site $\mathbf{j}$, let $k_{\mathbf{n}} = k_{\mathbf{n},\mathbf{j}} = \sum_{\mathbf{i}} 1_{[\|\mathbf{i}-\mathbf{j}\| \leq d_{\mathbf{n}}]}$ denote the number of neighbors $\mathbf{i}$ for which the distance between $\mathbf{i}$ and $\mathbf{j}$ is less than or equal to distance $d_{\mathbf{n}} > 0$ such that $d_{\mathbf{n}} \to \infty$ as $\mathbf{n} \to \infty$. This last assumes the correlation between locations (eventually) increases as the sample size increases. Taking the Euclidean distance and if $N = 2$ (square grid), we have $k_{\mathbf{n}} \leq 4d_{\mathbf{n}}^2 - 4d_{\mathbf{n}} + 4$ which leads to $k_{\mathbf{n}} = O(d_{\mathbf{n}}^2)$ and $k_{\mathbf{n}} = o(d_{\mathbf{n}}^{\eta}), \eta > 2$. Moreover, if $d_{\mathbf{n}} = o(\widehat{\mathbf{n}}^{\epsilon})$, $0 < \epsilon < 1$ then $k_{\mathbf{n}} = o(\widehat{\mathbf{n}}^{2\epsilon})$, see for instance Kelejian and Prucha (2007). Let $d_{\mathbf{n}} = n\rho_{\mathbf{n}}$; consequently, we have $d_{\mathbf{n}}^2 = \widehat{\mathbf{n}}\rho_{\mathbf{n}}^N$ and $k_{\mathbf{n}} = O(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N)$ as well as $k_{\mathbf{n}} = o((\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N)^{\eta/2})$, $\eta > 2$. Using this, we construct a spatial kernel estimator with

weight on the sites. The weights are assumed to decline as a measure of distance between corresponding sites (that are normalized) increases. We are interested in the regression estimation of $r(\cdot)$, in particular the prediction of $Y_{\mathbf{j}}$ under the condition that $X_{\mathbf{j}} = x$ (as in Wang et al. (2012)), denoted in what follows $x_{\mathbf{j}}$, related to the concerned location $\mathbf{j}$. Considering normalized sites, the kernel estimator of the regression function $r(\cdot)$ at this point is defined as

$$r_{\mathbf{n}}(x_{\mathbf{j}}) = \begin{cases} \dfrac{\varphi_{\mathbf{n}}(x_{\mathbf{j}})}{f_{\mathbf{n}}(x_{\mathbf{j}})} & \text{if} \quad f_{\mathbf{n}}(x_{\mathbf{j}}) \neq 0; \\ \overline{Y} & \text{otherwise,} \end{cases} \tag{3.1}$$

where $\overline{Y}$ denotes the empirical mean of the $Y_{\mathbf{i}}$, the functions $\varphi_{\mathbf{n}}(x_{\mathbf{j}})$ and $f_{\mathbf{n}}(x_{\mathbf{j}})$ are defined by

$$\varphi_{\mathbf{n}}(x_{\mathbf{j}}) = \frac{1}{a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j}-\mathbf{i}\|),$$
$$f_{\mathbf{n}}(x_{\mathbf{j}}) = \frac{1}{a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j}-\mathbf{i}\|),$$

with $a_{\mathbf{n},\mathbf{j}} = \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j}-\mathbf{i}\|)$ and $K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j}-\mathbf{i}\|) = K_2\left(\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{j}-\mathbf{i}}{\mathbf{n}}\right\|\right)$, (where $\frac{\mathbf{i}}{\mathbf{n}} = \left(\frac{i_1}{n}, \frac{i_2}{n}, \ldots, \frac{i_N}{n}\right)$), K_1 and K_2 are kernels respectively defined on $\mathbb{R}^d$ and $\mathbb{R}$, $b_{\mathbf{n}}$ and $\rho_{\mathbf{n}}$ are bandwidths tending to zero, such that $\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N \to \infty$. Then we have $\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{j}-\mathbf{i}}{\mathbf{n}}\right\| \leq 1$ means that $\|\mathbf{j}-\mathbf{i}\| \leq n\rho_{\mathbf{n}}$. The estimator $f_{\mathbf{n}}(x_{\mathbf{j}})$ is a function of the number $k_{\mathbf{n}}$ of neighbors $\mathbf{i}$ for which the distance $d_{\mathbf{n}}$ is chosen hereafter to be $n\rho_{\mathbf{n}}$, with $k_{\mathbf{n}} \to +\infty$, $k_{\mathbf{n}} = O(d_{\mathbf{n}}^N) = O(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N)$. If one assumes that $d_{\mathbf{n}} = o(\widehat{\mathbf{n}}^{\epsilon})$, $\epsilon \in (0,1)$, then $k_{\mathbf{n}}$ can be expressed in terms of $\widehat{\mathbf{n}}$. We notice that the kernel K_2 is here to handle the nearness between locations. In what follows, we assume that $k_{\mathbf{n}} = C_N d_{\mathbf{n}}^N + O(d_{\mathbf{n}}^{\beta})$ as $d_{\mathbf{n}} \to +\infty$, $0 < \beta < N$ and C_N is a constant that depends on N. In the literature, some works (e.g. Biau and Cadre (2004), Menezes et al. (2010), Wang et al. (2012)) deal with other estimates of the kernel regression function for spatial data, see Section 3.3.3 for more details.

Remark 3.1.

- Instead of the previous functions $\varphi_{\mathbf{n}}$, $f_{\mathbf{n}}$, one can consider simpler following versions

$$\varphi_{\mathbf{n}}(x_{\mathbf{j}}) = \frac{1}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j}-\mathbf{i}\|),$$
$$f_{\mathbf{n}}(x_{\mathbf{j}}) = \frac{1}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j}-\mathbf{i}\|),$$

  In this case, properties of Riemann sums of Lebesgue integrable functions and additional conditions on K_2 lead to

$$\lim_{\mathbf{n}\to\infty} \frac{1}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j}-\mathbf{i}\|) \longrightarrow \int_{\mathbb{R}^N} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j}-\mathbf{i}\|) d\mathbf{i} = 1.$$

  This permits to control in a simple manner the bias and variance terms of the above estimates.

- More generally, let $(X_{\mathbf{i}}, Y_{\mathbf{i}})$, $\mathbf{i} \in \mathcal{I}_{\mathbf{n}}$, a strictly stationary spatial process and let $\mathbf{i_0}$ a site that does not belong to $\mathcal{I}_{\mathbf{n}}$, one can extend $\varphi_{\mathbf{n}}(\cdot)$ and $f_{\mathbf{n}}(\cdot)$ estimates in the

following way

$$\widehat{\varphi}_{\mathbf{n}}(x_{\mathbf{i_0}}) = \frac{1}{a^{\star}_{\mathbf{n},\mathbf{i_0}} b^d_{\mathbf{n}}} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} K_1\left(\frac{x_{\mathbf{i_0}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_2\left(\frac{\mathbf{i}-\mathbf{i_0}}{\rho_{\mathbf{n}}}\right)$$
$$\widehat{f}_{\mathbf{n}}(x_{\mathbf{i_0}}) = \frac{1}{a^{\star}_{\mathbf{n},\mathbf{i_0}} b^d_{\mathbf{n}}} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_1\left(\frac{x_{\mathbf{i_0}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_2\left(\frac{\mathbf{i}-\mathbf{i_0}}{\rho_{\mathbf{n}}}\right)$$

where sites $\mathbf{i}$ and $\mathbf{i_0}$ are not normalized (see e.g. Wang et al. (2012)), $a^{\star}_{\mathbf{n},\mathbf{i_0}} = \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_2\left(\frac{\mathbf{i}-\mathbf{i_0}}{\rho_{\mathbf{n}}}\right)$ and $K_2(\cdot)$ is a kernel on $\mathbb{R}^N$.

- To give some examples where the assumption on $k_{\mathbf{n}}$ is reasonable, consider $q_{\mathbf{n}}$ the number of standard lattice (in $\mathbb{Z}^N$) points contained in a closed ball $\mathcal{B}(\mathbf{j}, d_{\mathbf{n}})$ with center $\mathbf{j}$ and radius $d_{\mathbf{n}}$ that is $q_{\mathbf{n}} = \text{Card}\{\mathbf{i}\in\mathbb{R}^N,\ \|\mathbf{i}-\mathbf{j}\| \le d_{\mathbf{n}}\}$ where $\mathbf{j}$ is any vector of $\mathbb{R}^N$. It is well known that

$$q_{\mathbf{n}} = \frac{\pi^{N/2}}{\Gamma(N/2+1)} d^N_{\mathbf{n}} + O(d^{N-1}_{\mathbf{n}}),$$

where $\Gamma(\cdot)$ is the gamma function; see, for instance, Mitchell (1966), Chamizo and Iwaniec (1995), Tsang (2000) and Meyer (2011). And notice that $k_{\mathbf{n}} = C_N q_{\mathbf{n}}$. In particular, if $N = 2$ we have $q_{\mathbf{n}} = \frac{\pi}{\Gamma(2)} d^2_{\mathbf{n}} + O(d_{\mathbf{n}})$, $q_{\mathbf{n}} = \frac{\pi}{\Gamma(2)} d^2_{\mathbf{n}} + o(d^{2/3}_{\mathbf{n}})$.

- In the following, we consider pointwise (for a fixed $x_{\mathbf{j}}$) convergence result of the regression estimate but one can extend the obtained results to uniform ones, on a set where corresponding sites are in a set S (that can be a subset of $\mathcal{I}_{\mathbf{n}}$ or a set larger than $\mathcal{I}_{\mathbf{n}}$) by considering $K_{\mathbf{n}} = \sup_{\mathbf{j}\in S} k_{\mathbf{n},\mathbf{j}}$ instead of $k_{\mathbf{n}} = k_{\mathbf{n},\mathbf{j}}$.
- Note that although the estimate $r_{\mathbf{n}}(\cdot)$ takes into account the spatial dependency, we do not measure this dependency. However, before using this estimate rather than the classical estimate, one could evaluate the importance of the dependence, for instance, in fitting a variogram (e.g. Gaetan and Guyon (2008)) on the data to be processed. Indeed the more higher the range parameter of the variogram is, the greater the spatial dependence.

3.3 Assumptions and main results

3.3.1 Dependency conditions

To account for spatial dependency, we assume that the process $(Z_{\mathbf{i}})$ satisfies a mixing condition defined, for example, in Carbon et al. (1997) as follows: there exists a function $\varphi(x) \searrow 0$ as $x \to \infty$, such that

$$\begin{aligned}\alpha\left(\sigma\left(S\right), \sigma\left(S'\right)\right) &= \sup\left\{\left|\mathbb{P}(A\cap B) - \mathbb{P}\left(A\right)\mathbb{P}\left(B\right)\right|,\quad A\in\sigma\left(S\right),\quad B\in\sigma\left(S'\right)\right\}\\ &\le \psi\left(\text{Card}(S), \text{Card}(S')\right)\varphi\left(\text{dist}(S,S')\right)\end{aligned}$$

where S and S' are two finite sets of sites, $\text{Card}(S)$ denotes the cardinality of the set S, $\sigma\left(S\right) = \{Z_{\mathbf{i}}, \mathbf{i}\in S\}$ and $\sigma\left(S'\right) = \{Z_{\mathbf{i}}, \mathbf{i}\in S'\}$ are σ-fields generated by $Z_{\mathbf{i}}$, $\text{dist}(S,S')$ is the Euclidean distance between S and S', and $\psi(\cdot)$ is a positive symmetric function nondecreasing in each variable. We recall that the process $(Z_{\mathbf{i}})$ is said to be strongly mixing if $\psi \equiv 1$. As usual, we will assume that one of both conditions on $\varphi(i)$ is verified:

$$\varphi\left(i\right) \le Ci^{-\theta}, \qquad \text{for some } \theta > 0$$

i.e. that $\varphi(i)$ tends to zero at a polynomial rate, or

$$\varphi(i) \leq C \exp(-si), \qquad \text{for some } s > 0$$

i.e. that $\varphi(i)$ tends to zero at an exponential rate. These conditions are satisfied, for instance, by several kernels with compact support such as triangular (Bartlett), biweight, circular (cosine), Epanechnikov, Parzen, Tukey-Hanning kernels. For all compact support kernels, both conditions are verified at least asymptotically. That is sufficient to ensure the control of the mixing condition. Concerning the function $\varphi(\cdot)$, for the sake of simplicity, we will only study the case where $\varphi(\cdot)$ tends to zero at a polynomial rate, that is $\varphi(i) \leq Ci^{-\theta}$. However, similar results to that of Theorems 3.3 and 3.4 (below) could be obtained with $\varphi(\cdot)$ tending to zero at an exponential rate (which implies the polynomial case).

Let $u_{\mathbf{n}} = \prod_{i=1}^{N} (\log n_i)(\log\log n_i)^{1+\epsilon}$, then $\sum_{\mathbf{n}\in\mathbb{N}^N} \frac{1}{\widehat{\mathbf{n}} u_{\mathbf{n}}} < \infty$.

3.3.2 General assumptions and results

The consistency results of $r_{\mathbf{n}}$ are achieved under the following assumptions on f, r, the kernels, bandwidths and local dependence condition.

H1: The density functions $f_{X,Y}$ and f are continuous on $\mathbb{R}^{d+1}$ and $\mathbb{R}^d$ respectively.

H2: The density function f and the regression function r are Lipschitzian.

H3: The functions $K_1(\cdot)$ and $K_2(\cdot)$ are bounded integrable kernels on $\mathbb{R}$. Moreover, the kernel $K_1(\cdot)$ satisfy some Lipschitz conditions.

H4: There exist some constants C_{1i} and C_{2i} with $0 < C_{1i} < C_{2i} < \infty$, for $i = 1, 2$, such that

$$C_{11}\mathbf{1}_{[0,1]}(s's) \leq K_1(s) \leq C_{21}\mathbf{1}_{[0,1]}(s's) \qquad \text{for } s \in \mathbb{R}^d$$
$$C_{12}\mathbf{1}_{[0,1]}(t) \leq K_2(t) \leq C_{22}\mathbf{1}_{[0,1]}(t) \qquad \text{for } t \in \mathbb{R}$$

where s' denotes the transpose of s.

H5: **Local dependence condition.** The joint probability density $f_{X_{\mathbf{i}},X_{\mathbf{j}}}$ of $(X_{\mathbf{i}}, X_{\mathbf{j}})$ exists, is bounded and $\forall u, v \in \mathbb{R}^d$, for some constant $C > 0$, verifies $|f_{X_{\mathbf{i}},X_{\mathbf{j}}}(u, v) - f_{X_{\mathbf{i}}}(u) f_{X_{\mathbf{j}}}(v)| < C$.

H6: $\psi(n, m) \leq C \min(n, m)$ and $\widehat{\mathbf{n}} b_{\mathbf{n}}^{d\theta_1} \rho_{\mathbf{n}}^{N\theta_1} \log \widehat{\mathbf{n}}^{\theta_2} u_{\mathbf{n}}^{\theta_3} \to \infty$ with the mixing coefficient $\theta > N(q + 2)$, $q > 1$ and

$$\theta_1 = \frac{qN - \theta}{N(q + 2) - \theta}; \quad \theta_2 = \frac{\theta - 2N}{N(q + 2) - \theta}; \quad \theta_3 = \frac{2N}{N(q + 2) - \theta}.$$

H7: $\psi(n, m) \leq C(n + m + 1)^{\widetilde{\beta}}$ and $\widehat{\mathbf{n}} b_{\mathbf{n}}^{d\theta'_1} \rho_{\mathbf{n}}^{N\theta'_1} \log \widehat{\mathbf{n}}^{\theta'_2} u_{\mathbf{n}}^{\theta'_3} \to \infty$ with the mixing coefficient $\theta > N(q + 2\tilde{\beta} + 1)$, $q > 1$, $\widetilde{\beta} > 1$ and

$$\theta'_1 = \frac{N(q - 1) - \theta}{N(q + 2\tilde{\beta} + 1) - \theta}; \quad \theta'_2 = \frac{\theta - N}{N(q + 2\tilde{\beta} + 1) - \theta}; \quad \theta'_3 = \frac{2N}{N(q + 2\tilde{\beta} + 1) - \theta}.$$

Remark 3.2. These assumptions are very standard in the context of spatial nonparametric modeling. Indeed, the assumptions **H2** and **H3** allow to control the bias of the estimator. The Lipschitz condition **H2** allows the precise rate of convergence to be found whereas a continuity-type model would give only convergence results. Assumption **H4** is imposed for the sake of simplicity and brevity of the proofs. The local dependence condition **H5** is a classical condition in kernel estimation based on dependent data (see, e.g., Bosq (1998) or Carbon et al. (1997)). This assumption controls the local dependence whereas the mixing condition controls the dependence of sites which are far from each other. The assumptions **H6** and **H7** are classical technical assumptions, which appear (in the calculations when studying the asymptotic behavior of the estimator) in the particular case where the mixing coefficient is such that φ tends to zero at a polynomial rate (see Neaderhouser (1980) and Rosenblatt (1985) for some examples). Each of these conditions is related to a specific case of mixing in the spatial context and are used respectively in Neaderhouser (1980) and Takahata (1983).

The two following theorems give some results about the consistency of the estimator proposed for the regression function. The almost complete convergence and the mean of order q consistency are obtained as well as some rates of consistency.

Theorem 3.3. Under assumptions **H1**-**H5** and **H6** or **H7**, $r_{\mathbf{n}}(\cdot)$ converges almost completely to $r(\cdot)$ and we have

$$|r_{\mathbf{n}}(x_{\mathbf{j}}) - r(x_{\mathbf{j}})| = O\left(b_{\mathbf{n}} + \sqrt{\frac{\log \widehat{\mathbf{n}}}{\widehat{\mathbf{n}}\, b_{\mathbf{n}}^{d}\, \rho_{\mathbf{n}}^{N}}}\right) \text{ a.c.}$$

Theorem 3.4. Under assumptions **H1**-**H5** and **H6** or **H7**, $r_{\mathbf{n}}(\cdot)$ converges in mean of order q to $r(\cdot)$ and we have

$$\|r_{\mathbf{n}}(x_{\mathbf{j}}) - r(x_{\mathbf{j}})\|_{q} = O\left(b_{\mathbf{n}} + \sqrt{\frac{1}{\widehat{\mathbf{n}} b_{\mathbf{n}}^{d} \rho_{\mathbf{n}}^{N}}}\right),\ q > 1.$$

The proofs of these theorems are given in Appendix.

Remark 3.5. This current work is supported by a particular sampling scheme, which only includes deterministic designs for the spatial locations. For this reason, the bound of Theorems 3.3 and 3.4 shows a dissymmetric contribution of $b_{\mathbf{n}}$ and $\rho_{\mathbf{n}}$ on the risk even though both kernels K_1 and K_2 play symmetric roles. One can generalize this work to random spatial sample such as in Menezes et al. (2010) (for real-valued regression) and in Kelejian and Prucha (2007) (for spatial HAC estimation) and have a bound including $\rho_{\mathbf{n}}^{\alpha}$.

The remainder of this section focuses on the application of the proposed regression function through an example, namely the spatial prediction.

3.3.3 Spatial prediction

We propose here a spatial prediction methodology taking explicitly into account the spatial locations. Let $(X_{\mathbf{i}}, Y_{\mathbf{i}})_{\mathbf{i}}$ be a spatial process and we want to predict $Y_{\mathbf{i}}$ in some unobserved locations. More precisely, we suppose that the field $(X_{\mathbf{i}}, Y_{\mathbf{i}})_{\mathbf{i}}$ is observed on the set $\mathcal{O}_{\mathbf{n}}$ contained in $\mathcal{I}_{\mathbf{n}}$. The main purpose is to predict the unobserved value $Y_{\mathbf{i_0}}$ given $X_{\mathbf{i_0}}$ for a location $\mathbf{i_0} \in \mathcal{I}_{\mathbf{n}}$ but $\mathbf{i_0} \notin \mathcal{O}_{\mathbf{n}}$.

To achieve the forecasting at the site $\mathbf{i_0}$, we propose to use the regression function estimator $r_{\mathbf{n}}$ suggested previously. Then, the prediction of the value of the field $(Y_{\mathbf{i}})_{\mathbf{i}\in(\mathbb{Z})^N}$ at the location $\mathbf{i_0} \notin \mathcal{O}_{\mathbf{n}}$ is written

$$\widehat{Y}_{\mathbf{i_0}} = r_{\mathbf{n}}(X_{\mathbf{i_0}}) = \frac{\sum_{\mathbf{i}\in\mathcal{O}_{\mathbf{n}}} Y_{\mathbf{i}} K_1\left(\frac{X_{\mathbf{i_0}}-X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i_0}-\mathbf{i}\|)}{\sum_{\mathbf{i}\in\mathcal{O}_{\mathbf{n}}} K_1\left(\frac{X_{\mathbf{i_0}}-X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i_0}-\mathbf{i}\|)}. \tag{3.2}$$

One can derive an asymptotic result such as almost complete convergence and consistency in L^q norm ($q \in \mathbb{N}^*$) for $\widehat{Y}_{\mathbf{i_0}}$ given by (3.2) from the kernel regression estimate (3.1) studied previously. The proof is immediate and then is not given in Appendix.

Remark 3.6.

- In the more general case where the site $\mathbf{i_0}$ does not belong to $\mathcal{I}_{\mathbf{n}}$, one can rewrite the predictor (3.2) as

$$\widehat{Y}_{\mathbf{i_0}} = r_{\mathbf{n}}(X_{\mathbf{i_0}}) = \frac{\sum_{\mathbf{i}\in\mathcal{O}_{\mathbf{n}}} Y_{\mathbf{i}} K_1\left(\frac{X_{\mathbf{i_0}}-X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_2\left(\frac{\mathbf{i}-\mathbf{i_0}}{\rho_{\mathbf{n}}}\right)}{\sum_{\mathbf{i}\in\mathcal{O}_{\mathbf{n}}} K_1\left(\frac{X_{\mathbf{i_0}}-X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_2\left(\frac{\mathbf{i}-\mathbf{i_0}}{\rho_{\mathbf{n}}}\right)}$$

  where sites $\mathbf{i}$ and $\mathbf{i_0}$ are not normalized and $K_2(\cdot)$ is a kernel on $\mathbb{R}^N$.
- Note that predictor (3.2) is similar to Wang et al. (2012)'s predictor in the local linear spatio-temporal context with the specific assumption that the error term of the model is autocorrelated. Indeed, in Wang et al. (2012) the estimate of the error term is composed of a kernel weight function taking into account the distance between sites.
- In Biau and Cadre (2004), it is assumed that a generic value $Y_{\mathbf{i_0}}$ only depends on the values taken by the field in a vicinity of $\mathbf{i_0}$, denoted $\mathcal{V}_{\mathbf{i_0}}$, not containing $\mathbf{i_0}$. In the time context, this means that the field is Markovian. We denote by $\widetilde{Y}_{\mathbf{i_0}}$ the vector whose the d-components are the $\{Y_{\mathbf{i}}, \mathbf{i} \in \mathcal{V}_{\mathbf{i_0}}\}$, concatenated and ordered according to an arbitrary order. In this situation, the proposed predictor of Y at an unobserved site $\mathbf{i_0}$, denoted $\widehat{Y}^{(BC)}_{\mathbf{i_0}}$, is written

$$\widehat{Y}^{(BC)}_{\mathbf{i_0}} = \frac{\sum_{\substack{\mathbf{i}\in\mathcal{O}_{\mathbf{n}} \\ \mathcal{V}_{\mathbf{i}\in\mathcal{O}_{\mathbf{n}}}}} Y_{\mathbf{i}} K\left(\frac{\widetilde{Y}_{\mathbf{i_0}}-\widetilde{Y}_{\mathbf{i}}}{h}\right)}{\sum_{\substack{\mathbf{i}\in\mathcal{O}_{\mathbf{n}} \\ \mathcal{V}_{\mathbf{i}\in\mathcal{O}_{\mathbf{n}}}}} K\left(\frac{\widetilde{Y}_{\mathbf{i_0}}-\widetilde{Y}_{\mathbf{i}}}{h}\right)}$$

  where the kernel $K : \mathbb{R}^d \to \mathbb{R}$ is a probability density and h is the bandwidth parameter. In this context, the proposed predictor (3.2) can be extended in the following way

$$\widehat{Y}'_{\mathbf{i_0}} = \frac{\sum_{\substack{\mathbf{i}\in\mathcal{O}_{\mathbf{n}} \\ \mathcal{V}_{\mathbf{i}\in\mathcal{O}_{\mathbf{n}}}}} Y_{\mathbf{i}} K_1\left(\frac{\widetilde{Y}_{\mathbf{i_0}}-\widetilde{Y}_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i_0}-\mathbf{i}\|)}{\sum_{\substack{\mathbf{i}\in\mathcal{O}_{\mathbf{n}} \\ \mathcal{V}_{\mathbf{i}\in\mathcal{O}_{\mathbf{n}}}}} K_1\left(\frac{\widetilde{Y}_{\mathbf{i_0}}-\widetilde{Y}_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i_0}-\mathbf{i}\|)}.$$

- The nonparametric predictor of Y at an unobserved site $\mathbf{i_0}$, denoted $\widehat{Y}^{(MGF)}_{\mathbf{i_0}}$, considered in Menezes et al. (2010) for spatial stochastic processes when a stochastic sampling design is assumed for selection of random locations, is defined by

$$\widehat{Y}^{(MGF)}_{\mathbf{i_0}} = \frac{\sum_{\mathbf{i}\in\mathcal{O}_{\mathbf{n}}} Y_{\mathbf{i}} K_d\left(\frac{\mathbf{i_0}-\mathbf{i}}{h}\right)}{\sum_{\mathbf{i}\in\mathcal{O}_{\mathbf{n}}} K_d\left(\frac{\mathbf{i_0}-\mathbf{i}}{h}\right)}$$

where K_d represents a d-dimensional kernel function and h is the bandwidth parameter.

Now that we have checked the theoretical behavior of our estimator, we are going to study its practical features through some numerical results. To this end, in the following section, a procedure to estimate $r(X_{\mathbf{j}})$, $\mathbf{j} \in \mathcal{I}_{\mathbf{n}}$ is given. Moreover, our estimator is illustrated by some simulations as well as an application to a multivariate soil data set related to heavy metal contamination in the Swiss Jura.

3.4 Numerical results

In this section, we study the performance of the proposed regression estimator towards some simulations which highlight the importance of taking into account the spatial locations of the data. We remind that our theoretical result is obtained under a mixing condition which can be taken into account by the kernel function on the locations. We compare our estimator with the one that ignores any spatial dependence between the observations in the structure of the regression estimate (see Dabo-Niang et al. (2011b)). We consider a sample of dependent realizations of some spatial multivariate variables $X_{\mathbf{i}}$ with same distribution as a random field X valued in some d-dimensional space. Before studying the numerical results, we describe a useful procedure to estimate the spatial regression function investigated in this work. We also carry out an environmental case study, that is the concentration prediction of heavy metal in the soil of the Swiss Jura.
To ease the reading, let $K_{1\mathbf{i}} = K_1\left(b_{\mathbf{n}}^{-1}(x_{\mathbf{j}} - X_{\mathbf{i}})\right)$ and $K_{2\mathbf{i}} = K_2\left(\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{j}-\mathbf{i}}{\mathbf{n}}\right\|\right)$.

3.4.1 Procedure of estimation of $r(X_{\mathbf{j}})$, $\mathbf{j} \in \mathcal{I}_{\mathbf{n}}$

Step 1: Specify sets of bandwidths $S(b)$ and $S(\rho)$ of respectively K_1 and K_2.

Step 2: For each $b_{\mathbf{n}} \in S(b)$ and $\rho_{\mathbf{n}} \in S(\rho)$ and each site $\mathbf{j} \in \mathcal{I}_{\mathbf{n}}$, compute

$$r_{\mathbf{n}}^{\sharp}(X_{\mathbf{j}}) = \frac{\sum_{\substack{\mathbf{i}\in\mathcal{I}_{\mathbf{n}} \\ \mathbf{i}\neq\mathbf{j}}} Y_{\mathbf{i}} K_{1\mathbf{i}} K_{2\mathbf{i}}}{\sum_{\substack{\mathbf{i}\in\mathcal{I}_{\mathbf{n}} \\ \mathbf{i}\neq\mathbf{j}}} K_{1\mathbf{i}} K_{2\mathbf{i}}}$$

Step 3: Compute optimal bandwidths $b_{\mathbf{n},opt}$ and $\rho_{\mathbf{n},opt}$ by applying a cross-validation procedure over $S(b)$ and $S(\rho)$. More precisely, consider the following minimization problem, i.e. determine $b_{\mathbf{n},opt}$ and $\rho_{\mathbf{n},opt}$ minimizing the mean squared error over the $\hat{\mathbf{n}}$ sites

$$\min_{b_{\mathbf{n}},\rho_{\mathbf{n}}} \frac{1}{\hat{\mathbf{n}}} \sum_{\mathbf{j}\in\mathcal{I}_{\mathbf{n}}} (r_{\mathbf{n}}^{\sharp}(X_{\mathbf{j}})) - Y_{\mathbf{j}})^2$$

Step 4: For each site $\mathbf{j}$, compute $r_{\mathbf{n},opt}^{\sharp}(X_{\mathbf{j}})$ corresponding to $b_{\mathbf{n},opt}$ and $\rho_{\mathbf{n},opt}$.

Thus, this procedure is used in the subsequent analysis, in which we aim to study the behavior of our estimator. All the following numerical analysis were carried out using the R software (version 3.0.1).

3.4.2 Simulation study

In order to illustrate the fact that our method works for multidimensional random variables, we focus on the case where X belongs to $\mathbb{R}^d$ with $d > 1$ (naturally the procedure is valid if $d = 1$). The procedure presented in Section 3.4.1 is used in the current study dealing with $N = 2$. We consider observations $(X_{(i,j)}, Y_{(i,j)})$, $1 \leq i \leq n_1$ and $1 \leq j \leq n_2$. We will denote by $GRF(m, \sigma^2, s)$ any stationary Gaussian random field with mean m and spatial covariance function defined by

$$C(h) = \sigma^2 \exp\left(-\left(\frac{\|h\|}{s}\right)^2\right), \qquad h \in \mathbb{R}^2 \text{ and } s > 0.$$

In the following, we deal with $d = 3$ and

$$\begin{aligned} Y_{(i,j)} &= r(X_{(i,j)}) + \epsilon_{(i,j)} \\ &= \sin\left(X^{(1)}_{(i,j)} + X^{(2)}_{(i,j)} + X^{(3)}_{(i,j)}\right) + \epsilon_{(i,j)} \end{aligned}$$

with $\epsilon = (\epsilon_{(i,j)})$ is a random variable such as $\epsilon = GRF(0, 0.01, 3)$ and

$$X^{(d)}_{(i,j)} = D_{(i,j)} \times \left(\frac{2}{M}\right)^{1/2} \sum_{k=1}^{M} \cos(w(1,k) \times i + w(2,k) \times j + q(k) \times t_d + r(k)),$$

where $t_1 = 1$, $t_2 = 5$, $t_3 = 9$. Moreover, $w(g,k)$, $g = 1$, 2 and $q(k)$, $k = 1, \ldots, M$ are independent and identically distributed (i.i.d.) from a standard normal distribution ($m = 0$ and $\sigma = 0.5$), independent of $r(k)$ which are i.i.d. uniform random variables on $[-\pi, \pi]$. Moreover, we define $D_{(i,j)} = \frac{1}{n_1 \times n_2} \sum_{1 \leq m \leq n_1, 1 \leq t \leq n_2} \exp\left(-\frac{\|(i,j)-(m,t)\|}{a}\right)$. The function D is here to ensure and control the spatial mixing condition: the greater a is, the weaker the spatial dependency. Accordingly, we provide simulation results obtained with different values of a which are $a = 2$, 5, 10, and 25 and different grid size ($\widehat{\mathbf{n}} = 25 \times 25 = 625$ and $\widehat{\mathbf{n}} = 35 \times 30 = 1050$). Note that the simulation scheme of $X_{(i,j)}$ is inspired by the work of Wang et al. (2012). An example of considered simulated variable X is given in Figure 3.1, for a grid size of $\widehat{\mathbf{n}} = 1050$ and a value of $a = 2$. Along this part, the spatial regression is computed based on the kernels K_1 as the multivariate Epanechnikov kernel and K_2 as the Parzen kernel.

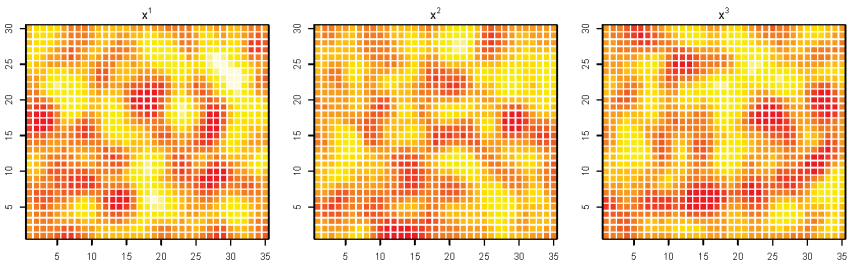


Figure 3.1 – Simulated field X^d
Situation where the dimension d is 3, the grid size is $\widehat{\mathbf{n}} = 1050$ and the value of a is 2

To assess the performance of the proposed regression estimator, now denoted by $r^{\sharp}_{\mathbf{n}}(\cdot)$ and to compare it with the one that does not directly take into account the distance between locations (e.g. Biau and Cadre (2004)) and denoted by $r^{\star}_{\mathbf{n}}(\cdot)$, the studied model

is replicated 30 times. Recall that $r^{\sharp}_{\mathbf{n}}(\cdot)$ and $r^{\star}_{\mathbf{n}}(\cdot)$ are defined by

$$r^{\sharp}_{\mathbf{n}}(X_{\mathbf{j}}) = \frac{\sum_{\substack{\mathbf{i}\in\mathcal{I}_{\mathbf{n}},\\ \mathbf{i}\neq\mathbf{j}}} Y_{\mathbf{i}} K_{1\mathbf{i}} K_{2\mathbf{i}}}{\sum_{\substack{\mathbf{i}\in\mathcal{I}_{\mathbf{n}},\\ \mathbf{i}\neq\mathbf{j}}} K_{1\mathbf{i}} K_{2\mathbf{i}}} \qquad \text{and} \qquad r^{\star}_{\mathbf{n}}(X_{\mathbf{j}}) = \frac{\sum_{\substack{\mathbf{i}\in\mathcal{I}_{\mathbf{n}},\\ \mathbf{i}\neq\mathbf{j}}} Y_{\mathbf{i}} K_1\left(h_{\mathbf{n}}^{-1}(X_{\mathbf{j}} - X_{\mathbf{i}})\right)}{\sum_{\substack{\mathbf{i}\in\mathcal{I}_{\mathbf{n}},\\ \mathbf{i}\neq\mathbf{j}}} K_1\left(h_{\mathbf{n}}^{-1}(X_{\mathbf{j}} - X_{\mathbf{i}})\right)}.$$

Note that if $\sum_{\substack{\mathbf{i}\in\mathcal{I}_{\mathbf{n}},\\ \mathbf{i}\neq\mathbf{j}}} K_{1\mathbf{i}} K_{2\mathbf{i}} = 0$ then $r^{\sharp}_{\mathbf{n}}(X_{\mathbf{j}}) = \frac{1}{\widehat{\mathbf{n}}-1}\sum_{\substack{\mathbf{i}\in\mathcal{I}_{\mathbf{n}},\\ \mathbf{i}\neq\mathbf{j}}} Y_{\mathbf{i}}$. In the same way, if $\sum_{\substack{\mathbf{i}\in\mathcal{I}_{\mathbf{n}},\\ \mathbf{i}\neq\mathbf{j}}} K_1\left(h_{\mathbf{n}}^{-1}(X_{\mathbf{j}} - X_{\mathbf{i}})\right) = 0$ then $r^{\star}_{\mathbf{n}}(X_{\mathbf{j}}) = \frac{1}{\widehat{\mathbf{n}}-1}\sum_{\substack{\mathbf{i}\in\mathcal{I}_{\mathbf{n}},\\ \mathbf{i}\neq\mathbf{j}}} Y_{\mathbf{i}}$.

At each replication $k = 1, \ldots, 30$, we compute the coefficient of determination over the $\widehat{\mathbf{n}}$ sites. The bandwidths used are those obtained using the previous procedure 3.4.1. Note that the optimal bandwidths are different at each replication. For the k^{th} replication, $1 \leq k \leq 30$, we define the coefficient of determination ($R^{2(k)}$) by

$$R^{2(k)} = 1 - \frac{\sum_{\mathbf{j}\in\mathcal{I}_{\mathbf{n}}}(Y_{\mathbf{j}} - r^{\dagger}_{\mathbf{n},opt}(X_{\mathbf{j}}))^2}{\sum_{\mathbf{j}\in\mathcal{I}_{\mathbf{n}}}(Y_{\mathbf{j}} - \overline{Y})^2}, \qquad \text{with } r^{\dagger}_{\mathbf{n}} = r^{\sharp}_{\mathbf{n}} \text{ or } r^{\star}_{\mathbf{n}} \tag{3.3}$$

where $\overline{Y}$ denotes the mean of the $Y_{\mathbf{i}}$. The obtained results are summarized in Table 3.1. For each value of a, this table provides the average of the $R^{2(k)}$ over the 30 replications of Equation (3.3), denoted $\overline{R}^2$, and, in brackets, the corresponding standard deviation. The column entitled "p-value" gives, for each considered situation, the p-value of a paired t-test performing in order to determine if the mean of $R^{2(k)\sharp}$ is significantly greater than that of $R^{2(k)\star}$ (the alternative hypothesis is then $H1 : \overline{R}^{2\sharp} > \overline{R}^{2\star}$).

a	$\widehat{\mathbf{n}}$	$\overline{R}^{2\sharp}$	$\overline{R}^{2\star}$	p-value
2	625	0.9241 (0.0275)	0.3797 (0.0847)	5.93×10^{-25}
	1050	0.9008 (0.0145)	0.2391 (0.0702)	2.16×10^{-31}
5	625	0.9599 (0.0091)	0.8748 (0.0281)	1.08×10^{-17}
	1050	0.9492 (0.0082)	0.7884 (0.0403)	1.82×10^{-21}
10	625	0.9677 (0.0060)	0.9533 (0.0107)	6.47×10^{-12}
	1050	0.9649 (0.0056)	0.9398 (0.0114)	5.51×10^{-17}
25	625	0.9614 (0.0079)	0.9595 (0.0087)	2.14×10^{-05}
	1050	0.9681 (0.0046)	0.9650 (0.0051)	5.71×10^{-13}

Table 3.1 – Simulation results according to the value of $a = 2$, 5, 10 and 25 and the grid size ($\widehat{\mathbf{n}} = 625$ or 1050)

The table gives the average coefficient of determination ($\overline{R}^2$) and, in brackets, the corresponding standard deviation. The column entitled "p-value" gives the p-value of a paired t-test performing in order to determine whether $\overline{R}^{2\sharp}$ is significantly greater than $\overline{R}^{2\star}$.

We note that our estimator $r^{\sharp}_{\mathbf{n}}$ performs better than the estimator $r^{\star}_{\mathbf{n}}$ especially when a is small, that is when the spatial dependence is important. In fact, for a value of $a = 2$, the average of the 30 coefficients of determination $R^{2\sharp}$ is equal to 0.9241 (respectively 0.9008) which is significantly higher than those of $R^{2\star}$ equals to 0.3797 (respectively 0.2391), for a grid size of 625 (respectively 1050). Insight into this performance can also be viewed from Figure 3.2 in which a field Y, from one replication k, is represented in the left caption whereas the squared errors obtained using functions $r^{\sharp}_{\mathbf{n}}$ and $r^{\star}_{\mathbf{n}}$ are represented in the middle and right captions respectively. The last two confirm that our methodology generates less errors than using the regression function $r^{\star}_{\mathbf{n}}$ since the more colorful the representation is, the greater the error. The low $p-$values (less than 2.14×10^{-05}) confirm

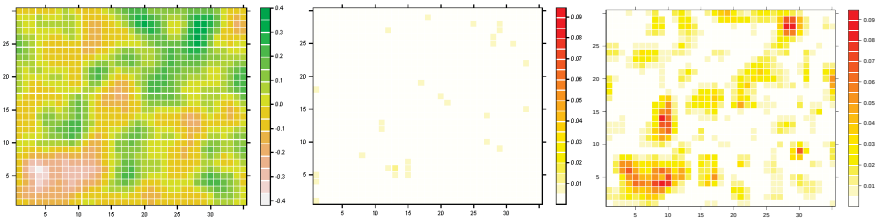


Figure 3.2 – Illustration of the results
A simulated field considering $\widehat{\mathbf{n}} = 1050$ and $a = 2$ with (left to right) an image of the field Y; the squared errors using $r^{\sharp}_{\mathbf{n}}$; the squared errors using $r^{\star}_{\mathbf{n}}$

that $r^{\sharp}_{\mathbf{n}}$ produces less errors than $r^{\star}_{\mathbf{n}}$. Nevertheless, the probability of erroneously rejecting the null hypothesis grows when the value of a increases. Indeed, when the value of a increases, the departure between $R^{2\sharp}$ and $R^{2\star}$ decreases because $r^{\sharp}_{\mathbf{n}}$ tends to behave in the same way that $r^{\star}_{\mathbf{n}}$. For example, for a value of $a = 25$, the average of the 30 coefficients of determination $R^{2\sharp}$ is equal to 0.9614 (respectively 0.9681) which is close to those of $R^{2\star}$ equals to 0.9595 (respectively 0.9650) for a grid size of 625 (respectively 1050). In other words, a reduction of the spatial dependence induces a rise of the bandwidth $\rho_{\mathbf{n}}$ and when the bandwidth $\rho_{\mathbf{n}}$ is at its maximum, the two estimators have similar behaviors.

This simulation study shows that our methodology improves the classical one in presence of highly dependent data. To complete the study, we deal, in the following, with a real case study.

3.4.3 Environmental dataset study

Here, we are interested in studying the behavior of our predictor through the famous Jura data set (*https://sites.google.com/site/goovaertspierre/pierregoovaertswebsite/download/jura-data*) which is often encountered in the geostatistic literature, for example, in Atteia et al. (1994), Webster et al. (1994), Atteia et al. (1995), Goovaerts (1997, 1998), Bel et al. (2009) and Allard et al. (2011). Data were collected by the Swiss Federal Institute of Technology at Lausanne. A complete exploratory data analysis of this multivariate soil data set is provided in the monograph of Goovaerts (1997) whereas a detailed description of the sampling field, and laboratory procedures is given in Atteia et al. (1994) and Webster et al. (1994). The data concern concentration of seven heavy metals (cadmium Cd, cobalt Co, chromium Cr, copper Cu, nickel Ni, lead Pb, and zinc Zn) of a 14.5 km^2 region in the Swiss Jura. All metal concentrations were measured at 359 locations but two subsets are considered: the first (prediction data set), composed of 259 locations, is used for parameters estimation whereas the second (validation data set), composed of 100 locations, will be used to check results provided by predictors. Figure 3.3 represents locations of the sites from each subset.

Regression

In Goovaerts (1998), prediction performances of co-kriging estimators are assessed on these data where three cases are considered according to the secondary variables used to estimate primary metal at the 100 test locations. Then, in order to compare performances of our nonparametric estimator with co-kriging estimator, we treat these same three cases, presented in Table 3.2. To this end, we use Equation (3.2) of the regression estimator where the $Y_{\mathbf{i}}$s are replaced by the observations of the primary variable and the secondary variables

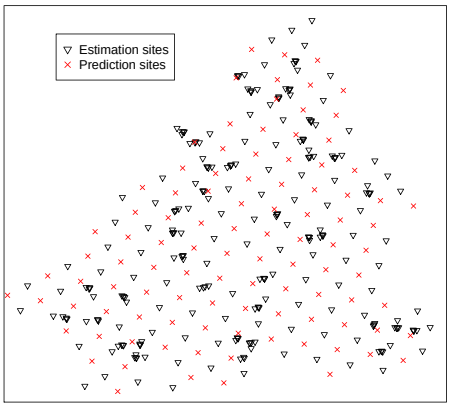


Figure 3.3 – Locations considered in the studied region of Swiss Jura

refer to the $X_{\mathbf{i}}$s.

Case	Primary variable	Secondary variable
1	Cadmium (Cd)	Nickel (Ni), Zinc (Zn)
2	Copper (Cu)	Lead (Pb), Nickel (Ni), Zinc (Zn)
3	Lead (Pb)	Copper (Cu), Nickel (Ni), Zinc (Zn)

Table 3.2 – Three considered cases

We are particularly interested in the heterotopic situation where secondary variables are available at the 259 primary data locations plus the 100 test locations but different situations are also treated in Goovaerts (1998). The performances are assessed using the mean absolute error (MAE) of prediction, that is the average absolute difference between the *true* values and the predictions. These results are presented in Table 3.3. In Cases 1 and 2, our method produces the better prediction results compared to the parametric prediction methods and the nonparametric predictor that does not take into account the distance between sites. For Case 3, our results are close to the best predictions obtained by the revisited co-kriging method. Note that according to the case, different kind of kernels are used (see Table 3.3).

Method	Case 1	Case 2	Case 3
Ordinary Cokriging	0.51	7.90	10.80
Revisited Cokriging (cov)	0.52	7.80	10.70
Revisited Cokriging (corr)	0.52	7.40	**10.60**
Nonparametric $r_{\mathbf{n}}^{\sharp}$	**0.42**	**7.02**	11.02
$r_{\mathbf{n}}^{\star}$	0.44	7.80	12.51
K_1	Silverman	Gaussian	Silverman
K_2	Biweight	Parzen	Parzen

Table 3.3 – Prediction performances measured by the mean absolute error (MAE) for the different methods on the three considered cases

Prediction

Now we consider the situation where only the primary variable is available at 259 locations. The purpose is to predict the concentration of the primary variable on the 100 unobserved locations from the 259 observations. In the parametric context, the kriging estimate allows to solve this problem. The kriging predictions for the three primary variables

of the Jura data set are obtained in Goovaerts (1997) and the results are reported in Table 3.4. We propose to use the predictor given by Equation (3.2). In this situation, the $X_{\mathbf{i}}$'s are composed by the nearest observations of the site $\mathbf{i}$ of the primary variable of interest and $Y_{\mathbf{i}}$ is the observation at the site $\mathbf{i}$. The nonparametric predictions are also reported in Table 3.4. We also show the prediction obtained by the nonparametric kernel predictor with one kernel on the observations. We notice that our approach produces the smallest mean absolute error in all three cases under examination. As in the previous section 3.4.3, according to the considered case, different kind of kernels are used (see Table 3.4).

Primary variable	Kriging	Nonparametric $r_{\mathbf{n}}^{\sharp}$	Nonparametric $r_{\mathbf{n}}^{\star}$	K_1	K_2
Cadmium (Cd)	0.58	0.56	0.61	Circular	Epanechnikov
Copper(Cu)	15.40	14.99	15.53	Epanechnikov	Indicator
Lead (Pb)	20.90	20.10	21.96	Circular	Biweight

Table 3.4 – MAE of the prediction results

3.5 Conclusion

In this chapter, we propose a new method to estimate nonparametrically the spatial regression function of a stationary $(d+1)$-dimensional process. The originality of the proposed estimator is to take into account both the distance between sites and between the observations. It is shown that our estimator converges almost completely and in mean of order q. After studying the theoretical behavior of the proposed methodology, we are interested in its practical use. We provide some simulation results concerning the estimation of the regression function. Finally, the predictor deduced from the regression function is applied through an environmental data set. These applications show that our method performs better than existing methods, particularly in presence of spatial dependence. Consequently, one can see the proposed methodology as a good alternative to the classical kernel approach for spatial data.

We notice that the Jura data set contains also categorical attributes which take only a limited number of states, usually non-ordered, e.g., rock types or land uses. This kind of data is not included within our approach but is the topic of a forthcoming work. Moreover, we could investigate the case of continuously indexed spatial processes with this approach. Also, an adaptation of this method to issues such as the spatial conditional mode or quantile regression estimation could be developed. Moreover, note that this work does not deal with the boundary problem (edge effect) for which points on the edge have less neighbors than the others. One solution would be to give less weight to data on the boundary. This issue will be considered in further investigations.

3.6 Appendix

Recall that $K_{1\mathbf{i}} = K_1\left(\frac{x_{\mathbf{j}}-X_{\mathbf{i}}}{b_{\mathbf{n}}}\right)$, $K_{2\mathbf{i}} = K_2\left(\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{j}-\mathbf{i}}{\mathbf{n}}\right\|\right)$ and let $W_{\mathbf{ni}} = \dfrac{K_{1\mathbf{i}}K_{2\mathbf{i}}}{\sum_{\mathbf{k}\in\mathcal{I}_{\mathbf{n}}} K_{1\mathbf{k}}K_{2\mathbf{k}}}$.

Some results for the proofs

Lemma 3.7. From an adjustment of Lemma 3.2 in Dabo-Niang and Yao (2007)
Let $(\zeta_{\mathbf{v}},\ \mathbf{v}\in\mathbb{N}^N)$ be a zero-mean real-values random spatial process such that $\sup_{\mathbf{v}}|\zeta_{\mathbf{v}}| \leq$

b, and let $S_{\mathbf{n}} = \sum_{\mathbf{v}\in I_{\mathbf{n}}} \zeta_{\mathbf{v}}$ for $\mathbf{n} \in (\mathbb{N}^*)^N$. Then for each $\mathbf{t} \in (\mathbb{N}^*)^N$ such that $1 \leq t_i \leq \frac{1}{2}n_i$ and each $\epsilon_0 > 0$,

$$\mathbb{P}(|S_{\mathbf{n}}| > \widehat{\mathbf{n}}\epsilon_0) \leq 2^{N+1} \exp\left(-\frac{\epsilon_0^2}{4v^2(\mathbf{t})}\widehat{\mathbf{t}}\right) + \frac{2^{N+2}b\psi([\widehat{\mathbf{t}}-1]p^N, p^N)\varphi(p)}{\epsilon_0},$$

where $v^2(\mathbf{t}) = \frac{4}{p^{2N}}\sigma^2(\mathbf{t}) + b\epsilon_0$ with p an integer such that $p = \frac{n_i}{2t_i}$, $\widehat{\mathbf{n}} = n_1 \times \ldots \times n_N$, $I_{\mathbf{n}} = \{\mathbf{i} = (i_1, \ldots, i_N) \in \mathbb{N}^N, 1 \leq i_k \leq n_k\}$ and $\sigma^2(\mathbf{t}) = Var(\sum_{1\leq v_k\leq p,\, k=1,\ldots,N} \zeta_{\mathbf{v}})$.

Lemma 3.8. Under the conditions of Theorem 3.3, we have

$$Var(f_{\mathbf{n}}(x_{\mathbf{j}}) - \mathbb{E}[f_{\mathbf{n}}(x_{\mathbf{j}})]) = O\left(\frac{1}{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\right)$$

Lemma 3.9. Under the conditions of Theorem 3.4, we have

$$\mathbb{E}^{1/q}\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}}\mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}}) - r(x_{\mathbf{j}})\right]^q = O(b_{\mathbf{n}})$$

Lemma 3.10. Under the conditions of Theorem 3.4, we have

$$\mathbb{E}^{1/q}\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}}\left[Y_{\mathbf{i}} - \mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}})\right]\right]^q = O\left(\left(\frac{1}{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\right)^{1/2}\right)$$

Lemma 3.11. Under the conditions of Theorem 3.4, we have

$$\mathbb{E}\left(\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \xi_{\mathbf{i}}\right)^q \leq O(\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N)^{q/2}$$

where $\xi_{\mathbf{i}} = K_{1\mathbf{i}}K_{2\mathbf{i}}\theta_{\mathbf{i}}$, with $\theta_{\mathbf{i}} = Y_{\mathbf{i}} - \mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}})$.

Lemma 3.12. Under the conditions of Theorem 3.4, we have

$$\left(\mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{1\mathbf{i}}K_{2\mathbf{i}} \leq \frac{ua_{\mathbf{n},\mathbf{j}}}{2}\right]\right)^{1/q} = O\left(\left(\frac{1}{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\right)^{1/2}\right)$$

with $u = \mathbb{E}[K_{1\mathbf{i}}]$.

Lemma 3.13. Under the conditions of Theorem 3.4, we have

$$\mathbb{E}^{1/q}\left[\left(\frac{1}{\widehat{\mathbf{n}}}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} - r(x)\right)\mathbf{1}_{\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}}=0\right]}\right]^q = O\left(\left(\frac{1}{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\right)^{1/2}\right)$$

Proofs

Proof of Theorem 3.3

We show that $r_{\mathbf{n}}(x_{\mathbf{j}})$ converges almost completely to $r(x_{\mathbf{j}})$ when $f(x_{\mathbf{j}}) > 0$. We note that

$$r_{\mathbf{n}}(x_{\mathbf{j}}) - r(x_{\mathbf{j}}) = \left(\frac{\varphi_{\mathbf{n}}(x_{\mathbf{j}}) - \varphi(x_{\mathbf{j}})}{f_{\mathbf{n}}(x_{\mathbf{j}})} - \varphi(x_{\mathbf{j}}) \frac{f_{\mathbf{n}}(x_{\mathbf{j}}) - f(x_{\mathbf{j}})}{f_{\mathbf{n}}(x_{\mathbf{j}}) f(x_{\mathbf{j}})} \right) \mathbf{1}_{\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}} \neq 0\right]} + \overline{Y} \mathbf{1}_{\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}} = 0\right]}$$

$$\begin{aligned} |r_{\mathbf{n}}(x_{\mathbf{j}}) - r(x_{\mathbf{j}})| \leq & \left(\frac{1}{f_{\mathbf{n}}(x_{\mathbf{j}})} |\varphi_{\mathbf{n}}(x_{\mathbf{j}}) - \varphi(x_{\mathbf{j}})| + \frac{\varphi(x_{\mathbf{j}})}{f_{\mathbf{n}}(x_{\mathbf{j}}) f(x_{\mathbf{j}})} |f_{\mathbf{n}}(x_{\mathbf{j}}) + f(x_{\mathbf{j}})| \right) \mathbf{1}_{\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}} \neq 0\right]} \\ & + \overline{Y} \mathbf{1}_{\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}} = 0\right]} \end{aligned} \tag{3.4}$$

where $\overline{Y}$ is the empirical mean of the $Y_{\mathbf{i}}$.
We have to study the following terms $|\varphi_{\mathbf{n}}(x_{\mathbf{j}}) - \varphi(x_{\mathbf{j}})|$ and $|f_{\mathbf{n}}(x_{\mathbf{j}}) - f(x_{\mathbf{j}})|$. We will concentrate our attention on the second term. In fact, the procedure is the same for each term since the second is a particular case of the first when $Y_{\mathbf{i}}$ is equal to 1. Let us show that for each $x_{\mathbf{j}}$, $f_{\mathbf{n}}(x_{\mathbf{j}})$ converges almost completely (denoted a.c. in the following) to $f(x_{\mathbf{j}})$, i.e. $\forall \epsilon > 0$, $\sum_{\mathbf{n}} \mathbb{P}(|f_{\mathbf{n}}(x_{\mathbf{j}}) - f(x_{\mathbf{j}})| > \epsilon) < \infty$. For $x_{\mathbf{j}} \in \mathbb{R}^d$ located at some site $\mathbf{j}$, we note that

$$|f_{\mathbf{n}}(x_{\mathbf{j}}) - f(x_{\mathbf{j}})| \leq |f_{\mathbf{n}}(x_{\mathbf{j}}) - \mathbb{E}[f_{\mathbf{n}}(x_{\mathbf{j}})]| + |\mathbb{E}[f_{\mathbf{n}}(x_{\mathbf{j}})] - f(x_{\mathbf{j}})| .$$

Then, as usual, the result is obtained studying separately the bias and the variance terms.

The bias $|\mathbb{E}[f_{\mathbf{n}}(x_{\mathbf{j}})] - f(x_{\mathbf{j}})|$**.**

Before going further note that

$$\begin{aligned} |\mathbb{E}[f_{\mathbf{n}}(x_{\mathbf{j}})] - f(x_{\mathbf{j}})| &= \left| \mathbb{E}\left(\frac{1}{b_{\mathbf{n}}^d a_{\mathbf{n},\mathbf{j}}} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{n}}} \right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j} - \mathbf{i}\|) \right) - f(x_{\mathbf{j}}) \right| \\ &= \left| \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \frac{1}{b_{\mathbf{n}}^d a_{\mathbf{n},\mathbf{j}}} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j} - \mathbf{i}\|) \int K_1\left(\frac{x_{\mathbf{j}} - u}{b_{\mathbf{n}}} \right) f(u) du - f(x_{\mathbf{j}}) \right| \\ &= \left| \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \frac{1}{a_{\mathbf{n},\mathbf{j}}} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j} - \mathbf{i}\|) \int K_1(v) f(x_{\mathbf{j}} - v b_{\mathbf{n}}) dv - f(x_{\mathbf{j}}) \right| \end{aligned}$$

Furthermore, $\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i} - \mathbf{j}\|)/a_{\mathbf{n},\mathbf{j}} = 1$, $\int K_1(v) dv = 1$ and

$$\begin{aligned} \int K_1(v) |f(x_{\mathbf{j}} + v b_{\mathbf{n}}) - f(x_{\mathbf{j}})| dv &\leq \int K_1(v) \|x_{\mathbf{j}} + v b_{\mathbf{n}} - x_{\mathbf{j}}\| dv \qquad \text{by assumption } \mathbf{H2} \\ &\leq \int K_1(v) \|v b_{\mathbf{n}}\| dv \\ &\leq C b_{\mathbf{n}} \end{aligned}$$

because $\int \|v\| K_1(v) dv < +\infty$. Consequently, $|\mathbb{E}[f_{\mathbf{n}}(x_{\mathbf{j}})] - f(x_{\mathbf{j}})| = O(b_{\mathbf{n}})$.

The study of the asymptotic behavior of the term $|f_{\mathbf{n}}(x_{\mathbf{j}}) - \mathbb{E}[f_{\mathbf{n}}(x_{\mathbf{j}})]|$**.**

Let $f_{\mathbf{n}}(x_{\mathbf{j}}) - \mathbb{E}[f_{\mathbf{n}}(x_{\mathbf{j}})] = \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \Lambda_{\mathbf{i}}(x_{\mathbf{j}}) = S_{\mathbf{n}}(x_{\mathbf{j}})$, with $\Lambda_{\mathbf{i}}(x_{\mathbf{j}}) = \lambda_{\mathbf{i}}(x_{\mathbf{j}}) - \mathbb{E}(\lambda_{\mathbf{i}}(x_{\mathbf{j}}))$ where

$$\lambda_{\mathbf{i}}(x_{\mathbf{j}}) = \frac{1}{a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{n}}} \right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i} - \mathbf{j}\|).$$

We are interested in studying $P = \mathbb{P}[|f_{\mathbf{n}}(x_{\mathbf{j}}) - \mathbb{E}[f_{\mathbf{n}}(x_{\mathbf{j}})]| > \epsilon]$. To this end, we use Lemma 3.7 with $\epsilon = \widehat{\mathbf{n}}\epsilon_0$, $\epsilon > 0$ and since K_1 and K_2 are bounded we can write $\sup_{\mathbf{j}} |\Lambda_{\mathbf{i}}(x_{\mathbf{j}})| \leq b =$

$\frac{C}{a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d}$, with $C = \|K_1\|_\infty \|K_2\|_\infty$ (where $\|\cdot\|_\infty$ is the sup norm). Then, for each $\mathbf{t} \in (\mathbb{N}^*)^N$ such that $1 \leq t_i \leq \frac{1}{2} n_i$ and p an integer such that $p = \frac{n_i}{2t_i}$, we deduce from Lemma 3.7 that

$$\mathbb{P}(|S_{\mathbf{n}}(x_{\mathbf{j}})| > \widehat{\mathbf{n}}\epsilon_0) \leq 2^{N+1} \exp\left(-\frac{\epsilon_0^2}{4v^2(\mathbf{t})}\widehat{\mathbf{t}}\right) + \frac{2^{N+2} b \psi\left([\widehat{\mathbf{t}}-1]p^N, p^N\right)\varphi(p)}{\epsilon_0}$$

$$\begin{aligned}
P &= \mathbb{P}(|S_{\mathbf{n}}(x_{\mathbf{j}})| > \epsilon) \\
&\leq 2^{N+1} \exp\left(\frac{-(\epsilon/\widehat{\mathbf{n}})^2}{4\left(\frac{4}{p^{2N}}\sigma^2(\mathbf{t}) + b(\epsilon/\widehat{\mathbf{n}})\right)}\widehat{\mathbf{t}}\right) + \frac{2^{N+2} b \psi\left([\widehat{\mathbf{t}}-1]p^N, p^N\right)\varphi(p)}{\epsilon/\widehat{\mathbf{n}}} \\
&\leq 2^{N+1} \exp\left(\frac{-\epsilon^2}{4\left(\widehat{\mathbf{n}}\frac{4}{p^{2N}}\sigma^2(\mathbf{t}) + \frac{C}{a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d}\epsilon\right)}\frac{\widehat{\mathbf{t}}}{\widehat{\mathbf{n}}}\right) + \frac{2^{N+2} C}{a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d}\psi\left([\widehat{\mathbf{t}}-1]p^N, p^N\right)\varphi(p)\frac{\widehat{\mathbf{n}}}{\epsilon} \\
&\leq 2^{N+1} \exp\left(\frac{-\epsilon^2}{2^{N+2}p^N\left(2^N p^N \widehat{\mathbf{t}}\frac{4\sigma^2(\mathbf{t})}{p^{2N}} + \frac{C\epsilon}{a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d}\right)}\right) + \frac{2^{N+2} C}{a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d}\psi\left([\widehat{\mathbf{t}}-1]p^N, p^N\right)\varphi(p)\frac{\widehat{\mathbf{n}}}{\epsilon}
\end{aligned}$$

Let $\delta > 0$, $\epsilon = \epsilon_{\mathbf{n}} = \delta\left(\frac{\log \widehat{\mathbf{n}}}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}\right)^{1/2}$ and $p = \left\lfloor \left(\frac{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}{\log \widehat{\mathbf{n}}}\right)^{\frac{1}{2N}} \right\rfloor$. Moreover,

$$\begin{aligned}
\sigma^2(\mathbf{t}) = Var\left(\sum_{1 \leq i_k \leq p,\ k=1,\ldots,N} \Lambda_{\mathbf{i}}(x_{\mathbf{j}})\right) &= \mathbb{E}\left[\left(\sum_{1 \leq i_k \leq p,\ k=1,\ldots,N} \Lambda_{\mathbf{i}}(x_{\mathbf{j}})\right)^2\right] \\
&\leq \frac{1}{\widehat{\widehat{\mathbf{t}}}}\left(Var(f_{\mathbf{n}}(x_{\mathbf{j}}) - \mathbb{E}[f_{\mathbf{n}}(x_{\mathbf{j}})])\right) \\
&\leq \frac{1}{\widehat{\widehat{\mathbf{t}}}} \times O\left(\frac{1}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}\right)
\end{aligned}$$

according to Lemma 3.8.

$$\begin{aligned}
P &\leq 2^{N+1} \exp\left(\frac{-\delta^2 \log \widehat{\mathbf{n}}}{(\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N)\left(2^{2N+4}\widehat{\mathbf{t}}\frac{C}{\widehat{\mathbf{t}}\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N} + \frac{2^{N+2}C}{a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d}\delta\right)}\right) + 2^{N+2}\frac{C}{a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d}\psi([\widehat{\mathbf{t}}-1]p^N, p^N)\varphi(p)\frac{\widehat{\mathbf{n}}}{\epsilon} \\
&\leq 2^{N+1} \exp\left(\frac{-\delta^2 \log \widehat{\mathbf{n}}}{2^{2N+4}C + \frac{2^{N+2}C}{a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d}\delta\left(\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N\right)}\right) + \frac{2^{N+2}C}{a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d}\psi([\widehat{\mathbf{t}}-1]p^N, p^N)\varphi(p)\widehat{\mathbf{n}}\delta^{-1}\left(\frac{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}{\log \widehat{\mathbf{n}}}\right)^{1/2} \\
&\leq 2^{N+1} \exp\left\{\log \widehat{\mathbf{n}}^{-a}\right\} + \frac{2^{N+2}C}{a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d}\psi([\widehat{\mathbf{t}}-1]p^N, p^N)\varphi(p)\widehat{\mathbf{n}}\delta^{-1}\left(\frac{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}{\log \widehat{\mathbf{n}}}\right)^{1/2} \\
&\leq C_N \widehat{\mathbf{n}}^{-a} + 2^{N+2}\frac{C}{a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d}\psi([\widehat{\mathbf{t}}-1]p^N, p^N)\varphi(p)\widehat{\mathbf{n}}\delta^{-1}\left(\frac{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}{\log \widehat{\mathbf{n}}}\right)^{1/2}
\end{aligned}$$

with $a = \dfrac{\delta^2}{2^{2N+4}C + 2^{N+2}C_N\delta}$ and using the fact that $a_{\mathbf{n},\mathbf{j}} \geq Ck_{\mathbf{n}}$, $k_{\mathbf{n}} = C_N d_{\mathbf{n}}^N + O(d_{\mathbf{n}}^\beta)$, $d_{\mathbf{n}}^N = \widehat{\mathbf{n}}\rho_{\mathbf{n}}^N$. The convergence of $C_N \sum_{\mathbf{n} \in \mathbb{N}^N} \widehat{\mathbf{n}}^{-a} < \infty$ is insured by an appropriate choice of $\delta > 2^{N+1} C_N$. The second term of the right-hand-side of the previous inequality is treated according to assumptions on $\psi(n, m)$.

First, we consider assumption **H6**, i.e. $\psi(n,m) \leq C\min(n,m)$ and $\widehat{\mathbf{n}} b_{\mathbf{n}}^{d\theta_1}\rho_{\mathbf{n}}^{N\theta_1}\log\widehat{\mathbf{n}}^{\theta_2}u_{\mathbf{n}}^{\theta_3} \to \infty$ with $\theta > N(q+2) \geq 4N$ and we have

$$\begin{aligned} g_{\mathbf{n}} &= 2^{N+2}\frac{C}{a_{\mathbf{n},\mathbf{j}}b_{\mathbf{n}}^d}\psi([\widehat{\mathbf{t}}-1]p^N,p^N)\varphi(p)\widehat{\mathbf{n}}\delta^{-1}\left(\frac{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}{\log\widehat{\mathbf{n}}}\right)^{1/2} \\ &\leq 2^{N+2}\frac{C}{a_{\mathbf{n},\mathbf{j}}b_{\mathbf{n}}^d}p^N p^{-\theta}\widehat{\mathbf{n}}\delta^{-1}\left(\frac{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}{\log\widehat{\mathbf{n}}}\right)^{1/2} \\ &\leq 2^{N+2}\frac{C}{a_{\mathbf{n},\mathbf{j}}b_{\mathbf{n}}^d}\left(\frac{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}{\log\widehat{\mathbf{n}}}\right)^{\frac{2N-\theta}{2N}}\widehat{\mathbf{n}}\delta^{-1} \end{aligned}$$

Let the following sequence of calculations

$$\begin{aligned} g_{\mathbf{n}}\widehat{\mathbf{n}}u_{\mathbf{n}} &\leq C_N 2^{N+2}\widehat{\mathbf{n}}^{\frac{4N-\theta}{2N}}b_{\mathbf{n}}^{\frac{-d\theta}{2N}}\rho_{\mathbf{n}}^{\frac{-N\theta}{2N}}\log\widehat{\mathbf{n}}^{\frac{\theta-2N}{2N}}\delta^{-1}u_{\mathbf{n}} \\ &\leq C_N\left(\widehat{\mathbf{n}}b_{\mathbf{n}}^{\frac{-d\theta}{4N-\theta}}\rho_{\mathbf{n}}^{\frac{-N\theta}{4N-\theta}}\log\widehat{\mathbf{n}}^{\frac{\theta-2N}{4N-\theta}}u_{\mathbf{n}}^{\frac{2N}{4N-\theta}}\right)^{\frac{4N-\theta}{2N}} \\ &\leq C\,\mathcal{A}_{\mathbf{n}}^{\frac{4N-\theta}{2N}}, \end{aligned}$$

for $\mathbf{n}$ large enough. Note that in the case where $\theta > N(q+2)$ and $q > 1$, one has $\dfrac{4N-\theta}{2N} < 0$. Furthermore, assumption **H6** leads to $\mathcal{A}_{\mathbf{n}} \to +\infty$, as $\mathbf{n} \to \infty$. So, $g_{\mathbf{n}}\widehat{\mathbf{n}}u_{\mathbf{n}} \to 0$ and consequently, $g_{\mathbf{n}} < \dfrac{1}{\widehat{\mathbf{n}}u_{\mathbf{n}}}$ leads to

$$\sum_{\mathbf{n}\in\mathbb{N}^N} g_{\mathbf{n}} < \sum_{\mathbf{n}\in\mathbb{N}^N}\frac{1}{\widehat{\mathbf{n}}u_{\mathbf{n}}} < +\infty.$$

Second, we consider assumption **H7**, i.e. $\psi(n,m) \leq C(n+m+1)^{\widetilde{\beta}}$ where $\widehat{\mathbf{n}}b_{\mathbf{n}}^{d\theta'_1}\rho_{\mathbf{n}}^{N\theta'_1}\log\widehat{\mathbf{n}}^{\theta'_2}u_{\mathbf{n}}^{\theta'_3} \to \infty$ with $\theta > N(q+2\widetilde{\beta}+1) \geq N(2\widetilde{\beta}+3)$ and noticing that $\psi([\widehat{\mathbf{t}}-1]p^N,p^N) \leq C([\widehat{\mathbf{t}}-1]p^N + p^N+1)^{\widetilde{\beta}} \leq C\,\widehat{\mathbf{n}}^{\widetilde{\beta}}$. Then, we have

$$\begin{aligned} h_{\mathbf{n}} &= 2^{N+2}\frac{C}{a_{\mathbf{n},\mathbf{j}}b_{\mathbf{n}}^d}\psi([\widehat{\mathbf{t}}-1]p^N,p^N)\varphi(p)\widehat{\mathbf{n}}\delta^{-1}\left(\frac{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}{\log\widehat{\mathbf{n}}}\right)^{1/2} \\ &\leq C2^{N+2}\widehat{\mathbf{n}}^{\widetilde{\beta}}\frac{\widehat{\mathbf{n}}}{a_{\mathbf{n},\mathbf{j}}b_{\mathbf{n}}^d}p^{-\theta}\delta^{-1}\left(\frac{\log\widehat{\mathbf{n}}}{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\right)^{-1/2} \\ &\leq C2^{N+2}\widehat{\mathbf{n}}^{\widetilde{\beta}}\frac{\widehat{\mathbf{n}}}{a_{\mathbf{n},\mathbf{j}}b_{\mathbf{n}}^d}\delta^{-1}\left(\frac{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}{\log\widehat{\mathbf{n}}}\right)^{\frac{N-\theta}{2N}} \end{aligned}$$

and

$$\begin{aligned} h_{\mathbf{n}}\widehat{\mathbf{n}}u_{\mathbf{n}} &\leq C_N\left(\widehat{\mathbf{n}}b_{\mathbf{n}}^{\frac{-d(\theta+N)}{N(3+2\widetilde{\beta})-\theta}}\rho_{\mathbf{n}}^{\frac{-N(\theta+N)}{N(3+2\widetilde{\beta})-\theta}}\log\widehat{\mathbf{n}}^{\frac{\theta-N}{N(3+2\widetilde{\beta})-\theta}}u_{\mathbf{n}}^{\frac{2N}{N(3+2\widetilde{\beta})-\theta}}\right)^{\frac{N(3+2\widetilde{\beta})-\theta}{2N}} \\ &\leq C_N\mathcal{B}_{\mathbf{n}}^{\frac{N(3+2\widetilde{\beta})-\theta}{2N}}. \end{aligned}$$

Now, since in this case $\theta > N(q+2\widetilde{\beta}+1)$, $q > 1$, we have $\dfrac{N(3+2\widetilde{\beta})-\theta}{2N} < 0$. Assumption **H7** allows us to say that $\mathcal{B}_{\mathbf{n}} \to +\infty$, as $\mathbf{n} \to \infty$, then $\widehat{\mathbf{n}}u_{\mathbf{n}}h_{\mathbf{n}} \to 0$. Consequently,

$$\sum_{\mathbf{n}\in\mathbb{N}^N} h_{\mathbf{n}} < \sum_{\mathbf{n}\in\mathbb{N}^N}\frac{1}{\widehat{\mathbf{n}}u_{\mathbf{n}}} < +\infty.$$

Then the two assumptions **H6** and **H7** lead to

$$\sum_{\mathbf{n}\in\mathbb{N}^N} C_N\widehat{\mathbf{n}}^{-a} + 2^{N+2}\frac{C}{a_{\mathbf{n},\mathbf{j}}b_{\mathbf{n}}^d}\psi([\widehat{\mathbf{t}}-1]p^N,p^N)\varphi(p)\widehat{\mathbf{n}}\delta^{-1}\left(\frac{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}{\log\widehat{\mathbf{n}}}\right)^{1/2} < \infty$$

Let us now consider $\mathbb{P}\left([\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} W_{\mathbf{ni}}=0]\right)$. We have

$$\begin{aligned}\mathbb{P}\left(\left[\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} W_{\mathbf{ni}}=0\right]\right) &\leq \mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} K_{1\mathbf{i}}K_{2\mathbf{i}} \leq \frac{ua_{\mathbf{n},\mathbf{j}}}{2}\right]\\ &= \mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} (K_{1\mathbf{i}}K_{2\mathbf{i}} - \mathbb{E}(K_{1\mathbf{i}}K_{2\mathbf{i}})) \leq \frac{ua_{\mathbf{n},\mathbf{j}}}{2} - ua_{\mathbf{n},\mathbf{j}}\right]\\ &\leq \mathbb{P}\left[\left|S_\mathbf{n}(x_\mathbf{j})\right| > \epsilon\right] \qquad \text{for } \mathbf{n} \text{ large enough,}\end{aligned}$$

where $S_\mathbf{n}(x_\mathbf{j}) = \sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} \mathbf{\Lambda}_\mathbf{i}(x_\mathbf{j}) = f_\mathbf{n}(x_\mathbf{j}) - \mathbb{E}[f_\mathbf{n}(x_\mathbf{j})]$, $\epsilon = \epsilon_\mathbf{n}$ and p are the same as above. So the last term of (3.4) is a.c. zero for large $\mathbf{n}$. This ends the proof.

■

Proof of Lemma 3.8

From the sequel of calculations $Var(f_\mathbf{n}(x_\mathbf{j}) - \mathbb{E}[f_\mathbf{n}(x_\mathbf{j})]) = \mathbb{E}\left[\left(f_\mathbf{n}(x_\mathbf{j}) - \mathbb{E}[f_\mathbf{n}(x_\mathbf{j})]\right)^2\right] = \mathbb{E}\left[\left(\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} \Lambda_\mathbf{i}(x_\mathbf{j})\right)^2\right] \leq \mathbf{I}_\mathbf{n}(x_\mathbf{j}) + \mathbf{R}_\mathbf{n}(x_\mathbf{j})$ where $\mathbf{I}_\mathbf{n}(x_\mathbf{j}) = \sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} \mathbb{E}\left[\left(\Lambda_\mathbf{i}(x_\mathbf{j})\right)^2\right]$ and $\mathbf{R}_\mathbf{n}(x_\mathbf{j}) = \sum_{\mathbf{i},\mathbf{k}\in\mathcal{I}_\mathbf{n}}\sum_{\mathbf{i}\neq\mathbf{k}} \left|\mathbb{E}\left[\Lambda_\mathbf{i}(x_\mathbf{j})\Lambda_\mathbf{k}(x_\mathbf{j})\right]\right|$.

$$\widehat{\mathbf{n}}b_\mathbf{n}^d\rho_\mathbf{n}^N\mathbf{I}_\mathbf{n}(x_\mathbf{j}) = \widehat{\mathbf{n}}b_\mathbf{n}^d\rho_\mathbf{n}^N\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}}\mathbb{E}\left(\lambda_\mathbf{i}(x_\mathbf{j})\right)^2 - \widehat{\mathbf{n}}b_\mathbf{n}^d\rho_\mathbf{n}^N\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}}\left(\mathbb{E}\lambda_\mathbf{i}(x_\mathbf{j})\right)^2$$

$$\begin{aligned}\diamondsuit\quad \widehat{\mathbf{n}}b_\mathbf{n}^d\rho_\mathbf{n}^N\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}}\mathbb{E}\left[\lambda_\mathbf{i}(x_\mathbf{j})\right]^2 &= \widehat{\mathbf{n}}b_\mathbf{n}^d\rho_\mathbf{n}^N\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}}\mathbb{E}\left[\frac{1}{a_{\mathbf{n},\mathbf{j}}b_\mathbf{n}^d}K_1\left(\frac{x_\mathbf{j}-X_\mathbf{i}}{b_\mathbf{n}}\right)K_{2,\rho_\mathbf{n}}(\|\mathbf{i}-\mathbf{j}\|)\right]^2\\ &= \frac{\widehat{\mathbf{n}}b_\mathbf{n}^d\rho_\mathbf{n}^N}{(a_{\mathbf{n},\mathbf{j}}b_\mathbf{n}^d)^2}\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}}K_{2,\rho_\mathbf{n}}^2(\|\mathbf{i}-\mathbf{j}\|)\mathbb{E}\left[K_1\left(\frac{x_\mathbf{j}-X_\mathbf{i}}{b_\mathbf{n}}\right)\right]^2\\ &\leq C\frac{\widehat{\mathbf{n}}b_\mathbf{n}^d\rho_\mathbf{n}^N}{(a_{\mathbf{n},\mathbf{j}}b_\mathbf{n}^d)^2}\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}}K_{2,\rho_\mathbf{n}}(\|\mathbf{i}-\mathbf{j}\|)\int K_1^2(t)f(x_\mathbf{j}-b_\mathbf{n}t)(-b_\mathbf{n}^d)dt\\ &\leq C\frac{\widehat{\mathbf{n}}\rho_\mathbf{n}^N}{a_{\mathbf{n},\mathbf{j}}}\int K_1^2(t)f(x_\mathbf{j}-b_\mathbf{n}t)dt.\end{aligned}$$

Since $\frac{\widehat{\mathbf{n}}\rho_\mathbf{n}^N}{a_{\mathbf{n},\mathbf{j}}} \to C_N$, $\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}}\mathbb{E}\left[\lambda_\mathbf{i}(x_\mathbf{j})\right]^2 = O\left(\frac{1}{\widehat{\mathbf{n}}b_\mathbf{n}^d\rho_\mathbf{n}^N}\right)$ by assumptions on K_1 and f.

$$\begin{aligned}\diamondsuit\quad \widehat{\mathbf{n}}b_\mathbf{n}^d\rho_\mathbf{n}^N\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}}\left(\mathbb{E}\lambda_\mathbf{i}(x_\mathbf{j})\right)^2 &= \frac{\widehat{\mathbf{n}}b_\mathbf{n}^d\rho_\mathbf{n}^N}{(a_{\mathbf{n},\mathbf{j}}b_\mathbf{n}^d)^2}\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}}K_{2,\rho_\mathbf{n}}^2(\|\mathbf{i}-\mathbf{j}\|)\left(\mathbb{E}\left[K_1\left(\frac{x_\mathbf{j}-X_\mathbf{i}}{b_\mathbf{n}}\right)\right]\right)^2\\ &\leq C\frac{\widehat{\mathbf{n}}b_\mathbf{n}^d\rho_\mathbf{n}^N}{(a_{\mathbf{n},\mathbf{j}}b_\mathbf{n}^d)^2}\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}}K_{2,\rho_\mathbf{n}}(\|\mathbf{i}-\mathbf{j}\|)\left(\mathbb{E}\left[K_1\left(\frac{x_\mathbf{j}-X_\mathbf{i}}{b_\mathbf{n}}\right)\right]\right)^2\\ &\leq C\frac{\widehat{\mathbf{n}}\rho_\mathbf{n}^N}{a_{\mathbf{n},\mathbf{j}}b_\mathbf{n}^d}\left(\int K_1(u)f(x_\mathbf{j}-ub_\mathbf{n})(-b_\mathbf{n}^d)du\right)^2 = O(b_\mathbf{n}^d)\end{aligned}$$

and we have by assumption $\lim_{\mathbf{n}\to\infty} \widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \left(\mathbb{E}\lambda_{\mathbf{i}}(x_{\mathbf{j}})\right)^2 = 0$. We now treat the term $\mathbf{R}_{\mathbf{n}}(x_{\mathbf{j}})$. Let $D_{\mathbf{n}}$ be a sequence of real numbers tending to ∞ as $\mathbf{n} \to \infty$. Let $S = \{\mathbf{i}, \mathbf{k} \in \mathcal{I}_{\mathbf{n}}, \|\mathbf{i}-\mathbf{k}\| \leq D_{\mathbf{n}}\}$ and denote by S^c the complementary of S. Let $\mathbf{R}_{\mathbf{n}}^{(1)}(x_{\mathbf{j}}) = \sum_{\mathbf{i},\mathbf{k}\in S} |\mathbb{E}\Lambda_{\mathbf{i}}(x_{\mathbf{j}})\Lambda_{\mathbf{k}}(x_{\mathbf{j}})|$ and $\mathbf{R}_{\mathbf{n}}^{(2)}(x_{\mathbf{j}}) = \sum_{\mathbf{i},\mathbf{k}\in S^c} |\mathbb{E}\Lambda_{\mathbf{i}}(x_{\mathbf{j}})\Lambda_{\mathbf{k}}(x_{\mathbf{j}})|$. Hence, $\mathbf{R}_{\mathbf{n}}(x_{\mathbf{j}}) \leq \mathbf{R}_{\mathbf{n}}^{(1)}(x_{\mathbf{j}}) + \mathbf{R}_{\mathbf{n}}^{(2)}(x_{\mathbf{j}})$.

$$\begin{aligned}
\diamondsuit\ \mathbf{R}_{\mathbf{n}}^{(1)}(x_{\mathbf{j}}) &= \sum_{\mathbf{i},\mathbf{k}\in S} |\mathbb{E}[\lambda_{\mathbf{i}}(x_{\mathbf{j}})\lambda_{\mathbf{k}}(x_{\mathbf{j}})] - \mathbb{E}\lambda_{\mathbf{i}}(x_{\mathbf{j}})\mathbb{E}\lambda_{\mathbf{k}}(x_{\mathbf{j}})| \\
&= \sum_{\mathbf{i},\mathbf{k}\in S} |\mathbf{A}-\mathbf{B}|
\end{aligned}$$

$$\begin{aligned}
\mathbf{A} &= \mathbb{E}\left(\frac{1}{a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|) \frac{1}{a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{k}}}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{k}-\mathbf{j}\|)\right) \\
&= \frac{1}{(a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d)^2} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{k}-\mathbf{j}\|)\mathbf{A}'. \\
\mathbf{A}' &= \int\int K_1\left(\frac{x_{\mathbf{j}} - u}{b_{\mathbf{n}}}\right) K_1\left(\frac{x_{\mathbf{j}} - v}{b_{\mathbf{n}}}\right) f_{X_{\mathbf{i}}X_{\mathbf{k}}}(u,v)dudv \\
\mathbf{B} &= \mathbb{E}\left[\frac{1}{a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|)\right] \mathbb{E}\left[\frac{1}{a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{k}}}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{k}-\mathbf{j}\|)\right] \\
&= \frac{1}{(a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d)^2} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{k}-\mathbf{j}\|)\mathbf{B}' \\
\mathbf{B}' &= \left(\int K_1\left(\frac{x_{\mathbf{j}} - u}{b_{\mathbf{n}}}\right) f_{X_{\mathbf{i}}}(u)du\right)\left(\int K_1\left(\frac{x_{\mathbf{j}} - v}{b_{\mathbf{n}}}\right) f_{X_{\mathbf{k}}}(v)dv\right)
\end{aligned}$$

By assumption **H5**, we have $\sup_{u,\,v} \sup_{\mathbf{i},\,\mathbf{k}} |f_{X_{\mathbf{i}}X_{\mathbf{k}}}(u,v) - f_{X_{\mathbf{i}}}(u) f_{X_{\mathbf{k}}}(v)| \leq C$, then

$$\begin{aligned}
|\mathbf{A}' - \mathbf{B}'| &\leq \int\int K_1\left(\frac{x_{\mathbf{j}} - u}{b_{\mathbf{n}}}\right) K_1\left(\frac{x_{\mathbf{j}} - v}{b_{\mathbf{n}}}\right) |f_{X_{\mathbf{i}}X_{\mathbf{k}}}(u,v) - f_{X_{\mathbf{i}}}(u) f_{X_{\mathbf{k}}}(v)| dudv \\
&\leq C \int\int K_1\left(\frac{x_{\mathbf{j}} - u}{b_{\mathbf{n}}}\right) K_1\left(\frac{x_{\mathbf{j}} - v}{b_{\mathbf{n}}}\right) dudv
\end{aligned}$$

Thus

$$\mathbf{R}_{\mathbf{n}}^{(1)}(x_{\mathbf{j}}) \leq C \sum_{\mathbf{i},\mathbf{k}\in S} \frac{1}{(a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d)^2} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{k}-\mathbf{j}\|) \int\int K_1\left(\frac{x_{\mathbf{j}} - u}{b_{\mathbf{n}}}\right) K_1\left(\frac{x_{\mathbf{j}} - v}{b_{\mathbf{n}}}\right) dudv$$

and assumption **H4** leads to

$$\begin{aligned}
\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N \mathbf{R}_{\mathbf{n}}^{(1)}(x_{\mathbf{j}}) &\leq C \frac{\widehat{\mathbf{n}} b_{\mathbf{n}}^{3d} \rho_{\mathbf{n}}^N}{(a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d)^2} \sum_{\mathbf{i},\mathbf{k}\in S} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{k}-\mathbf{j}\|) \left(\int |K_1(u)|\, du\right)^2 \\
&\leq C \frac{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}{a_{\mathbf{n},\mathbf{j}}^2} \sum_{\mathbf{i},\mathbf{k}\in S} \mathbf{1}_{[0,1]}\left(\rho_{\mathbf{n}}^{-1} \left\|\frac{\mathbf{i}-\mathbf{j}}{\mathbf{n}}\right\|\right) \mathbf{1}_{[0,1]}\left(\rho_{\mathbf{n}}^{-1} \left\|\frac{\mathbf{k}-\mathbf{j}}{\mathbf{n}}\right\|\right) \left(\int |K_1(u)|\, du\right)^2 \\
&\leq C \frac{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}{a_{\mathbf{n},\mathbf{j}}^2} \sum_{\mathbf{i},\mathbf{k}\in V_{\mathbf{j}}} \mathbf{1}_{[0,1]}\left(\frac{\|\mathbf{k}-\mathbf{i}\|}{D_{\mathbf{n}}}\right) \left(\int |K_1(u)|\, du\right)^2 \\
&\leq 2^N C \frac{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}{a_{\mathbf{n},\mathbf{j}}^2} \sum_{\mathbf{i}\in V_{\mathbf{j}}} \sum_{\mathbf{i}-\mathbf{u}\in V_{\mathbf{j}}} \mathbf{1}_{\{\mathbf{u};\|\mathbf{u}\|\leq D_{\mathbf{n}}\}} \left(\int |K_1(u)|\, du\right)^2
\end{aligned}$$

where $V_{\mathbf{j}} = \left\{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}, \left\|\frac{\mathbf{i}-\mathbf{j}}{\mathbf{n}}\right\| \leq \rho_{\mathbf{n}}\right\}$ with $\mathrm{Card}(V_{\mathbf{j}}) = k_{\mathbf{n}}$. Then $\sum_{\mathbf{i}\in V_{\mathbf{j}}} \sum_{\mathbf{i}-\mathbf{u}\in V_{\mathbf{j}}} \mathbf{1}_{\{\mathbf{u};\|\mathbf{u}\|\leq D_{\mathbf{n}}\}} \leq k_{\mathbf{n}}[D_{\mathbf{n}}]^N$. So, $\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N \mathbf{R}_{\mathbf{n}}^{(1)}(x_{\mathbf{j}}) \leq C_N b_{\mathbf{n}}^d D_{\mathbf{n}}^N$, for $\mathbf{n}$ large enough.

$\diamondsuit$ Since the functions $K_1(\cdot)$ and $K_2(\cdot)$ are bounded, we get by applying Lemma A.1

$$|\mathbb{E}\Lambda_{\mathbf{i}}(x_{\mathbf{j}})\Lambda_{\mathbf{k}}(x_{\mathbf{j}})| \leq C \frac{K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{k}-\mathbf{j}\|)}{(a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d)^2} \psi(1,1)\varphi(\|\mathbf{i}-\mathbf{k}\|)$$

$$\begin{aligned}
\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N \mathbf{R}_{\mathbf{n}}^{(2)}(x_{\mathbf{j}}) &\leq \widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N \sum_{\mathbf{i},\mathbf{k}\in S^c} C \frac{K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i}-\mathbf{j}\|) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{k}-\mathbf{j}\|)}{(a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d)^2} \psi(1,1)\varphi(\|\mathbf{i}-\mathbf{k}\|) \\
&\leq C \frac{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}{(a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d)^2} \sum_{\mathbf{i},\mathbf{k}\in S^c\cap V_{\mathbf{j}}} \varphi(\|\mathbf{i}-\mathbf{k}\|) \\
&\leq 2^N C \frac{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}{(a_{\mathbf{n},\mathbf{j}} b_{\mathbf{n}}^d)^2} \sum_{\mathbf{k}\in V_{\mathbf{j}}} \sum_{\mathbf{k}-\mathbf{u}\in V_{\mathbf{j}}, \|\mathbf{u}\|>D_{\mathbf{n}}} \varphi(\|\mathbf{u}\|) \\
&\leq C \frac{k_{\mathbf{n}} \widehat{\mathbf{n}} \rho_{\mathbf{n}}^N}{(a_{\mathbf{n},\mathbf{j}})^2 b_{\mathbf{n}}^d} \sum_{\|\mathbf{i}\|>D_{\mathbf{n}}} \varphi(\|\mathbf{i}\|) \\
&\left(\text{because} \forall\, \mathbf{k} \in \mathbb{N}^N, 2^N \sum_{\mathbf{k}-\mathbf{u}\in V_{\mathbf{j}}, \|\mathbf{u}\|>D_{\mathbf{n}}} \varphi(\|\mathbf{u}\|) = \sum_{\|\mathbf{i}\|>D_{\mathbf{n}}} \varphi(\|\mathbf{i}\|)\right)
\end{aligned}$$

Since $\sum_{\mathbf{i}>D_{\mathbf{n}}} \varphi(\mathbf{i}) \leq C \sum_{\mathbf{i}>D_{\mathbf{n}}} \mathbf{i}^{-\theta} \leq C \sum_{\mathbf{i}>D_{\mathbf{n}}} \mathbf{i}^{-\theta}\mathbf{i}^{-N}\mathbf{i}^N$ and $\|\mathbf{i}\| > D_{\mathbf{n}}$, $\|\mathbf{i}\|^{-N} \leq (D_{\mathbf{n}})^{-N}$, we have

$$C \sum_{\mathbf{i}>D_{\mathbf{n}}} \mathbf{i}^{-\theta}\mathbf{i}^{-N}\mathbf{i}^N \leq C D_{\mathbf{n}}^{-N} \sum_{\mathbf{i}>D_{\mathbf{n}}} \mathbf{i}^{-\theta}\mathbf{i}^N \leq C D_{\mathbf{n}}^{-N} \sum_{\mathbf{i}>D_{\mathbf{n}}} \mathbf{i}^{N-\theta}.$$

Consequently, we have

$$\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N \mathbf{R}_{\mathbf{n}}^{(2)}(x_{\mathbf{j}}) \leq \frac{C_N}{b_{\mathbf{n}}^d} D_{\mathbf{n}}^{-N} \sum_{\mathbf{i}>D_{\mathbf{n}}} \mathbf{i}^{N-\theta}, \text{ for } \mathbf{n} \text{ large enough.}$$

The fact that $\theta > N(2+q) > N+1$ leads to choose $D_{\mathbf{n}} = (b_{\mathbf{n}}^d)^{-1/N}$ which gives the expected result. In fact, we have then $\lim_{\mathbf{n}\to\infty} \widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N \mathbf{R}_{\mathbf{n}}^{(2)}(x_{\mathbf{j}}) = C_N$ and $\lim_{\mathbf{n}\to\infty} \widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N \mathbf{R}_{\mathbf{n}}^{(1)}(x_{\mathbf{j}}) = C_N$. So, $\mathbf{R}_{\mathbf{n}}^{(1)}(x_{\mathbf{j}}) = O\left(\frac{1}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}\right)$, $\mathbf{R}_{\mathbf{n}}^{(2)} = O\left(\frac{1}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}\right)$ and then, $\mathbf{R}_{\mathbf{n}}(x_{\mathbf{j}}) = O\left(\frac{1}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}\right)$. That shows that $\mathbf{I}_{\mathbf{n}}(x_{\mathbf{j}}) + \mathbf{R}_{\mathbf{n}}(x_{\mathbf{j}}) = O\left(\frac{1}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N}\right)$ for $\mathbf{n}$ enough large. ■

Proof of Theorem 3.4.

We can write $\left(\mathbb{E}|r_{\mathbf{n}}(x_{\mathbf{j}}) - r(x_{\mathbf{j}})|^q\right)^{1/q} = \|r_{\mathbf{n}}(x_{\mathbf{j}}) - r(x_{\mathbf{j}})\|_q$. Recall that $W_{\mathbf{ni}} = \frac{K_{1\mathbf{i}} K_{2\mathbf{i}}}{\sum_{\mathbf{k}\in\mathcal{I}_{\mathbf{n}}} K_{1\mathbf{k}} K_{2\mathbf{k}}}$. By adopting the convention $0/0 = 0$, we have $\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}} = 0$ or 1. Therefore

$$\begin{aligned}
\mathbb{E}^{1/q}\left[r_{\mathbf{n}}(x_{\mathbf{j}}) - r(x_{\mathbf{j}})\right]^q &\leq \mathbb{E}^{1/q}\left[\left(\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}}\left[\mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}}) - r(x_{\mathbf{j}})\right]\right)\mathbf{1}_{\left[\sum_{\mathbf{i}} W_{\mathbf{ni}}=1\right]}\right]^q \\
&+ \mathbb{E}^{1/q}\left[\left(\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}}\left[Y_{\mathbf{i}} - \mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}})\right]\right)\mathbf{1}_{\left[\sum_{\mathbf{i}} W_{\mathbf{ni}}=1\right]}\right]^q \\
&+ \mathbb{E}^{1/q}\left[\left(\frac{1}{\widehat{\mathbf{n}}}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} - r(x_{\mathbf{j}})\right)\mathbf{1}_{\left[\sum_{\mathbf{i}} W_{\mathbf{ni}}=0\right]}\right]^q
\end{aligned}$$

applying Minkowski's inequality. The terms on the right of the previous inequality are treated in Lemmas 3.9, 3.10, and 3.13 respectively and give the result of Theorem 3.4. ■

Proof of Lemma 3.9

$$\begin{aligned}
\mathbb{E}^{1/q}\left[\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} W_{\mathbf{ni}}\mathbb{E}(Y_\mathbf{i}|X_\mathbf{i}) - r(x_\mathbf{j})\right]^q &= \mathbb{E}^{1/q}\left[\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} W_{\mathbf{ni}} r(X_\mathbf{i}) - r(x_\mathbf{j})\right]^q \\
&= \mathbb{E}^{1/q}\left[\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} W_{\mathbf{ni}}\mathbf{1}_{[\|X_\mathbf{i}-x_\mathbf{j}\|\le b_\mathbf{n}]}\left(r(X_\mathbf{i}) - r(x_\mathbf{j})\right)\right]^q \\
&\le \mathbb{E}^{1/q}\left[C\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} W_{\mathbf{ni}} b_\mathbf{n}\right]^q \\
&\le \mathbb{E}^{1/q}\left[Cb_\mathbf{n}\right]^q \le Cb_\mathbf{n} = O(b_\mathbf{n})
\end{aligned}$$

by assumptions **H2** (Lipschitz condition) and **H4**. ■

Proof of Lemma 3.10

Let $D = \sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} W_{\mathbf{ni}}\left(Y_\mathbf{i} - \mathbb{E}(Y_\mathbf{i}|X_\mathbf{i})\right) = \dfrac{e_\mathbf{n}(x_\mathbf{j})}{f_\mathbf{n}(x_\mathbf{j})}$ with

$$e_\mathbf{n}(x_\mathbf{j}) = \frac{1}{a_{\mathbf{n},\mathbf{j}} b_\mathbf{n}^d}\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} K_{1\mathbf{i}}K_{2\mathbf{i}}\left[Y_\mathbf{i} - \mathbb{E}(Y_\mathbf{i}|X_\mathbf{i})\right].$$

We note that $\quad \forall \mathbf{i}\colon 0 \le |Y_\mathbf{i} - \mathbb{E}(Y_\mathbf{i}|X_\mathbf{i})| \le C$ then, $|D| \le \sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} W_{\mathbf{ni}}\, C \le C$.

$$\begin{aligned}
|D| &= |D|\mathbf{1}_{\left[\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} K_{1\mathbf{i}}K_{2\mathbf{i}} > c\right]} + |D|\mathbf{1}_{\left[\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} K_{1\mathbf{i}}K_{2\mathbf{i}} \le c\right]} \\
&\le \frac{|e_\mathbf{n}(x_\mathbf{j})|}{f_\mathbf{n}(x_\mathbf{j})}\mathbf{1}_{\left[\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} K_{1\mathbf{i}}K_{2\mathbf{i}} > c\right]} + C\mathbf{1}_{\left[\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} K_{1\mathbf{i}}K_{2\mathbf{i}} \le c\right]}
\end{aligned}$$

where c is a given constant. We take $c = \dfrac{ua_{\mathbf{n},\mathbf{j}}}{2}$ with $u = \mathbb{E}[K_{1\mathbf{i}}] = \int K_1\left(\frac{x_\mathbf{j}-v}{b_\mathbf{n}}\right) f(v)dv$. By assumption **H4**,

$$\begin{aligned}
C\int \mathbf{1}_{[0,1]}\left(\left\|\frac{x_\mathbf{j}-v}{b_\mathbf{n}}\right\|\right) f(v)dv \le u \le C'\int \mathbf{1}_{[0,1]}\left(\left\|\frac{x_\mathbf{j}-v}{b_\mathbf{n}}\right\|\right) f(v)dv \\
Cb_\mathbf{n}^d\int \mathbf{1}_{[0,1]}\left(\|t\|\right) f(x_\mathbf{j} - tb_\mathbf{n})dt \le u \le C'b_\mathbf{n}^d\int \mathbf{1}_{[0,1]}\left(\|t\|\right) f(x_\mathbf{j} - tb_\mathbf{n})dt.
\end{aligned}$$

Since $f(\cdot)$ is bounded, there exist two constants c_1 and c_2 such as $c_1 b_\mathbf{n}^d \le u \le c_2 b_\mathbf{n}^d$. If $\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} K_{1\mathbf{i}}K_{2\mathbf{i}} > \frac{ua_{\mathbf{n},\mathbf{j}}}{2}$ then, $f_\mathbf{n}(x_\mathbf{j}) > \frac{ua_{\mathbf{n},\mathbf{j}}}{2a_{\mathbf{n},\mathbf{j}}b_\mathbf{n}^d} > \frac{u}{2b_\mathbf{n}^d} > \frac{c_1 b_\mathbf{n}^d}{2b_\mathbf{n}^d} > C$ and $\frac{|e_\mathbf{n}(x_\mathbf{j})|}{f_\mathbf{n}(x_\mathbf{j})} < C|e_\mathbf{n}(x_\mathbf{j})|$, for $\mathbf{n}$ wide enough. Consequently,

$$\|D\|_q \le C\|e_\mathbf{n}(x_\mathbf{j})\|_q + C\left(\mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} K_{1\mathbf{i}}K_{2\mathbf{i}} \le \frac{ua_{\mathbf{n},\mathbf{j}}}{2}\right]\right)^{1/q}$$

$$\|e_\mathbf{n}(x_\mathbf{j})\|_q = \frac{1}{a_{\mathbf{n},\mathbf{j}}b_\mathbf{n}^d}\left[\mathbb{E}\left(\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} K_{1\mathbf{i}}K_{2\mathbf{i}}\theta_\mathbf{i}\right)^q\right]^{1/q} = \frac{1}{a_{\mathbf{n},\mathbf{j}}b_\mathbf{n}^d}\left[\mathbb{E}\left(\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} \xi_\mathbf{i}\right)^q\right]^{1/q},$$

where $\xi_{\mathbf{i}} = K_{1\mathbf{i}}K_{2\mathbf{i}}\theta_{\mathbf{i}}$, with $\theta_{\mathbf{i}} = Y_{\mathbf{i}} - \mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}})$. We note that $\theta_{\mathbf{i}}$ and $\xi_{\mathbf{i}}$ introduced above are similar as those used in Lemma 2.2 in Gao et al. (2008). According to Lemma 3.11 and Lemma 3.12 respectively, from a certain rank, we have (for $\mathbf{n}$ wide enough) $\|e_{\mathbf{n}}(x_{\mathbf{j}})\|_q = O\left(\left(\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N\right)^{-1/2}\right)$ and $\left(\mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{1\mathbf{i}}K_{2\mathbf{i}} \leq \frac{ua_{\mathbf{n},\mathbf{j}}}{2}\right]\right)^{1/q} = O\left(\left(\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N\right)^{-1/2}\right)$. To conclude, we have

$$\left\|\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}}\left(Y_{\mathbf{i}} - \mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}})\right)\right\|_q = O\left(\left(\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N\right)^{-1/2}\right).$$

■

Proof of Lemma 3.11

We are interested in the following term

$$\mathbb{E}\left(\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}\xi_{\mathbf{i}}\right)^q = \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}\mathbb{E}[\xi_{\mathbf{i}}^q] + \sum_{s=1}^{q-1}\sum_{v_0+v_1+\ldots+v_s=q}\sum_{\mathbf{i}_0\neq\ldots\neq\mathbf{i}_s}\mathbb{E}[\xi_{\mathbf{i}_0}^{v_0}\ldots\xi_{\mathbf{i}_s}^{v_s}].$$

Note that $\sum_{v_0+v_1+\ldots+v_s=q,}$ is the summation over $(v_0, v_1, \ldots, v_s)$ with positive integer components satisfying $v_0 + v_1 + \ldots + v_s = q$ and the summation $\sum_{\mathbf{i}_0\neq\ldots\neq\mathbf{i}_s}$ is over indexes $(\mathbf{i}_0, \mathbf{i}_1, \ldots, \mathbf{i}_s)$ with each index $\mathbf{i}_j$ taking values in $\mathbb{Z}^N$ from $\mathbf{1}$ to $\mathbf{n}$ and satisfying $\mathbf{i}_j \neq \mathbf{i}_l$ for any $j \neq l$, $0 \leq j, l \leq s$.
First, we have

$$\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}\mathbb{E}[\xi_{\mathbf{i}}^q] \leq \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}\mathbb{E}[(K_{1\mathbf{i}}K_{2\mathbf{i}}|\theta_{\mathbf{i}}|)^q] \leq Ck_{\mathbf{n}}\mathbb{E}[(K_{1\mathbf{i}})^q] \leq Ck_{\mathbf{n}}b_{\mathbf{n}}^d.$$

In the following, we assume that $q = 2r$, $r \geq 1$. If q takes values different than $2r$, just apply Hölder's inequality. Secondly, it is necessary to show, for positive integers $v_0, v_1, \ldots, v_s$, $(s = 1, \ldots, q-1)$, the following results

- $\mathbb{E}|\xi_{\mathbf{i}_1}^{v_1}\xi_{\mathbf{i}_2}^{v_2}\ldots\xi_{\mathbf{i}_s}^{v_s}| \leq Cb_{\mathbf{n}}^{ds}$,
- $\sum_{\mathbf{i}_0\neq\ldots\neq\mathbf{i}_s}\mathbb{E}[\xi_{\mathbf{i}_0}^{v_0}\ldots\xi_{\mathbf{i}_s}^{v_s}] = C((\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N)^{s+1})$, for $s = 1, 2, \ldots, r-1$ and $r > 1$,
- $\sum_{\mathbf{i}_0\neq\ldots\neq\mathbf{i}_s}\mathbb{E}[\xi_{\mathbf{i}_0}^{v_0}\ldots\xi_{\mathbf{i}_s}^{v_s}] = C((\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N)^{r})$, for $r \leq s \leq 2r-1$.

We have

$$\begin{aligned}
\diamondsuit \quad \mathbb{E}|\xi_{\mathbf{i}_1}^{v_1}\xi_{\mathbf{i}_2}^{v_2}\ldots\xi_{\mathbf{i}_s}^{v_s}| &= \mathbb{E}\left[\prod_{j=1}^s K_{1\mathbf{i}_j}^{v_j}K_{2\mathbf{i}_j}^{v_j}\prod_{j=1}^s|\theta_{\mathbf{i}_j}^{v_j}|\right] \leq \mathbb{E}\left[\prod_{j=1}^s K_{1\mathbf{i}_j}^{v_j}\prod_{j=1}^s|\theta_{\mathbf{i}_j}^{v_j}|\right] \leq C\,\mathbb{E}\left[\prod_{j=1}^s K_{1\mathbf{i}_j}^{v_j}\right] \\
&\leq C\int\cdots\int K_1^{v_1}\left(\frac{x_{\mathbf{j}}-u_1}{b_{\mathbf{n}}}\right)\ldots K_1^{v_s}\left(\frac{x_{\mathbf{j}}-u_s}{b_{\mathbf{n}}}\right)f(u_1,\ldots,u_s)du_1\ldots du_s \\
&\leq C\int\cdots\int K_1^{v_1}(t_1)\ldots K_1^{v_s}(t_s)f(x_{\mathbf{j}}-t_1b_{\mathbf{n}},\ldots,x_{\mathbf{j}}-t_sb_{\mathbf{n}})(-b_{\mathbf{n}}^d)^s dt_1\ldots dt_s \\
&\leq C\,(b_{\mathbf{n}}^d)^s.
\end{aligned}$$

For $s = 1, 2, \ldots, r-1$, we have

$$\begin{aligned}
\diamondsuit \quad \sum_{\mathbf{i}_0\neq\ldots\neq\mathbf{i}_s}\mathbb{E}[\xi_{\mathbf{i}_0}^{v_0}\ldots\xi_{\mathbf{i}_s}^{v_s}] &= \sum_{\mathbf{i}_0\neq\ldots\neq\mathbf{i}_s}\left[\mathbb{E}\left(\prod_{j=0}^s\xi_{\mathbf{i}_j}^{v_j}\right) - \prod_{j=0}^s\mathbb{E}(\xi_{\mathbf{i}_j}^{v_j})\right] + \sum_{\mathbf{i}_0\neq\ldots\neq\mathbf{i}_s}\prod_{j=0}^s\mathbb{E}(\xi_{\mathbf{i}_j}^{v_j}) \\
&= V_{s1} + V_{s2}.
\end{aligned}$$

$$
\begin{aligned}
|V_{s2}| &\leq \sum_{\mathbf{i}_0 \neq \ldots \neq \mathbf{i}_s} C(b_{\mathbf{n}}^d)^{s+1} \prod_{j=0}^{s} K_{2\mathbf{i}_j} \\
&\leq C(k_{\mathbf{n}} b_{\mathbf{n}}^d)^{s+1}
\end{aligned}
$$

Defining $\prod_{j=l}^{s} = 1$ if $l > s$, we have

$$
\begin{aligned}
V_{s1} &= \sum_{\mathbf{i}_0 \neq \ldots \neq \mathbf{i}_s} \left[\sum_{l=0}^{s-1} \left(\prod_{j=0}^{l-1} \mathbb{E}\xi_{\mathbf{i}_j}^{v_j} \right) \left(\mathbb{E}\left[\prod_{j=l}^{s} \xi_{\mathbf{i}_j}^{v_j} \right] - \mathbb{E}\left[\xi_{\mathbf{i}_l}^{v_l} \right] \mathbb{E}\left[\prod_{j=l+1}^{s} \xi_{\mathbf{i}_j}^{v_j} \right] \right) \right] \\
|V_{s1}| &\leq C \sum_{l=0}^{s-1} \left(k_{\mathbf{n}} b_{\mathbf{n}}^d \right)^l \sum_{\mathbf{i}_l \neq \ldots \neq \mathbf{i}_s} \left| \mathbb{E}\left[\prod_{j=l}^{s} \xi_{\mathbf{i}_j}^{v_j} \right] - \mathbb{E}\left[\xi_{\mathbf{i}_l}^{v_l} \right] \mathbb{E}\left[\prod_{j=l+1}^{s} \xi_{\mathbf{i}_j}^{v_j} \right] \right| \\
&\leq C \sum_{l=0}^{s-1} \left(k_{\mathbf{n}} b_{\mathbf{n}}^d \right)^l V_{ls1}
\end{aligned}
$$

Let D a real positive number, we have

$$
\begin{aligned}
V_{ls1} &= \sum_{0 < \mathrm{dist}(\{\mathbf{i}_l\}, \{\mathbf{i}_{l+1}, \ldots, \mathbf{i}_s\}) \leq D} \mathfrak{V}_{\mathbf{i}} + \sum_{\mathrm{dist}(\{\mathbf{i}_l\}, \{\mathbf{i}_{l+1}, \ldots, \mathbf{i}_s\}) > D} \mathfrak{V}_{\mathbf{i}} \\
&= V_{ls11} + V_{ls12}
\end{aligned}
$$

$$
\begin{aligned}
\mathfrak{V}_{\mathbf{i}} &= \left| \mathbb{E}\left[\prod_{j=l}^{s} \xi_{\mathbf{i}_j}^{v_j} \right] - \mathbb{E}\left[\xi_{\mathbf{i}_l}^{v_l} \right] \mathbb{E}\left[\prod_{j=l+1}^{s} \xi_{\mathbf{i}_j}^{v_j} \right] \right| \\
&\leq \left| \mathbb{E}\left[\prod_{j=l}^{s} \xi_{\mathbf{i}_j}^{v_j} \right] \right| + \left| \mathbb{E}\left[\xi_{\mathbf{i}_l}^{v_l} \right] \mathbb{E}\left[\prod_{j=l+1}^{s} \xi_{\mathbf{i}_j}^{v_j} \right] \right| \\
&\leq C(b_{\mathbf{n}}^d)^{s-l+1} \prod_{j=l}^{s} K_{2\mathbf{i}_j} + C K_{2\mathbf{i}_l} b_{\mathbf{n}}^d \times (b_{\mathbf{n}}^d)^{s-l} \prod_{j=l+1}^{s} K_{2\mathbf{i}_j} \\
&\leq C(b_{\mathbf{n}}^d)^{s-l+1} \prod_{j=l}^{s} K_{2\mathbf{i}_j}
\end{aligned}
$$

$$
\begin{aligned}
V_{ls11} &\leq \sum_{0 < \mathrm{dist}(\{\mathbf{i}_l\}, \{\mathbf{i}_{l+1}, \ldots, \mathbf{i}_s\}) \leq D} C(b_{\mathbf{n}}^d)^{s-l+1} \prod_{j=l}^{s} K_{2\mathbf{i}_j} \\
&\leq C(b_{\mathbf{n}}^d)^{s-l+1} \sum_{k=1}^{D} \sum_{k \leq \mathrm{dist}(\{\mathbf{i}_l\}, \{\mathbf{i}_{l+1}, \ldots, \mathbf{i}_s\}) = t < k+1} \prod_{j=l}^{s} K_{2\mathbf{i}_j}
\end{aligned}
$$

Since $\mathrm{dist}(\{\mathbf{i}_l\}, \{\mathbf{i}_{l+1}, \ldots, \mathbf{i}_s\}) = t$, then there exists a location $\mathbf{i}_j \in \{\mathbf{i}_{l+1}, \ldots, \mathbf{i}_s\}$, say $\mathbf{i}_{l+1}$, such that $\mathrm{dist}(\mathbf{i}_l, \mathbf{i}_{l+1}) = t$, and consequently

$$
\begin{aligned}
\sum_{k=1}^{D} \sum_{k \leq \mathrm{dist}(\{\mathbf{i}_l\}, \{\mathbf{i}_{l+1}, \ldots, \mathbf{i}_s\}) = t < k+1} \prod_{j=l}^{s} K_{2\mathbf{i}_j} &\leq \sum_{k=1}^{D} \sum_{\substack{\mathbf{i}_j = \mathbf{1}, \\ j = l+2, \ldots, s.}}^{\mathbf{n}} \sum_{k \leq \mathrm{dist}(\{\mathbf{i}_l\}, \{\mathbf{i}_{l+1}\}) = t < k+1} \prod_{j=l}^{s} K_{2\mathbf{i}_j} \\
&\leq k_{\mathbf{n}}^{s-l-1} \sum_{k=1}^{D} \sum_{k \leq \mathrm{dist}(\{\mathbf{i}_l\}, \{\mathbf{i}_{l+1}\}) = t < k+1} \prod_{j=l}^{s} K_{2\mathbf{i}_j}.
\end{aligned}
$$

Then,

$$\begin{aligned}
V_{ls11} &\leq C(b_{\mathbf{n}}^d)^{s-l+1}k_{\mathbf{n}}^{s-l-1}\sum_{k=1}^{D}\sum_{k\leq \mathrm{dist}(\{\mathbf{i}_l\},\{\mathbf{i}_{l+1}\})=t<k+1}\prod_{j=l}^{s}K_{2\mathbf{i}_j}\\
&\leq C(b_{\mathbf{n}}^d)^{s-l+1}k_{\mathbf{n}}^{s-l}\sum_{k=1}^{D}\sum_{k\leq \|\mathbf{u}\|=t<k+1}1\\
&\leq C(b_{\mathbf{n}}^d)^{s-l+1}k_{\mathbf{n}}^{s-l}\sum_{t=1}^{D}t^{N-1}\\
&\leq C(b_{\mathbf{n}}^d)^{s-l+1}k_{\mathbf{n}}^{s-l}D^N.
\end{aligned}$$

We deal now with the term V_{ls12}. Let $t=\mathrm{dist}(\{\mathbf{i}_l\},\{\mathbf{i}_{l+1},\ldots,\mathbf{i}_s\})$ and consider $\psi(n,m)\leq C\min(n,m)$ then by Lemma A.1, since the variables $\xi_{\mathbf{i}}$ are bounded, we have

$$\left|\mathbb{E}\left[\prod_{j=l}^{s}\xi_{\mathbf{i}_j}^{v_j}\right]-\mathbb{E}\left[\xi_{\mathbf{i}_l}^{v_l}\right]\mathbb{E}\left[\prod_{j=l+1}^{s}\xi_{\mathbf{i}_j}^{v_j}\right]\right|\leq C\,\psi\left(1,s-l\right)\varphi(t)\leq C\,\varphi(t)$$

$$\begin{aligned}
V_{ls12} &\leq \sum_{\mathrm{dist}(\{\mathbf{i}_l\},\{\mathbf{i}_{l+1},\ldots,\mathbf{i}_s\})>D}CK_{2\mathbf{i}_l}\,\varphi(t)\\
&\leq C\sum_{k=D+1}^{\infty}K_{2\mathbf{i}_l}\sum_{k\leq \mathrm{dist}(\{\mathbf{i}_l\},\{\mathbf{i}_{l+1}\})=t<k+1}\varphi(t)\\
&\leq C\sum_{k=D+1}^{\infty}k_{\mathbf{n}}^{s-l}\sum_{k\leq \|\mathbf{i}\|=t<k+1}\varphi(t)\\
&\leq C\,k_{\mathbf{n}}^{s-l}\sum_{t=D+1}^{\infty}t^{N-1}\,\varphi(t)
\end{aligned}$$

Combining the higher bound of V_{ls11} and of V_{ls12}, we have

$$\begin{aligned}
|V_{s1}| &\leq \sum_{l=0}^{s-1}\left(k_{\mathbf{n}}b_{\mathbf{n}}^d\right)^l\left(|C(b_{\mathbf{n}}^d)^{s-l+1}k_{\mathbf{n}}^{s-l}D^N|+|C\,k_{\mathbf{n}}^{s-l}\sum_{t=D+1}^{\infty}t^{N-1}\,\varphi(t)|\right)\\
&\leq C\,\left(k_{\mathbf{n}}b_{\mathbf{n}}^d\right)^{s+1}\sum_{l=0}^{s-1}\left(k_{\mathbf{n}}b_{\mathbf{n}}^d\right)^{l-s-1}\left((b_{\mathbf{n}}^d)^{s-l+1}k_{\mathbf{n}}^{s-l}D^N+k_{\mathbf{n}}^{s-l}\sum_{t=D+1}^{\infty}t^{N-1}\,\varphi(t)\right)\\
&\leq C\,\left(k_{\mathbf{n}}b_{\mathbf{n}}^d\right)^{s+1}\sum_{l=0}^{s-1}\left(k_{\mathbf{n}}^{-1}D^N+k_{\mathbf{n}}^{-1}\left(b_{\mathbf{n}}^d\right)^{l-s-1}\sum_{t=D+1}^{\infty}t^{N-1}\,\varphi(t)\right)\\
&\leq C\,\left(k_{\mathbf{n}}b_{\mathbf{n}}^d\right)^{s+1}\sum_{l=0}^{s-1}\left(k_{\mathbf{n}}^{-1}D^N+k_{\mathbf{n}}^{-1}\left(b_{\mathbf{n}}^d\right)^{l-s-1}D^{-sN}\sum_{t=D+1}^{\infty}t^{sN+N-1}\,\varphi(t)\right).
\end{aligned}$$

Taking $D=b_{\mathbf{n}}^{-d/N}$, we obtain

$$\begin{aligned}
|V_{s1}| &\leq C\,\left(k_{\mathbf{n}}b_{\mathbf{n}}^d\right)^{s+1}\sum_{l=0}^{s-1}\left(k_{\mathbf{n}}^{-1}(b_{\mathbf{n}}^{-d/N})^N+k_{\mathbf{n}}^{-1}\left(b_{\mathbf{n}}^d\right)^{l-s-1}(b_{\mathbf{n}}^{-d/N})^{-sN}\sum_{t=D+1}^{\infty}t^{sN+N-1}\,\varphi(t)\right)\\
&\leq C\,\left(k_{\mathbf{n}}b_{\mathbf{n}}^d\right)^{s+1}\sum_{l=0}^{s-1}\left(\frac{1}{k_{\mathbf{n}}b_{\mathbf{n}}^d}+k_{\mathbf{n}}^{-1}\left(b_{\mathbf{n}}^d\right)^{l-1}\sum_{t=D+1}^{\infty}t^{sN+N-1}\,\varphi(t)\right)
\end{aligned}$$

$$|V_{s1}| \leq C \left(k_{\mathbf{n}} b_{\mathbf{n}}^d\right)^{s+1} \sum_{l=0}^{s-1} \left(\frac{1}{k_{\mathbf{n}} b_{\mathbf{n}}^d} + \frac{1}{k_{\mathbf{n}} b_{\mathbf{n}}^d} \left(b_{\mathbf{n}}^d\right)^l \sum_{t=D+1}^{\infty} t^{sN+N-1} \varphi(t) \right)$$
$$\leq C \left(\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N\right)^{s+1} \sum_{l=0}^{s-1} \left(\frac{k_{\mathbf{n}}^s}{(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N)^{s+1} b_{\mathbf{n}}^d} + \frac{k_{\mathbf{n}}^s}{(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N)^{s+1} b_{\mathbf{n}}^d} \left(b_{\mathbf{n}}^d\right)^l \sum_{t=D+1}^{\infty} t^{sN+N-1} \varphi(t) \right)$$
$$\leq C \left(\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N\right)^{s+1} \sum_{l=0}^{s-1} \left(\left(\frac{k_{\mathbf{n}}}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N}\right)^s \frac{1}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N b_{\mathbf{n}}^d} + \left(\frac{k_{\mathbf{n}}}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N}\right)^s \frac{1}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N b_{\mathbf{n}}^d} \left(b_{\mathbf{n}}^d\right)^l \sum_{t=D+1}^{\infty} t^{sN+N-1-\theta} \right)$$

Then, $V_{s1} = O\left(\left(\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N\right)^{-s-1}\right)$ since $\varphi(t) \leq C t^{-\theta}$ and $\theta > N(2+q)$, $\frac{k_{\mathbf{n}}}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N} \to C$, $b_{\mathbf{n}} \to 0$, $\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N b_{\mathbf{n}}^d \to \infty$.

♢ For $r \leq s \leq 2r-1$, $\sum_{\mathbf{i}_0 \neq \ldots \neq \mathbf{i}_s} \mathbb{E}[\xi_{\mathbf{i}_0}^{v_0} \ldots \xi_{\mathbf{i}_s}^{v_s}]$.

As indicated in Gao et al. (2008) the arguments are similar for all $r \leq s \leq 2r-1$, therefore for simplicity, the proof is only given for $s = q-1$ and $N = 2$. Also, we suppose $v_0 = v_1 = \ldots = v_s = 1$. We use the following notation $W = \sum_{\mathbf{i}_0 \neq \ldots \neq \mathbf{i}_{2r-1}} \mathbb{E}[\xi_{\mathbf{i}_0} \ldots \xi_{\mathbf{i}_{2r-1}}]$ and we will denote $\mathbf{i} = (i,j) \in \mathbb{Z}^2$ and $\mathbf{i_k} = (i_k, j_k) \in \mathbb{Z}^2$.

$$\sum_{\mathbf{i}_0 \neq \ldots \neq \mathbf{i}_{2r-1}} \mathbb{E}[\xi_{\mathbf{i}_0} \ldots \xi_{\mathbf{i}_{2r-1}}] = \sum_{(i_0,j_0) \neq (i_1,j_1) \neq \ldots \neq (i_{2r-1}, j_{2r-1})} \mathbb{E}[\xi_{i_0,j_0} \xi_{i_1,j_1} \ldots \xi_{i_{2r-1},j_{2r-1}}].$$

To continue, we use a new order in $\mathbb{Z}^2$ (see Gao et al. (2008)). This order allows to separate the indices in two sets whose distances between locations are larger or smaller than P. Let us arrange each set of indices $\{i_0, i_1, \ldots, i_{2r-1}\}$ and $\{j_0, j_1, \ldots, j_{2r-1}\}$ in an increasing order such as $i_0 \leq i_1 \leq \ldots \leq i_{2r-1}$ and $j_{m_0} \leq j_{m_1} \leq \ldots \leq j_{m_{2r-1}}$. The number of such arrangements is at most $(2r)!$.

Let $\Delta_{i_k} = i_k - i_{k-1}$ and $\Delta_{j_{m_k}} = j_{m_k} - j_{m_{k-1}}$. Let us arrange $\{\Delta_{i_1}, \ldots, \Delta_{i_{2r-1}}\}$ and $\{\Delta_{j_{m_1}}, \ldots, \Delta_{j_{m_{2r-1}}}\}$ in decreasing order such as $\Delta_{i_{a_1}} \geq \Delta_{i_{a_2}} \geq \ldots \geq \Delta_{i_{a_{2r-1}}}$ and $\Delta_{j_{m_{b_1}}} \geq \Delta_{j_{m_{b_2}}} \geq \ldots \geq \Delta_{j_{m_{b_{2r-1}}}}$.

Let $t_1 = \Delta_{i_{a_r}}$, $t_2 = \Delta_{j_{m_{b_r}}}$ and $t = \max(t_1, t_2)$. If $t_1 > t_2$, then $t = t_1$, and
$0 \leq i_{a_k} - i_{a_k - 1} \leq t \leq n_1$, for $k = r+1, \ldots, 2r-1$,
$0 \leq j_{m_{b_k}} - j_{m_{b_k - 1}} \leq t \leq n_2$, for $k = r, r+1, \ldots, 2r-1$.
Therefore,
$i_{a_k - 1} \leq i_{a_k} \leq t + i_{a_k - 1}$, $j_{m_{b_k - 1}} \leq j_{m_{b_k}} \leq t + j_{m_{b_k - 1}}$, for $k = r+1, \ldots, 2r-1$.
Let us arrange $\mathbf{i}_0 \neq \mathbf{i}_1 \neq \ldots \neq \mathbf{i}_{2r-1}$ according to the order of $i_0 \leq i_1 \leq \ldots \leq i_{2r-1}$. We have

$$W = \sum_{\mathbf{i}_0 \neq \ldots \neq \mathbf{i}_{2r-1}} \mathbb{E}[\xi_{\mathbf{i}_0} \ldots \xi_{\mathbf{i}_{2r-1}}]$$
$$\leq C \sum_{1 \leq i_0 \leq i_1 \leq \ldots \leq i_{2r-1} \leq n_1} \sum_{1 \leq j_{m_0} \leq j_{m_1} \leq \ldots \leq j_{m_{2r-1}} \leq n_2} |\mathbb{E}[\xi_{i_0,j_0} \ldots \xi_{i_{2r-1},j_{2r-1}}]|.$$

We define
$\mathcal{I} = \{i_1, \ldots, i_{2r-1}\}$, $\mathcal{I}_a = \{i_{a_1}, \ldots, i_{a_r}\}$ and $\mathcal{I}_a^c = \{i_{a_{r+1}}, \ldots, i_{a_{2r-1}}\}$
$\mathcal{J} = \{j_{m_1}, \ldots, j_{m_{2r-1}}\}$, $\mathcal{J}_b = \{j_{m_{b_1}}, \ldots, j_{m_{b_r}}\}$ and $\mathcal{J}_b^c = \{j_{m_{b_{r+1}}}, \ldots, j_{m_{b_{2r-1}}}\}$
We note that $(i_{a_1}, \ldots, i_{a_{2r-1}})$ is a permutation of $\mathcal{I}$ and $(j_{m_{b_1}}, \ldots, j_{m_{b_{2r-1}}})$ is a permutation of $\mathcal{J}$. Since $i_{a_k - 1} \leq i_{a_k} \leq t + i_{a_k - 1}$, $j_{m_{b_k - 1}} \leq j_{m_{b_k}} \leq t + j_{m_{b_k - 1}}$, for

$k = r+1, \ldots, 2r-1$ and since $t = i_{a_r} - i_{a_{r-1}}$ and $t \geq j_{m_{b_r}} - j_{m_{b_{r-1}}} \geq 0$ we have

$$W \leq C \sum_{t=1}^{\max\{n_1,n_2\}} \sum_{i_0=1}^{n_1} \sum_{\substack{i=1,\\ i\in\mathcal{I}_a-\{i_{a_r}\}}}^{n_1} \sum_{\substack{i_{a_k}=i_{a_k-1},\\ k=r+1,\ldots,2r-1.}}^{i_{a_k-1}+t} \sum_{j_{m_0}=1}^{n_2} \sum_{\substack{j=1,\\ j\in\mathcal{J}_b-\{j_{m_{b_r}}\}}}^{n_2} \sum_{\substack{j_{m_{b_k}}=j_{m_{b_k-1}},\\ k=r,r+1,\ldots,2r-1.}}^{j_{m_{b_k-1}}+t} |\mathbb{E}[\xi_{\mathbf{i}_0}\xi_{\mathbf{i}_1}\ldots\xi_{\mathbf{i}_{2r-1}}]|.$$

Take a positive constant P such as $1 \leq P \leq \max(n_1, n_2)$ and separate into two parts the right term of the previous inequality in denoting them W_1 et W_2 according to $1 \leq t \leq P$ and $t > P$. Therefore, $W \leq W_1 + W_2$.

$$\begin{aligned}
W_1 &= C\sum_{t=1}^{P}\sum_{i_0=1}^{n_1} \sum_{\substack{i=1,\\ i\in\mathcal{I}_a-\{i_{a_r}\}}}^{n_1} \sum_{\substack{i_{a_k}=i_{a_k-1},\\ k=r+1,\ldots 2r-1.}}^{i_{a_k-1}+t} \sum_{j_{m_0}=1}^{n_2} \sum_{\substack{j=1,\\ j\in\mathcal{J}_b-\{j_{m_{b_r}}\}}}^{n_2} \sum_{\substack{j_{m_{b_k}}=j_{m_{b_k-1}},\\ k=r,r+1,\ldots 2r-1.}}^{j_{m_{b_k-1}}+t} |\mathbb{E}[\xi_{\mathbf{i}_0}\xi_{\mathbf{i}_1}\ldots\xi_{\mathbf{i}_{2r-1}}]| \\
&\leq C\sum_{t=1}^{P}\sum_{i_0=1}^{n_1} \sum_{\substack{i=1,\\ i\in\mathcal{I}_a-\{i_{a_r}\}}}^{n_1} \sum_{\substack{i_{a_k}=i_{a_k-1},\\ k=r+1,\ldots 2r-1.}}^{i_{a_k-1}+t} \sum_{j_{m_0}=1}^{n_2} \sum_{\substack{j=1,\\ j\in\mathcal{J}_b-\{j_{m_{b_r}}\}}}^{n_2} \sum_{\substack{j_{m_{b_k}}=j_{m_{b_k-1}},\\ k=r,r+1,\ldots 2r-1.}}^{j_{m_{b_k-1}}+t} b_{\mathbf{n}}^{2dr} K_{2\mathbf{i}_0}K_{2\mathbf{i}_1}\ldots K_{2\mathbf{i}_{2r-1}} \\
&\leq C(k_{\mathbf{n}})^r \sum_{t=1}^{P} t^{2r-1} b_{\mathbf{n}}^{2dr} \\
&\leq C_N(n_1n_2\rho_{\mathbf{n}}^N)^r P^{2r} b_{\mathbf{n}}^{2dr}
\end{aligned}$$

for $\mathbf{n}$ large enough using $k_{\mathbf{n}} = O(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N)$.
On the other hand, we assume that neither i_1 or i_{2r-1} belongs to $\mathcal{I}_a$. In this case, $\mathcal{I}_a$ is a subset of length r chosen among the $2r-3$ indices remaining (apart i_0, i_1 and i_{2r-1}). The first components of $\mathbf{i}_j$'s are stored in the following manner $i_0 \leq i_1 \leq \ldots \leq i_{k^*-1} \leq i_{k^*} \leq i_{k^*+1} \leq \ldots \leq i_{2r-1}$ for all k^* and $\Delta i_j = i_j - i_{j-1}$. Therefore we have

- $\text{dist}(\{\mathbf{i}_0, \ldots, \mathbf{i}_{k^*-1}\}, \{\mathbf{i}_{k^*}\}) \geq \Delta i_{k^*} \geq t$,
- $\text{dist}(\{\mathbf{i}_{k^*}\}, \{\mathbf{i}_{k^*+1}, \ldots, \mathbf{i}_{2r-1}\}) \geq \Delta i_{k^*+1} \geq t$, and
- $\text{dist}(\{\mathbf{i}_0, \ldots, \mathbf{i}_{k^*-1}\}, \{\mathbf{i}_{k^*}, \ldots, \mathbf{i}_{2r-1}\}) \geq \Delta i_{k^*} \geq t$, or
- $\text{dist}(\{\mathbf{i}_0\}, \{\mathbf{i}_1, \ldots, \mathbf{i}_{2r-1}\}) \geq \Delta i_1 \geq t$, or
- $\text{dist}(\{\mathbf{i}_0, \ldots, \mathbf{i}_{2r-2}\}, \{\mathbf{i}_{2r-1}\}) \geq \Delta i_{2r-1} \geq t$.

Let $A_{\mathbf{i}_{k^*-1}} = \xi_{\mathbf{i}_0}\ldots\xi_{\mathbf{i}_{k^*-1}}$ and $B_{\mathbf{i}_{k^*+1}} = \xi_{\mathbf{i}_{k^*+1}}\ldots\xi_{\mathbf{i}_{2r-1}}$, for the case i_{k^*} and i_{k^*+1} in $\mathcal{I}_a$. We have

$$\begin{aligned}
|\mathbb{E}\left[\xi_{\mathbf{i}_0}\ldots\xi_{\mathbf{i}_{2r-1}}\right]| &= |\mathbb{E}\left[A_{\mathbf{i}_{k^*-1}}\xi_{\mathbf{i}_{k^*}}B_{\mathbf{i}_{k^*+1}}\right]| \\
&\leq |\mathbb{E}\left[\left(A_{\mathbf{i}_{k^*-1}} - \mathbb{E}[A_{\mathbf{i}_{k^*-1}}]\right)\left(\xi_{\mathbf{i}_{k^*}}B_{\mathbf{i}_{k^*+1}} - \mathbb{E}[\xi_{\mathbf{i}_{k^*}}B_{\mathbf{i}_{k^*+1}}]\right)\right]| \\
&+ \quad |\mathbb{E}\left[A_{\mathbf{i}_{k^*-1}}\right]\mathbb{E}\left[\xi_{\mathbf{i}_{k^*}}B_{\mathbf{i}_{k^*+1}}\right]| \\
&= |Cov(A_{\mathbf{i}_{k^*-1}}, \xi_{\mathbf{i}_{k^*}}B_{\mathbf{i}_{k^*+1}}) + |\mathbb{E}\left[A_{\mathbf{i}_{k^*-1}}\right]| \; |Cov(\xi_{\mathbf{i}_{k^*}}B_{\mathbf{i}_{k^*+1}})| \\
&\leq C\varphi(t) + Cb_{\mathbf{n}}^{dk^*}\varphi(t) \leq C\varphi(t)
\end{aligned}$$

with $\left\|\frac{\mathbf{i}_l-\mathbf{j}}{\mathbf{n}}\right\| \leq \rho_{\mathbf{n}}, l=0,\dots,2r-1$. So,

$$W_2 = C\sum_{t=P+1}^{\infty}\sum_{i_0=1}^{n_1}\sum_{\substack{i=1,\\ i\in\mathcal{I}_a-\{i_{a_r}\}}}^{n_1}\sum_{\substack{i_{a_k}=i_{a_k-1},\\ k=r+1,\dots 2r-1.}}^{i_{a_k-1}+t}\sum_{j_{m_0}=1}^{n_2}\sum_{\substack{j=1,\\ j\in\mathcal{J}_b-\{j_{m_{b_r}}\}}}^{n_2}\sum_{\substack{j_{m_{b_k}}=j_{m_{b_k-1}},\\ k=r,r+1,\dots 2r-1.}}^{j_{m_{b_k-1}}+t}|\mathbb{E}[\xi_{\mathbf{i}_0}\xi_{\mathbf{i}_1}\dots\xi_{\mathbf{i}_{2r-1}}]|$$

$$\leq C(k_{\mathbf{n}})^r\sum_{t=P+1}^{\infty}t^{2r-1}\varphi(t) \leq C_N(n_1n_2\rho_{\mathbf{n}}^N)^r,$$

$\mathbf{n}$ large enough. It follows that

$$W \leq W_1+W_2 \leq C(n_1n_2\rho_{\mathbf{n}}^N)^rP^{2r}b_{\mathbf{n}}^{2dr} + C(n_1n_2\rho_{\mathbf{n}}^N)^r\sum_{t=P+1}^{\infty}t^{2r-1}\varphi(t)$$

$$\leq C(n_1n_2\rho_{\mathbf{n}}^Nb_{\mathbf{n}}^d)^r\left(P^{2r}b_{\mathbf{n}}^{dr}+b_{\mathbf{n}}^{-dr}\sum_{t=P+1}^{\infty}t^{2r-1-\theta}\right).$$

We obtained the result for $N=2$, and therefore for the general case N, we will obtain

$$W \leq C(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^Nb_{\mathbf{n}}^d)^r\left(P^{Nr}b_{\mathbf{n}}^{dr}+b_{\mathbf{n}}^{-dr}\sum_{t=P+1}^{\infty}t^{Nr-1-\theta}\right).$$

Taking $P=b_{\mathbf{n}}^{-d/N}$, we have

$$W \leq C(\widehat{\mathbf{n}}b_{\mathbf{n}}^d)^r\left(P^{Nr}b_{\mathbf{n}}^{dr}+b_{\mathbf{n}}^{-dr}P^{-Nr}\sum_{t=P+1}^{\infty}t^{2Nr-1-\theta}\right)$$

$$\leq C(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^Nb_{\mathbf{n}}^d)^r\left((b_{\mathbf{n}}^{-d/N})^{Nr}b_{\mathbf{n}}^{dr}+b_{\mathbf{n}}^{-dr}(b_{\mathbf{n}}^{-d/N})^{-Nr}\sum_{t=P+1}^{\infty}t^{2Nr-1-\theta}\right)$$

$$\leq C(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^Nb_{\mathbf{n}}^d)^r\left(1+\sum_{t=P+1}^{\infty}t^{2Nr-1-\theta}\right)$$

$$\leq C(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^Nb_{\mathbf{n}}^d)^r, \quad \text{since } \theta > N(2+q).$$

To conclude, we have

$$\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}\mathbb{E}[\xi_{\mathbf{i}}^q] \leq C\widehat{\mathbf{n}}\rho_{\mathbf{n}}^Nb_{\mathbf{n}}^d$$

$$\mathbb{E}|\xi_{\mathbf{i}_1}^{v_1}\xi_{\mathbf{i}_2}^{v_2}\dots\xi_{\mathbf{i}_s}^{v_s}| \leq C(b_{\mathbf{n}}^d)^s$$

$$\sum_{\mathbf{i}_0\neq\dots\neq\mathbf{i}_s}\mathbb{E}[\xi_{\mathbf{i}_0}^{v_0}\dots\xi_{\mathbf{i}_s}^{v_s}] \leq C\left(\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N\right)^{s+1}, \qquad \text{for} \quad s=1,2,\dots,r-1$$

$$\sum_{\mathbf{i}_0\neq\dots\neq\mathbf{i}_s}\mathbb{E}[\xi_{\mathbf{i}_0}^{v_0}\dots\xi_{\mathbf{i}_s}^{v_s}] \leq C(\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N)^r \qquad \text{for} \quad r\leq s\leq 2r-1$$

So, $\sum_{\mathbf{i}_0\neq\dots\neq\mathbf{i}_s}\mathbb{E}[\xi_{\mathbf{i}_0}^{v_0}\dots\xi_{\mathbf{i}_s}^{v_s}] \leq C(\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N)^r$. Finally, $\mathbb{E}\left(\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}\xi_{\mathbf{i}}\right)^q \leq C(\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N)^r$ and since $q=2r$, we have

$$\mathbb{E}\left(\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}\xi_{\mathbf{i}}\right)^q \leq C(\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N)^{\frac{q}{2}}.$$

■

Proof of Lemma 3.12

$$\begin{aligned}
\mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} K_{1\mathbf{i}}K_{2\mathbf{i}} \leq \frac{ua_{\mathbf{n},\mathbf{j}}}{2}\right] &= \mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} (K_{1\mathbf{i}}K_{2\mathbf{i}} - \mathbb{E}(K_{1\mathbf{i}}K_{2\mathbf{i}})) \leq \frac{ua_{\mathbf{n},\mathbf{j}}}{2} - ua_{\mathbf{n},\mathbf{j}}\right] \\
&\leq \mathbb{P}\left[\left|\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} (K_{1\mathbf{i}}K_{2\mathbf{i}} - \mathbb{E}(K_{1\mathbf{i}}K_{2\mathbf{i}}))\right| \geq \frac{ua_{\mathbf{n},\mathbf{j}}}{2}\right] \\
&\leq \mathbb{P}\left[\left|\frac{1}{a_{\mathbf{n},\mathbf{j}}b_\mathbf{n}^d}\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} (K_{1\mathbf{i}}K_{2\mathbf{i}} - \mathbb{E}(K_{1\mathbf{i}}K_{2\mathbf{i}}))\right| \geq C\right] \\
&\leq \mathbb{P}\left[|S_\mathbf{n}(x_\mathbf{j})| > \epsilon\right] \qquad \text{for } \mathbf{n} \text{ large enough,}
\end{aligned}$$

where $S_\mathbf{n}(x_\mathbf{j}) = \sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} \mathbf{\Lambda}_\mathbf{i}(x_\mathbf{j}) = f_\mathbf{n}(x_\mathbf{j}) - \mathbb{E}[f_\mathbf{n}(x_\mathbf{j})]$, $\epsilon = \epsilon_\mathbf{n}$ and p are the same as in the proof of Theorem 3.3. Considering $\mathbf{n}$ enough large and $a > 0$, we show

1. $\mathbb{P}\left[|S_\mathbf{n}(x_\mathbf{j})| > \delta\sqrt{\dfrac{\log\widehat{\mathbf{n}}}{\widehat{\mathbf{n}}b_\mathbf{n}^d\rho_\mathbf{n}^N}}\right] \leq C_N\widehat{\mathbf{n}}^{-a} + C2^{N+2}\left(\dfrac{\widehat{\mathbf{n}}b_\mathbf{n}^d\rho_\mathbf{n}^N}{\log\widehat{\mathbf{n}}}\right)^{\frac{2N-\theta}{2N}} \dfrac{\widehat{\mathbf{n}}}{a_{\mathbf{n},\mathbf{j}}b_\mathbf{n}^d}\delta^{-1}$
2. $\mathbb{P}\left[|S_\mathbf{n}(x_\mathbf{j})| > \delta\sqrt{\dfrac{\log\widehat{\mathbf{n}}}{\widehat{\mathbf{n}}b_\mathbf{n}^d\rho_\mathbf{n}^N}}\right] \leq C_N\widehat{\mathbf{n}}^{-a} + C2^{N+2}\widehat{\mathbf{n}}^{\widetilde{\beta}}\dfrac{\widehat{\mathbf{n}}}{a_{\mathbf{n},\mathbf{j}}b_\mathbf{n}^d}\delta^{-1}\left(\dfrac{\widehat{\mathbf{n}}b_\mathbf{n}^d\rho_\mathbf{n}^N}{\log\widehat{\mathbf{n}}}\right)^{\frac{N-\theta}{2N}}$.

We note that $\widehat{\mathbf{n}}^{\frac{q}{2}-a}b_\mathbf{n}^{\frac{q}{2}d}\rho_\mathbf{n}^{\frac{q}{2}N}$ tends to 0 when $a > \frac{q}{2}$ and therefore $\widehat{\mathbf{n}}^{-a} = o\left((\widehat{\mathbf{n}}b_\mathbf{n}^d\rho_\mathbf{n}^N)^{\frac{-q}{2}}\right)$. In the first case, considering assumption **H6**,

$$\begin{aligned}
(\widehat{\mathbf{n}}b_\mathbf{n}^d\rho_\mathbf{n}^N)^{\frac{q}{2}}C2^{N+2}\left(\frac{\widehat{\mathbf{n}}b_\mathbf{n}^d\rho_\mathbf{n}^N}{\log\widehat{\mathbf{n}}}\right)^{\frac{2N-\theta}{2N}} \frac{\widehat{\mathbf{n}}}{a_{\mathbf{n},\mathbf{j}}b_\mathbf{n}^d}\delta^{-1} &\leq C_N(\widehat{\mathbf{n}}b_\mathbf{n}^d\rho_\mathbf{n}^N)^{\frac{q}{2}}\left(\frac{\widehat{\mathbf{n}}b_\mathbf{n}^d\rho_\mathbf{n}^N}{\log\widehat{\mathbf{n}}}\right)^{\frac{2N-\theta}{2N}} \frac{1}{b_\mathbf{n}^d\rho_\mathbf{n}^N} \\
&\leq C_N\widehat{\mathbf{n}}^{\frac{N(q+2)-\theta}{2N}}b_\mathbf{n}^{\frac{d(qN-\theta)}{2N}}\rho_\mathbf{n}^{\frac{N(qN-\theta)}{2N}}\log\widehat{\mathbf{n}}^{\frac{\theta-2N}{2N}} \\
&\leq C_N\left(\widehat{\mathbf{n}}b_\mathbf{n}^{\frac{d(qN-\theta)}{N(q+2)-\theta}}\rho_\mathbf{n}^{\frac{N(qN-\theta)}{N(q+2)-\theta}}\log\widehat{\mathbf{n}}^{\frac{\theta-2N}{N(q+2)-\theta}}\right)^{\frac{N(q+2)-\theta}{2N}}
\end{aligned}$$

which tends to 0. In the second case, considering assumption **H7**,

$$\begin{aligned}
(\widehat{\mathbf{n}}b_\mathbf{n}^d\rho_\mathbf{n}^N)^{\frac{q}{2}}C2^{N+2}\widehat{\mathbf{n}}^{\widetilde{\beta}}\frac{\widehat{\mathbf{n}}}{a_{\mathbf{n},\mathbf{j}}b_\mathbf{n}^d}\delta^{-1}&\left(\frac{\widehat{\mathbf{n}}b_\mathbf{n}^d\rho_\mathbf{n}^N}{\log\widehat{\mathbf{n}}}\right)^{\frac{N-\theta}{2N}} \\
&\leq C_N\widehat{\mathbf{n}}^{\frac{N(q+2\tilde{\beta}+1)-\theta}{2N}}b_\mathbf{n}^{\frac{d(N(q-1)-\theta)}{2N}}\rho_\mathbf{n}^{\frac{N(N(q-1)-\theta)}{2N}}\log\widehat{\mathbf{n}}^{\frac{\theta-N}{2N}} \\
&\leq C_N\left(\widehat{\mathbf{n}}b_\mathbf{n}^{\frac{d(N(q-1)-\theta)}{N(q+2\tilde{\beta}+1)-\theta}}\rho_\mathbf{n}^{\frac{N(N(q-1)-\theta)}{N(q+2\tilde{\beta}+1)-\theta}}\log\widehat{\mathbf{n}}^{\frac{\theta-N}{N(q+2\tilde{\beta}+1)-\theta}}\right)^{\frac{N(q+2\tilde{\beta}+1)-\theta}{2N}}
\end{aligned}$$

which tends to 0. Therefore, we have

$$\begin{aligned}
\mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} K_{1\mathbf{i}}K_{2\mathbf{i}} \leq \frac{ua_{\mathbf{n},\mathbf{j}}}{2}\right] &= o\left(\left(\frac{1}{\widehat{\mathbf{n}}\rho_\mathbf{n}^N b_\mathbf{n}^d}\right)^{\frac{q}{2}}\right) \\
\left(\mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} K_{1\mathbf{i}}K_{2\mathbf{i}} \leq \frac{ua_{\mathbf{n},\mathbf{j}}}{2}\right]\right)^{1/q} &= O\left(\left(\frac{1}{\widehat{\mathbf{n}}\rho_\mathbf{n}^N b_\mathbf{n}^d}\right)^{1/2}\right).
\end{aligned}$$

■

Proof of Lemma 3.13:

Since $Y_{\mathbf{i}}$ and $r(\cdot)$ are bounded, we have

$$\begin{aligned}
\mathbb{E}^{1/q}\left[\left(\frac{1}{\widehat{\mathbf{n}}}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} - r(x_{\mathbf{j}})\right)\mathbf{1}_{\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}}=0\right]}\right]^q &\leq \mathbb{E}^{1/q}\left[\left|\frac{1}{\widehat{\mathbf{n}}}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} - r(x_{\mathbf{j}})\right|\mathbf{1}_{\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}}=0\right]}\right]^q \\
&\leq C\mathbb{E}^{1/q}\left[\mathbf{1}_{\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}}=0\right]}\right]^q \\
&\leq C\left(\mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{1\mathbf{i}}K_{2\mathbf{i}}=0\right]\right)^{1/q} \\
&\leq C\left(\mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{1\mathbf{i}}K_{2\mathbf{i}}\leq\frac{ua_{\mathbf{n},\mathbf{j}}}{2}\right]\right)^{1/q} \\
&= O\left(\left(\frac{1}{\widehat{\mathbf{n}}b_{\mathbf{n}}^d\rho_{\mathbf{n}}^N}\right)^{1/2}\right),
\end{aligned}$$

using Lemma 3.12. ■

Part II

Nonparametric statistics for functional data

Chapter 4

Spatial regression estimation for functional data with spatial dependency

Contents

Résumé en français

Dans ce chapitre, nous proposons une adaptation de l'estimateur présenté au chapitre 3 au cadre de données fonctionnelles. On se propose d'étudier un champ aléatoire strictement stationnaire $\{X_{\mathbf{i}}, Y_{\mathbf{i}}\}$ indexé par $\mathbf{i}$ dans $\mathbb{Z}^N$ dont les éléments ont la même distribution que le couple (X, Y) et qui est défini sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. Nous nous intéressons plus particulièrement à la situation où la variable explicative X prend ses valeurs dans un espace semi-métrique, noté $(\mathcal{E}, d(\cdot,\cdot))$ (de dimension éventuellement infinie) (i.e. X est une variable aléatoire fonctionnelle et d est une semi-métrique) et la variable réponse Y est à valeurs dans $\mathbb{R}$. On s'intéresse au modèle de régression défini par $Y_{\mathbf{i}} = r(X_{\mathbf{i}}) + \epsilon_{\mathbf{i}}$ où le bruit $\epsilon_{\mathbf{i}}$ est centré, α-mélangeant et indépendant de $X_{\mathbf{i}}$. L'ensemble spatial $\mathcal{I}_{\mathbf{n}} := \{\mathbf{i} = (i_1, \ldots, i_N), 1 \leq i_k \leq n_k, k = 1, \ldots, N\}$ sur lequel nous travaillons est supposé discret et appartient à $\mathbb{Z}^N$, $N \geq 1$.

Nous dénotons par $\widehat{\mathbf{n}} := n_1 \times \cdots \times n_N$ la taille de l'échantillon, par $\|\cdot\|$ une norme sur $\mathbb{Z}^N$ et par $\mathcal{B}(x, \tau)$ une boule ouverte de centre x et de rayon τ. Par souci de simplicité, on suppose que $n_1 = n_2 = \ldots = n_N = n$ (e.g. El Machkouri (2007, 2011), El Machkouri and Stoica (2010)), mais les résultats suivants peuvent être étendus à un cadre plus général.

En considérant les sites normalisés, l'estimateur à noyau proposé de $r(x_{\mathbf{j}}) = \mathbb{E}(Y_{\mathbf{j}}|X_{\mathbf{j}} =$

$x_{\mathbf{j}}$), pour $x_{\mathbf{j}} \in (\mathcal{E}, d(\cdot,\cdot))$ fixée et localisée au site $\mathbf{j} \in \mathcal{I}_{\mathbf{n}}$, est défini par

$$r_{\mathbf{n}}(x_{\mathbf{j}}) = \begin{cases} \dfrac{g_{\mathbf{n}}(x_{\mathbf{j}})}{f_{\mathbf{n}}(x_{\mathbf{j}})} & \text{si} \quad f_{\mathbf{n}}(x_{\mathbf{j}}) \neq 0; \\ \overline{Y} & \text{sinon,} \end{cases}$$

où $\overline{Y}$ est la moyenne empirique des $Y_{\mathbf{i}}$, les fonctions $g_{\mathbf{n}}(x_{\mathbf{j}})$ et $f_{\mathbf{n}}(x_{\mathbf{j}})$ sont définies par

$$g_{\mathbf{n}}(x_{\mathbf{j}}) = \frac{1}{a_{\mathbf{n},\mathbf{j}}\mathbb{E}\left[K_1\left(\frac{d(x_{\mathbf{j}},X_1)}{b_{\mathbf{n}}}\right)\right]} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} K_1\left(\frac{d(x_{\mathbf{j}},X_{\mathbf{i}})}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j}-\mathbf{i}\|)$$

$$f_{\mathbf{n}}(x_{\mathbf{j}}) = \frac{1}{a_{\mathbf{n},\mathbf{j}}\mathbb{E}\left[K_1\left(\frac{d(x_{\mathbf{j}},X_1)}{b_{\mathbf{n}}}\right)\right]} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_1\left(\frac{d(x_{\mathbf{j}},X_{\mathbf{i}})}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j}-\mathbf{i}\|)$$

avec $a_{\mathbf{n},\mathbf{j}} = \sum_{\mathbf{j}\in\mathcal{I}_{\mathbf{n}}} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j}-\mathbf{j}\|) = \sum_{\mathbf{j}\in\mathcal{I}_{\mathbf{n}}} K_2\left(\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{j}-\mathbf{i}}{\mathbf{n}}\right\|\right)$, (où $\frac{\mathbf{i}}{\mathbf{n}} = (\frac{i_1}{n}, \frac{i_2}{n}, \ldots, \frac{i_N}{n})$). On peut aussi écrire $K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j}-\mathbf{i}\|) = K_2\left(\frac{\|\mathbf{j}-\mathbf{i}\|}{n\rho_{\mathbf{n}}}\right)$. De plus, K_1 et K_2 sont des noyaux définis sur $\mathbb{R}$, $b_{\mathbf{n}}$ et $\rho_{\mathbf{n}}$ sont des fenêtres qui tendent vers zéro. Pour chaque site $\mathbf{j}$, on pose $k_{\mathbf{n}} = k_{\mathbf{n},\mathbf{j}} = \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \mathbf{1}_{[\|\mathbf{i}-\mathbf{j}\|\leq d_{\mathbf{n}}]}$ où $d_{\mathbf{n}} > 0$ est telle que $d_{\mathbf{n}} \to \infty$ lorsque $\mathbf{n} \to \infty$. On remarque que $k_{\mathbf{n},\mathbf{j}}$ est le nombre de voisins $\mathbf{i}$ pour lesquels la distance entre $\mathbf{i}$ et $\mathbf{j}$ est inférieure ou égale à la distance $d_{\mathbf{n}}$. L'estimateur $r_{\mathbf{n}}(x_{\mathbf{j}})$ est une fonction du nombre $k_{\mathbf{n}}$ pour lequel la distance $d_{\mathbf{n}}$ est choisie égale à $d_{\mathbf{n}} = n\rho_{\mathbf{n}}$ avec $k_{\mathbf{n}} \to \infty$ lorsque $\mathbf{n} \to \infty$, $k_{\mathbf{n}} = O(d_{\mathbf{n}}^N) = O(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N)$.

Plus généralement, soit $X_{\mathbf{i}}$, $\mathbf{i} \in \mathcal{I}_{\mathbf{n}}$, un processus spatial strictement stationnaire et soit $\mathbf{i_0}$ un site n'appartenant pas à $\mathcal{I}_{\mathbf{n}}$, on peut étendre les estimateurs $g_{\mathbf{n}}(\cdot)$ et $f_{\mathbf{n}}(\cdot)$ de la manière suivante

$$\widehat{g}_{\mathbf{n}}(x_{\mathbf{i_0}}) = \frac{1}{a^{\star}_{\mathbf{n},\mathbf{i_0}}\mathbb{E}\left[K_1\left(\frac{d(x_{\mathbf{i_0}},X_1)}{b_{\mathbf{n}}}\right)\right]} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} K_1\left(\frac{d(x_{\mathbf{i_0}},X_{\mathbf{i}})}{b_{\mathbf{n}}}\right) K_2\left(\frac{\mathbf{i}-\mathbf{i_0}}{\rho_{\mathbf{n}}}\right)$$

$$\widehat{f}_{\mathbf{n}}(x_{\mathbf{i_0}}) = \frac{1}{a^{\star}_{\mathbf{n},\mathbf{i_0}}\mathbb{E}\left[K_1\left(\frac{d(x_{\mathbf{i_0}},X_1)}{b_{\mathbf{n}}}\right)\right]} \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_1\left(\frac{d(x_{\mathbf{i_0}},X_{\mathbf{i}})}{b_{\mathbf{n}}}\right) K_2\left(\frac{\mathbf{i}-\mathbf{i_0}}{\rho_{\mathbf{n}}}\right)$$

où les sites $\mathbf{i}$, $\mathbf{i_0}$ ne sont pas normalisés (voir e.g. Wang et al. (2012)), $a^{\star}_{\mathbf{n},\mathbf{i_0}} = \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_2\left(\frac{\mathbf{i}-\mathbf{i_0}}{\rho_{\mathbf{n}}}\right)$ et $K_2(\cdot)$ est un noyau sur $\mathbb{R}^N$.

Dans la suite de ce chapitre, après avoir énoncé les hypothèses considérées pour ce travail, nous étudions la convergence en moyenne quadratique avec vitesse de l'estimateur à noyau proposé. Plus particulièrement, nous obtenons

$$\|r_{\mathbf{n}}(x_{\mathbf{j}}) - r(x_{\mathbf{j}})\|_2 = O\left(b_{\mathbf{n}} + \sqrt{\frac{1}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N \varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}})}}\right).$$

où $\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}}) = \mathbb{P}[X \in B(x_{\mathbf{j}}, b_{\mathbf{n}})]$, appelé probabilité de petites boules dans la littérature (e.g. Dabo-Niang (2002), Dabo-Niang (2004), Ferraty and Vieu (2006)). Ensuite, nous proposons une manière d'étendre l'estimateur de $r(\cdot)$ au cadre de la prévision spatiale. Avant de conclure, des résultats numériques issus d'une étude de simulation sont présentés. Ces résultats soulignent qu'en pratique, lorsque les données présentent une forte dépendance spatiale, notre méthode permet d'améliorer ceux obtenus avec l'estimateur à noyau classique.

Le travail présenté dans ce chapitre fait l'objet d'un article (Ternynck (2014)) publié dans le journal de la *Société Française de Statistique*.

4.1 Introduction

The spatial indexing, which provides geographical reference of data, is encountered in many subject areas such as oceanography, epidemiology, forestry survey and economy. As a consequence, the scientific research community is increasingly interested in analyzing spatial data and then in developing more and more efficient spatial statistical tools. Early spatial models appeared at the beginning of the 19^{th} century and are mainly related to parametric spatial statistics modeling (see Ripley (1981); Cressie (1993); Guyon (1995); Anselin and Florax (1995); Chilès and Delfiner (1999) for more details on statistics for spatial data). The nonparametric methods are able to reveal structure in data that might be missed by classical parametric ones. Nowadays, a dynamic concerns the deployment of nonparametric methods to spatial statistics such as density estimation, regression, prediction ...(e.g. Journel (1983); Tran (1990); Carbon et al. (1997); Biau and Cadre (2004); Hallin et al. (2009); Menezes et al. (2010)). However, most of nonparametric spatial contributions deal with univariate or multivariate data whereas recent advances of real-time measurement instruments and data storage resources led to the emergence of functional data. The studied objects can then be curves, not numbers or vectors. This kind of data is more and more frequently involved in statistical problems since the 1990's. For an introduction to this field, the reader is directed to the books of Bosq (2000); Ramsay and Silverman (2005); Ferraty and Vieu (2006).

Currently, the literature on spatial statistics for functional data is not extensive (see Laksaci and Maref (2009); Nerini et al. (2010); Delicado et al. (2010); Dabo-Niang et al. (2010); Laksaci and Mechab (2010); Dabo-Niang et al. (2011b); Attouch et al. (2011); Dabo-Niang et al. (2012a); Dabo-Niang and Yao (2013)) and is the baseline of this current work. Indeed, we are interested in estimating the nonparametric regression for functional data presenting spatial dependence. More particularly, this regression estimator aims at taking into account the spatial dependency directly in its construction. To the best of our knowledge, very little research deals with this issue. Among the nonparametric methods, the usual kernel density estimator (see Rosenblatt (1956b)) is often used in order to estimate the regression operator. In Menezes et al. (2010), a nonparametric kernel prediction is considered for spatial stochastic processes when a stochastic sampling design is assumed for selection of random locations. The particularity of this predictor is to be constructed with a kernel function on the locations. In the kernel-type estimator suggested in García-Soidán and Menezes (2012), the dependence structure is reduced to the estimation of one indicator variogram, as a nonparametric alternative to Matheron's indicator variogram. Wang et al. (2012) proposed a local linear spatio-temporal prediction model, using a kernel weight function taking into account the distance between sites. The works of Dabo-Niang et al. (2014a) and Dabo-Niang et al. (2014c) proposed, respectively, a spatial density and regression estimators, for multivariate data, depending on two kernels, one of which controls the distance between observations and the other controls the spatial dependence structure. All these previous works concern real valued data. The spatial kernel density estimator proposed in Dabo-Niang et al. (2011b) for functional data does not directly take into account the spatial dependency in the form of the estimator but the authors explained how this can be done by introducing a second kernel, based on distances between sites. Here, we combine these three last works since the regression operator is constructed from the kernel density and regression estimators introduced respectively in Dabo-Niang et al. (2014a) and Dabo-Niang et al. (2014c), when the explanatory variables are defined on $\mathbb{R}^d$, but here adapted to the functional data framework.

Denote the integer lattice points in the N-dimensional Euclidean space by $\mathbb{Z}^N$, $N \geq 1$.

Consider a strictly stationary random field $\{X_{\mathbf{i}}, Y_{\mathbf{i}}\}$ indexed by $\mathbf{i}$ in $\mathbb{Z}^N$ whose elements have the same distribution as a variable (X, Y) and defined over some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A point in bold $\mathbf{i} = (i_1, \ldots, i_N) \in \mathbb{Z}^N$ will be referred as a site. Suppose X takes values in a separable semi-metric space $(\mathcal{E}, d(\cdot,\cdot))$ (of eventually infinite dimension) (i.e. X is a functional random variable and d a semi-metric) and Y takes values in $\mathbb{R}$. We are interested in the regression model defined by $Y_{\mathbf{i}} = r(X_{\mathbf{i}}) + \epsilon_{\mathbf{i}}$ where the noise $\epsilon_{\mathbf{i}}$ is centered, α-mixing and independent of $X_{\mathbf{i}}$. Then, the main goal of this chapter is to estimate the regression function $r(\cdot)$.

In the following, we will assume, without loss of generality, that the data are observed over a rectangular region, defined by $\mathcal{I}_{\mathbf{n}} := \{\mathbf{i} = (i_1, \ldots, i_N), 1 \leq i_k \leq n_k, k = 1, \ldots, N\}$. Such regions are used in the literature to estimate nonparametrically the spatial density (Tran (1990); Biau and Cadre (2004); Wang and Wang (2009)). Let us recall that, as in any nonparametric spatial density model (see, e.g., El Machkouri (2011)), the method proposed here remains valid when the observed region has a more general form (e.g. subset of a large family of lattices of $\mathbb{R}^N$ or $\mathcal{I}_{\mathbf{n}} \subset \mathbb{R}^2$ is a closed convex domain with non-empty interior). Let $\widehat{\mathbf{n}} := n_1 \times \ldots \times n_N$ be the sample size. The letter C will be used to denote constants whose values are unimportant, $\|\cdot\|$ will denote any norm over $\mathbb{Z}^N$ and $\mathcal{B}(x, \tau)$ the opened ball of center x and radius τ. We will write $\mathbf{n} \to \infty$ if $\min\limits_{k=1,\ldots,n} n_k \to \infty$ and for all $1 \leq j, k \leq N$, for some constant $0 < C < \infty$, we assume $|n_j/n_k| < C$. This means that the number of observations on the rectangular region expands to infinity at the same rate along all directions. Such an expansion is called isotropic divergence. An other case could be considered, it is the less restrictive non-isotropic case where $\mathbf{n} \to \infty$ if $\min\limits_{k=1,\ldots,n} n_k \to \infty$. Note that the proof of the result obtained here is similar in the non-isotropic case.

Thereafter, we assume for simplicity that $n_1 = n_2 = \ldots = n_N = n$ (e.g. El Machkouri (2007, 2011), El Machkouri and Stoica (2010)), but the following results can be extended to a more general framework. For each site $\mathbf{j}$, let $k_{\mathbf{n}} = k_{\mathbf{n},\mathbf{j}} = \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} \mathbf{1}_{[\|\mathbf{i}-\mathbf{j}\| \leq d_{\mathbf{n}}]}$ where $d_{\mathbf{n}} > 0$ is such that $d_{\mathbf{n}} \to \infty$ as $\mathbf{n} \to \infty$. Note that $k_{\mathbf{n},\mathbf{j}}$ is the number of neighbors $\mathbf{i}$ for which the distance between $\mathbf{i}$ and $\mathbf{j}$ is less or equal to distance $d_{\mathbf{n}}$. Taking the Euclidean distance and if $N = 2$, we have $k_{\mathbf{n}} \leq 4d_{\mathbf{n}}^2 - 4d_{\mathbf{n}} + 4$ which leads to $k_{\mathbf{n}} = O(d_{\mathbf{n}}^2)$ and $k_{\mathbf{n}} = o(d_{\mathbf{n}}^{\eta})$, $\eta > 2$. Moreover, if $d_{\mathbf{n}} = o(\widehat{\mathbf{n}}^{\epsilon})$, $0 < \epsilon < 1$ then we have $k_{\mathbf{n}} = o(\widehat{\mathbf{n}}^{2\epsilon})$, see e.g. Kelejian and Prucha (2007).

Considering normalized sites, the proposed kernel regression estimator of r, for a fixed $x_{\mathbf{j}} \in (\mathcal{E}, d(\cdot,\cdot))$ located at a site $\mathbf{j} \in \mathcal{I}_{\mathbf{n}}$, is defined as

$$r_{\mathbf{n}}(x_{\mathbf{j}}) = \begin{cases} \dfrac{g_{\mathbf{n}}(x_{\mathbf{j}})}{f_{\mathbf{n}}(x_{\mathbf{j}})} & \text{if} \quad f_{\mathbf{n}}(x_{\mathbf{j}}) \neq 0; \\ \overline{Y} & \text{otherwise,} \end{cases}$$

where $\overline{Y}$ denotes the empirical mean of the $Y_{\mathbf{i}}$, the functions $g_{\mathbf{n}}(x_{\mathbf{j}})$ and $f_{\mathbf{n}}(x_{\mathbf{j}})$ are defined by

$$g_{\mathbf{n}}(x_{\mathbf{j}}) = \frac{1}{a_{\mathbf{n},\mathbf{j}} \mathbb{E}\left[K_1\left(\frac{d(x_{\mathbf{j}}, X_1)}{b_{\mathbf{n}}}\right)\right]} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} K_1\left(\frac{d(x_{\mathbf{j}}, X_{\mathbf{i}})}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j} - \mathbf{i}\|)$$

$$f_{\mathbf{n}}(x_{\mathbf{j}}) = \frac{1}{a_{\mathbf{n},\mathbf{j}} \mathbb{E}\left[K_1\left(\frac{d(x_{\mathbf{j}}, X_1)}{b_{\mathbf{n}}}\right)\right]} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K_1\left(\frac{d(x_{\mathbf{j}}, X_{\mathbf{i}})}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j} - \mathbf{i}\|)$$

with $a_{\mathbf{n},\mathbf{j}} = \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j} - \mathbf{i}\|) = \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K_2\left(\rho_{\mathbf{n}}^{-1} \left\|\frac{\mathbf{j}-\mathbf{i}}{\mathbf{n}}\right\|\right)$, (where $\frac{\mathbf{i}}{\mathbf{n}} = (\frac{i_1}{n}, \frac{i_2}{n}, \ldots, \frac{i_N}{n})$). It can also be written that $K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j} - \mathbf{i}\|) = K_2\left(\frac{\|\mathbf{j}-\mathbf{i}\|}{n\rho_{\mathbf{n}}}\right)$. Moreover, K_1 and K_2 are kernels

defined on $\mathbb{R}$, $b_{\mathbf{n}}$ and $\rho_{\mathbf{n}}$ are the bandwidths tending to zero. The estimator $r_{\mathbf{n}}(x_{\mathbf{j}})$ is a function of the number $k_{\mathbf{n}}$ for which distance $d_{\mathbf{n}}$ is chosen to be $d_{\mathbf{n}} = n\rho_{\mathbf{n}}$ with $k_{\mathbf{n}} \to \infty$ as $\mathbf{n} \to \infty$, $k_{\mathbf{n}} = O(d_{\mathbf{n}}^N) = O(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N)$. Hereinafter, we assume that $k_{\mathbf{n}} = C_N d_{\mathbf{n}}^N + O(d_{\mathbf{n}}^\beta)$ as $d_{\mathbf{n}} \to \infty$, $0 < \beta < N$ and C_N is a constant that depends on N. This is based on the well-known problem of counting points with lattice coordinates in the N-dimensional ball (see the first point of Remark 1 for further explanations). Similar conditions on the number of observations $\mathbf{i}$ in $\mathcal{I}_{\mathbf{n}}$ with $\left\|\frac{\mathbf{j}-\mathbf{i}}{\mathbf{n}}\right\| \leq \rho_{\mathbf{n}}$ are used in Wang and Wang (2009) who studied a local linear fitting method for real spatio-temporal data using some weights. Then, in this latter article, additional conditions concern time characteristics.

Remark 4.1.

- To give some examples where the assumption on $k_{\mathbf{n}}$ is reasonable, consider $q_{\mathbf{n}}$ the number of standard lattice (in $\mathbb{Z}^N$) points contained in a closed ball $\mathcal{B}(\mathbf{j}, d_{\mathbf{n}})$ that is $q_{\mathbf{n}} = \text{Card}\{\mathbf{i} \in \mathbb{Z}^N, \|\mathbf{i}-\mathbf{j}\| \leq d_{\mathbf{n}}\}$ where $\mathbf{j}$ is any vector of $\mathbb{R}^N$. It is well known that

$$q_{\mathbf{n}} = \frac{\pi^{N/2}}{\Gamma(N/2+1)} d_{\mathbf{n}}^N + O(d_{\mathbf{n}}^{N-1}),$$

  where $\Gamma(\cdot)$ is the Gamma function, see for instance Mitchell (1966); Chamizo and Iwaniec (1995); Tsang (2000); Meyer (2011). And notice that $k_{\mathbf{n}} = C_N q_{\mathbf{n}}$. In particular, if $N = 2$, $q_{\mathbf{n}} = \frac{\pi}{\Gamma(2)} d_{\mathbf{n}}^2 + O(d_{\mathbf{n}})$ or $q_{\mathbf{n}} = \frac{\pi}{\Gamma(2)} d_{\mathbf{n}}^2 + o(d_{\mathbf{n}}^{2/3})$.
- Instead of the previous functions $g_{\mathbf{n}}$ and $f_{\mathbf{n}}$, one can consider the simpler following versions

$$g_{\mathbf{n}}(x_{\mathbf{j}}) = \frac{1}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N \mathbb{E}\left[K_1\left(\frac{d(x_{\mathbf{j}}, X_{\mathbf{1}})}{b_{\mathbf{n}}}\right)\right]} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} K_1\left(\frac{d(x_{\mathbf{j}}, X_{\mathbf{i}})}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j}-\mathbf{i}\|),$$

$$f_{\mathbf{n}}(x_{\mathbf{j}}) = \frac{1}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N \mathbb{E}\left[K_1\left(\frac{d(x_{\mathbf{j}}, X_{\mathbf{1}})}{b_{\mathbf{n}}}\right)\right]} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K_1\left(\frac{d(x_{\mathbf{j}}, X_{\mathbf{i}})}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j}-\mathbf{i}\|).$$

  Such functions allow the following result to remain valid with some minor changes in conditions on $k_{\mathbf{n}}$.
- More generally, let $X_{\mathbf{i}}$, $\mathbf{i} \in \mathcal{I}_{\mathbf{n}}$, a strictly stationary spatial process and let $\mathbf{i_0}$ a site that does not belong to $\mathcal{I}_{\mathbf{n}}$, one can extend $g_{\mathbf{n}}(\cdot)$ and $f_{\mathbf{n}}(\cdot)$ estimates in the following way

$$\widehat{g}_{\mathbf{n}}(x_{\mathbf{i_0}}) = \frac{1}{a^\star_{\mathbf{n},\mathbf{i_0}} \mathbb{E}\left[K_1\left(\frac{d(x_{\mathbf{i_0}}, X_{\mathbf{1}})}{b_{\mathbf{n}}}\right)\right]} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} K_1\left(\frac{d(x_{\mathbf{i_0}}, X_{\mathbf{i}})}{b_{\mathbf{n}}}\right) K_2\left(\frac{\mathbf{i}-\mathbf{i_0}}{\rho_{\mathbf{n}}}\right)$$

$$\widehat{f}_{\mathbf{n}}(x_{\mathbf{i_0}}) = \frac{1}{a^\star_{\mathbf{n},\mathbf{i_0}} \mathbb{E}\left[K_1\left(\frac{d(x_{\mathbf{i_0}}, X_{\mathbf{1}})}{b_{\mathbf{n}}}\right)\right]} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K_1\left(\frac{d(x_{\mathbf{i_0}}, X_{\mathbf{i}})}{b_{\mathbf{n}}}\right) K_2\left(\frac{\mathbf{i}-\mathbf{i_0}}{\rho_{\mathbf{n}}}\right)$$

  where sites $\mathbf{i}$ and $\mathbf{i_0}$ are not normalized (see e.g. Wang et al. (2012)), $a^\star_{\mathbf{n},\mathbf{i_0}} = \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K_2\left(\frac{\mathbf{i}-\mathbf{i_0}}{\rho_{\mathbf{n}}}\right)$ and $K_2(\cdot)$ is a kernel on $\mathbb{R}^N$.

The rest of the chapter is organized as follows. In Section 4.2, we provide the assumptions, state our main result and present an example of application of $r_{\mathbf{n}}(x_{\mathbf{j}})$ to prevision. To check the performance of our estimator, numerical results are reported in Section 4.3. Conclusion is given in Section 4.4 while proofs and technical lemmas are postponed in the Appendix section 3.6.

4.2 Assumptions and main result

We first introduce some mixing assumptions. In fact, to take into account the spatial dependency, we assume that the process $Z_{\mathbf{i}} = (X_{\mathbf{i}}, Y_{\mathbf{i}})$ satisfies a mixing condition defined in Carbon et al. (1997) as follows: there exists a function $\gamma(t) \searrow 0$ as $t \to \infty$, such that

$$\begin{aligned}\alpha(\sigma(S), \sigma(S')) &= \sup\{|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|, A \in \sigma(S), B \in \sigma(S')\}, \\ &\leq \psi(\mathrm{Card}(S), \mathrm{Card}(S'))\gamma(\mathrm{dist}(S, S')),\end{aligned}$$

where $\mathrm{dist}(S, S')$ is the Euclidean distance between the two finite sets of sites S and S', $\mathrm{Card}(S)$ denotes the cardinality of the set S, $\sigma(S) = \{Z_{\mathbf{i}}, \mathbf{i} \in S\}$ and $\sigma(S') = \{Z_{\mathbf{i}}, \mathbf{i} \in S'\}$ are σ-fields generated by $Z_{\mathbf{i}}$, $\psi(\cdot)$ is a positive symmetric function nondecreasing in each variable. We recall that the process $(Z_{\mathbf{i}})$ is said to be strongly mixing if $\psi \equiv 1$. As usual, we will assume that one of both following conditions on $\gamma(i)$ is verified. These conditions are defined by

$$\gamma(i) \leq Ci^{-\theta}, \text{ for some } \theta > 0,$$

i.e. that $\gamma(i)$ tends to zero at a polynomial rate, or

$$\gamma(i) \leq C\exp(-si), \text{ for some } s > 0,$$

i.e. that $\gamma(i)$ tends to zero at an exponential rate. Concerning the function $\gamma(\cdot)$, for the sake of simplicity, we will only study the case where $\gamma(\cdot)$ tends to zero at a polynomial rate. However, similar result to that of Theorem 4.3 could be obtained with $\gamma(\cdot)$ tending to zero at an exponential rate (which implies the polynomial case). Throughout the chapter, it will be assumed that ψ satisfies either

$$\forall n, m \in \mathbb{N}, \quad \psi(n, m) \leq C\min(n, m) \qquad \text{or} \qquad \psi(n, m) \leq C(n + m + 1)^{\tilde{\beta}}$$

for some $C > 0$, and some $\tilde{\beta} \geq 1$. Such functions $\psi(n, m)$ can be found, for instance, in Tran (1990); Carbon et al. (1997); Hallin et al. (2004b); Biau and Cadre (2004); Dabo-Niang and Yao (2013).

The consistency result of $r_{\mathbf{n}}$ is obtained under the following assumptions (**H1**-**H6**) on r, the kernels, the bandwidths and local dependence condition. We will denote by p the probability distribution of the $(X_{\mathbf{i}})$'s and by $p_{\mathbf{i},\mathbf{j}}$ the joint probability distribution of $(X_{\mathbf{i}}, X_{\mathbf{j}})$, for all $\mathbf{i}$ and $\mathbf{j}$.

H1: The kernels $K_i : \mathbb{R} \to \mathbb{R}^+$, $i = 1, 2$, are of integral 1 and are such that there exist two constants C_1 and C_2 with $0 < C_1 < C_2 < \infty$, such that

$$C_1\mathbf{1}_{[0,1]}(t) \leq K_i(t) \leq C_2\mathbf{1}_{[0,1]}(t).$$

H2: $r(\cdot)$ is a Lipschitz function, that is $r \in Lip_{\mathcal{E}}$ where

$$Lip_{\mathcal{E}} = \{f : \mathcal{E} \to \mathbb{R}, \exists C \in \mathbb{R}_*^+, \forall x, x' \in \mathcal{E}, |f(x) - f(x')| < C_3 d(x, x')\}.$$

H3: **Local dependence condition** For all $\mathbf{i} \neq \mathbf{j} \in \mathbb{N}^N$, the joint probability distribution $p_{\mathbf{i},\mathbf{j}}$ of $X_{\mathbf{i}}$ and $X_{\mathbf{j}}$ satisfies

$$\exists \epsilon_1 \in (0, 1], p_{\mathbf{i},\mathbf{j}}(\mathcal{B}(x_{\mathbf{j}}, b_{\mathbf{n}}) \times \mathcal{B}(x_{\mathbf{j}}, b_{\mathbf{n}})) \leq C_4(\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}}))^{1+\epsilon_1},$$

where $\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}}) = \mathbb{P}[X \in \mathcal{B}(x_{\mathbf{j}}, b_{\mathbf{n}})]$, called small ball probability in the literature (e.g. Ferraty and Vieu (2006)).

H4: $\forall n, m \in \mathbb{N}$, $\psi(n,m) \leq C\min(n,m)$ and $\widehat{\mathbf{n}}\varphi_{x_\mathbf{j}}(b_\mathbf{n})^{\theta_1}\rho_\mathbf{n}^{N\theta_1}\log\widehat{\mathbf{n}}^{-\theta_1} \to \infty$ with the mixing coefficient $\theta > 4N$ and with $\theta_1 = \dfrac{2N-\theta}{4N-\theta}$.

H5: $\forall n, m \in \mathbb{N}$, for some $\tilde{\beta} \geq 1$, $\psi(n,m) \leq C(n+m+1)^{\tilde{\beta}}$ and $\widehat{\mathbf{n}}\varphi_{x_\mathbf{j}}(b_\mathbf{n})^{\theta_1^*}\rho_\mathbf{n}^{N\theta_1^*}\log\widehat{\mathbf{n}}^{-\theta_1^*} \to \infty$ with the mixing coefficient $\theta > N(3+2\tilde{\beta})$ and with $\theta_1^* = \dfrac{N-\theta}{N(3+2\tilde{\beta})-\theta}$.

H6: The variable Y is bounded almost surely and $|Y| < M$.

Remark 4.2.

- Assumptions **H1** and **H2** allow to control the bias of the estimator.
  - ☐ Assumption **H1** concerns the kernels K_i, $i = 1, 2$. More general kernels such as Gaussian or Silverman can also be used but for simplicity of calculations, we consider such kernels usually considered in nonparametric regression. For example, this condition is verified by e.g. triangular (Bartlett), biweight, circular (cosine), Epanechnikov, Parzen, Tukey-Hanning kernels.
  - ☐ A nonparametric assumption on the regression function is considered through hypothesis **H2**. In fact, this Lipschitz condition allows the precise rate of convergence to be found whereas a continuity-type model would give only convergence results. Assuming the continuity condition, one can obtain that

$$r_\mathbf{n}(x_\mathbf{j}) - r(x_\mathbf{j}) \xrightarrow{m.s.} 0 \text{ with } \widehat{\mathbf{n}}\rho_\mathbf{n}^N\varphi_{x_\mathbf{j}}(b_\mathbf{n}) \to \infty,\ b_\mathbf{n} \to 0 \text{ and } \rho_\mathbf{n} \to 0$$

    where $\xrightarrow{m.s.}$ denotes the mean square convergence.
- Assumption **H3** concerns the local dependency and a consequence is

$$\begin{aligned}|p_{\mathbf{i},\mathbf{j}}(\mathcal{B}(x_\mathbf{j}, b_\mathbf{n}) \times \mathcal{B}(x_\mathbf{j}, b_\mathbf{n})) - (\varphi_{x_\mathbf{j}}(b_\mathbf{n}))^2| &\leq |C_4(\varphi_{x_\mathbf{j}}(b_\mathbf{n}))^{1+\epsilon_1} - (\varphi_{x_\mathbf{j}}(b_\mathbf{n}))^2| \\ &\leq C_4(\varphi_{x_\mathbf{j}}(b_\mathbf{n}))^{1+\epsilon_1} \leq 1.\end{aligned}$$

  As it is noticed in Dabo-Niang et al. (2011b), this result can be linked with the classical local dependence condition met in the literature of real valued data when X and $(X_\mathbf{i}, X_\mathbf{j})$ admit, respectively, the densities f and $f_{\mathbf{i},\mathbf{j}}$. Such assumption can be also found in Ferraty and Vieu (2006) (Chapter 11, page 163) and in Dabo-Niang and Yao (2013).
- Assumptions **H4** and **H5** concern the mixing dependency and are similar to those of Carbon et al. (1997).

The following theorem states the pointwise mean square convergence of the proposed regression function estimator, whose proof is given in Appendix. We will denote $\|r_\mathbf{n}(x_\mathbf{j}) - r(x_\mathbf{j})\|_2 = \left(\mathbb{E}\left[(r_\mathbf{n}(x_\mathbf{j}) - r(x_\mathbf{j}))^2\right]\right)^{1/2}$.

Theorem 4.3. Under assumptions **H1-H3**, **H4** or **H5** and **H6**, we have

$$\|r_\mathbf{n}(x_\mathbf{j}) - r(x_\mathbf{j})\|_2 = O\left(b_\mathbf{n} + \sqrt{\frac{1}{\widehat{\mathbf{n}}\rho_\mathbf{n}^N\varphi_{x_\mathbf{j}}(b_\mathbf{n})}}\right).$$

Precisely, we have

$$\|r_\mathbf{n}(x_\mathbf{j}) - r(x_\mathbf{j})\|_2 = C_3 \times b_\mathbf{n} + \left(2C(2MC_2 + 2M\sqrt{C_4} + C_0) + 4M\right) \times \sqrt{\frac{1}{\widehat{\mathbf{n}}\rho_\mathbf{n}^N\varphi_{x_\mathbf{j}}(b_\mathbf{n})}},$$

where C depends on N (see e.g. Chamizo and Iwaniec (1995)) and is such that $k_\mathbf{n} \leq C\widehat{\mathbf{n}}\rho_\mathbf{n}^N$ whereas C_0 is a constant depending on the constant appearing in Lemma A.1.

This result permits to have a bound of the mean squared error of $r_{\mathbf{n}}(\cdot)$ that depends on $\rho_{\mathbf{n}}$. This is linked with the fact that $r_{\mathbf{n}}(\cdot)$ incorporates the dependence between sites compared to the result of Dabo-Niang et al. (2011b). In the foregoing paper, the authors used the following functional regression estimator

$$r^{\star}_{\mathbf{n}}(x) = \frac{\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} K_1\left(\frac{d(X_{\mathbf{i}},x)}{b^{\star}_{\mathbf{n}}}\right)}{\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_1\left(\frac{d(X_{\mathbf{i}},x)}{b^{\star}_{\mathbf{n}}}\right)}$$

where K_1 is a kernel and $b^{\star}_{\mathbf{n}}$ is the corresponding bandwidth. They gave an uniform almost sure bound of $|r^{\star}_{\mathbf{n}}(x) - r(x)|$ on a specific set $\mathcal{C}$, that is $O\left(b^{\star}_{\mathbf{n}} + \sqrt{\frac{\log \widehat{\mathbf{n}}}{\Gamma(b^{\star}_{\mathbf{n}})\widehat{\mathbf{n}}}}\right)$ with $\Gamma(b^{\star}_{\mathbf{n}}) = \sup_{x\in\mathcal{C}} \varphi_x(b_{\mathbf{n}})^{\star}$. For multivariate data, Dabo-Niang et al. (2014a) focus on a rate of almost complete convergence of the density estimator $f_{\mathbf{n}}(v_{\mathbf{i}})$ of $f(v_{\mathbf{i}})$, for $\mathbb{R}^d$-valued spatial data $v_{\mathbf{i}}$, depending on two kernels. They obtained that

$$|f_{\mathbf{n}}(v_{\mathbf{i}}) - f(v_{\mathbf{i}})| = O\left(a_{\mathbf{n}} + \sqrt{\frac{\log \widehat{\mathbf{n}}}{\widehat{\mathbf{n}}\tau^N_{\mathbf{n}} a^d_{\mathbf{n}}}}\right), \qquad \text{a. c.}$$

where $a_{\mathbf{n}}$ and $\tau_{\mathbf{n}}$ are the bandwidths corresponding to the kernels on the observations and on the sites respectively. This work is extended to regression estimation for multivariate data by Dabo-Niang et al. (2014c).

Remark 4.4.

- This current work is supported by a particular sampling scheme, which only includes deterministic designs for the spatial locations. For this reason, the bound of Theorem 4.3 shows a dissymmetric contribution of $b_{\mathbf{n}}$ and $\rho_{\mathbf{n}}$ on the risk even though both kernels K_1 and K_2 play symmetric roles. One can generalize this work to random spatial sample such as in Menezes et al. (2010) (for real-valued regression) and in Kelejian and Prucha (2007) (for spatial HAC estimation) and have a bound including $\rho^{\alpha}_{\mathbf{n}}$.
- Theorem 4.3 deals with local convergence (for a fixed $x_{\mathbf{j}}$) of the regression estimate but one can extend the obtained result to uniform one, on a set where corresponding sites are in a set S (that can be a subset of $\mathcal{I}_{\mathbf{n}}$ or a set larger than $\mathcal{I}_{\mathbf{n}}$) by considering $l_{\mathbf{n}} = \sup_{\mathbf{j}\in S} k_{\mathbf{n},\mathbf{j}}$.

The remainder of this section focuses on the application of the proposed regression function through an example, namely the spatial prediction.

Application to spatial prediction

In spatial statistic, an important topic, encountered in the literature, concerns the spatial prediction. One of the most popular method is kriging, which was developed at the beginning of the 1950's and studied in the scope of geostatistics. More recently, some works proposed nonparametric predictors for spatial fields indexed by lattices. The first results in this direction are those of Biau and Cadre (2004) and concerned the kernel prediction of a strictly stationary random field indexed in $(\mathbb{N}^*)^N$. Later, Dabo-Niang and Yao (2007) contribute to Biau and Cadre (2004)'s investigations since they are interested in the kernel regression estimation and prediction of continuously indexed random fields. In Menezes et al. (2010), nonparametric kernel prediction is considered for spatial stochastic processes when a stochastic sampling design is assumed for selection of random locations. These

contributions, but also Dabo-Niang et al. (2014c), dealt with multivariate data. In Dabo-Niang et al. (2011b), the authors stated that their work, dealing with the spatial regression estimator for functional data, offers some interesting perspectives of investigation, namely in spatial forecasting and real data problem. In continuation of these works, we propose a spatial prediction methodology dealing with functional data, taking explicitly into account the spatial locations.

In this application, we consider a $\mathbb{R}$-valued strictly stationary random spatial process $(\eta_{\mathbf{i}}, \mathbf{i} \in (\mathbb{R}^*)^N)$. This process is assumed to be bounded and observed over a subset $\mathcal{O}_{\mathbf{n}} \subset \mathcal{I}_{\mathbf{n}}$. We are interested in predicting $\eta_{\mathbf{i_0}}$ at an unobserved given location $\mathbf{i_0} \in \mathcal{I}_{\mathbf{n}} \setminus \mathcal{O}_{\mathbf{n}}$.

In practice, we expect that $\eta_{\mathbf{i_0}}$ depends only on the values of the process in a bounded neighborhood $\mathcal{V}_{\mathbf{i_0}} = \mathbf{i_0} + \mathcal{V} \subset \mathcal{O}_{\mathbf{n}}$, where $\mathbf{0} = (0, 0, \ldots, 0) \notin \mathcal{V}$. Consequently, we can construct a function $\tilde{\eta}_{\mathbf{i_0}}$ from the observations in a continuous vicinity $\mathcal{V}_{\mathbf{i_0}}$ of $\mathbf{i_0}$ and define $\tilde{\eta}_{\mathbf{i}} = \{\eta_{\mathbf{j}}, \mathbf{j} \in \mathcal{V}_{\mathbf{i}} = \mathbf{i} + \mathcal{V} \subset \mathbb{R}^N\}$ which belongs to the space of continuous and bounded functions. For more details on the choice of $\mathcal{V}$, see Dabo-Niang and Yao (2007).

To achieve the forecasting at the site $\mathbf{i_0} \notin \mathcal{O}_{\mathbf{n}}$, we propose to use the regression function estimator $r_{\mathbf{n}}$ suggested previously. Then, the value to be predicted of the field $(\eta_{\mathbf{i}})_{\mathbf{i} \in (\mathbb{R}^*)^N}$ at a site $\mathbf{i_0}$ becomes

$$\widehat{\eta}_{\mathbf{i_0}} = r_{\mathbf{n}}(\tilde{\eta}_{\mathbf{i_0}}) = \frac{\sum_{\mathbf{i} \in \mathcal{O}_{\mathbf{n}}} \eta_{\mathbf{i}} K_1\left(\frac{d(\tilde{\eta}_{\mathbf{i_0}}, \tilde{\eta}_{\mathbf{i}})}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i_0} - \mathbf{i}\|)}{\sum_{\mathbf{i} \in \mathcal{O}_{\mathbf{n}}} K_1\left(\frac{d(\tilde{\eta}_{\mathbf{i_0}}, \tilde{\eta}_{\mathbf{i}})}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i_0} - \mathbf{i}\|)}.$$

One can derive an asymptotic result such as mean square convergence of $\widehat{\eta}_{\mathbf{i_0}}$ by considering a kernel regression estimate of functional spatial random variables continuously indexed. Having checked the theoretical behavior of our estimator and presented a potential application, we are going to study its practical behavior through some numerical results.

4.3 Numerical results

In this section, we study the performance of the proposed regression estimator through some simulations which point out the importance of taking into account the spatial locations of the data. We remind that our theoretical result is obtained under a mixing condition which can be considered by the kernel function on the locations. We compare our estimator with the one that ignores any spatial dependence in the structure of the regression estimate (see Dabo-Niang et al. (2011b)). We consider a sample of dependent realizations of some spatial functional variables $X_{\mathbf{i}}$ with the same distribution as a random field X valued in some infinite dimensional semi-metric space $(\mathcal{E}, d(\cdot,\cdot))$. That is, on each site $\mathbf{i}$, we have a curve $X_{\mathbf{i}}$ such that $X_{\mathbf{i}} = \{X_{\mathbf{i}}(t), t \in [0, T]\}$. Before studying the numerical results, we propose a useful procedure for estimating the spatial regression function.

4.3.1 Procedure of estimation of $r(X_{\mathbf{j}})$, $\mathbf{j} \in \mathcal{I}_{\mathbf{n}}$

1. Specify sets of bandwidths $S(b)$ and $S(\rho)$ of respectively K_1 and K_2.
2. For each $b_{\mathbf{n}} \in S(b)$ and $\rho_{\mathbf{n}} \in S(\rho)$ and each $\mathbf{j} \in \mathcal{I}_{\mathbf{n}}$, compute

$$r_{\mathbf{n}}(X_{\mathbf{j}}) = \frac{\sum_{\substack{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}, \\ \mathbf{i} \neq \mathbf{j}}} Y_{\mathbf{i}} K_1\left(\frac{d(X_{\mathbf{i}}, X_{\mathbf{j}})}{b_{\mathbf{n}}}\right) K_2\left(\rho_{\mathbf{n}}^{-1} \left\|\frac{\mathbf{i}-\mathbf{j}}{\mathbf{n}}\right\|\right)}{\sum_{\substack{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}, \\ \mathbf{i} \neq \mathbf{j}}} K_1\left(\frac{d(X_{\mathbf{i}}, X_{\mathbf{j}})}{b_{\mathbf{n}}}\right) K_2\left(\rho_{\mathbf{n}}^{-1} \left\|\frac{\mathbf{i}-\mathbf{j}}{\mathbf{n}}\right\|\right)}$$

3. Compute $b_{\mathbf{n},opt}$ and $\rho_{\mathbf{n},opt}$ by applying a cross-validation procedure over $S(b)$ and $S(\rho)$. More precisely, consider the following minimization problem, i.e. determine $b_{\mathbf{n},opt}$ and $\rho_{\mathbf{n},opt}$ which minimize the mean squared error over the $\widehat{\mathbf{n}}$ sites

$$\min_{b_{\mathbf{n}},\rho_{\mathbf{n}}} \quad \frac{1}{\widehat{\mathbf{n}}} \sum_{\mathbf{j}\in\mathcal{I}_{\mathbf{n}}} (r_{\mathbf{n}}(X_{\mathbf{j}}) - Y_{\mathbf{j}})^2$$

4. For each $\mathbf{j}$, compute $r_{\mathbf{n},opt}(X_{\mathbf{j}})$ corresponding to $b_{\mathbf{n},opt}$ and $\rho_{\mathbf{n},opt}$.

4.3.2 Simulation

This last procedure is used in the following simulation study dealing with $N = 2$. We consider observations $(X_{(i,j)}, Y_{(i,j)})$, $1 \leq i, j \leq 25$, such that

$$\begin{aligned} Y_{(i,j)} &= r(X_{(i,j)}) + \epsilon_{(i,j)} \\ &= 4 \times A^2_{(i,j)} + \epsilon_{(i,j)} \end{aligned}$$

and for $t \in [0,1]$, $X_{(i,j)}(t)$ is defined according to the following cases

Case 1: $X_{(i,j)}(t) = A^2_{(i,j)} \times (t - 0.5)^2 + A_{(i,j)} \times B_{(i,j)}$;

Case 2: $X_{(i,j)}(t) = A_{(i,j)} \times \cos(2\pi t)$,

where $A = (A_{(i,j)})$, $B = (B_{(i,j)})$ and $\epsilon = (\epsilon_{(i,j)})$ are random variables which will be specified according to the following considered model on $A = (A_{(i,j)})$. Several curve examples of $X_{(i,j)}(t)$, for each case, are drawn on Figure 4.1. More precisely, the figure on the left displays some curves simulated from Case 1 and that on the right concerns Case 2. In Case 1, an example of the function $r(\cdot)$ could be $r(X) = 2X''$ (where X'' denotes the second derivative of X with respect to t) whereas in Case 2, it could be $r(X) = \frac{A}{\pi^2} X''$ with $t = \frac{1}{2}$. We will denote by $GRF(m, \sigma^2, s)$ any stationary Gaussian Random Field with mean m and spatial covariance function defined by

$$C(h) = \sigma^2 \exp\left(-\left(\frac{\|h\|}{s}\right)^2\right), \qquad h \in \mathbb{R}^2 \text{ and } s > 0.$$

Then, we define the two considered models on $A = (A_{(i,j)})$ by

Model A: $A_{i,j} = D_{i,j} \times (\sin(2G_{i,j}) + 2\exp(-16G^2_{i,j}))$;

Model B: $A_{i,j} = D_{i,j} \times (2\cos(2G_{i,j}) + \exp(-4G^2_{i,j}))$.

Here, the number of observations $\widehat{\mathbf{n}}$ is equal to 25×25, i.e. 625. The several fields are defined by $D_{i,j} = \frac{1}{625} \sum_{1\leq m,t\leq 25} \exp\left(-\frac{\|(i,j)-(m,t)\|}{a}\right)$, $G_{i,j} = GRF(0, 5, 3)$, $B_{i,j} = GRF(2.5, 5, 3)$ and $\epsilon_{i,j} = GRF(0, 0.1, 5)$. We note that the spatial dependence is controlled by the value of a. In fact, the greater a is, the weaker the spatial dependency is. According to this fact, we provide simulation results obtained with different values of a which are $a = 5$, 20 and 50.

Along this part, the spatial regression is computed based on the kernels K_1 as the Epanechnikov kernel and K_2 as the Parzen kernel. The choice of the semi-metric $d(\cdot,\cdot)$ is important and depends on the information one gets on the data. Ferraty and Vieu (2006) present three families of semi-metrics. The first is built from functional principal component analysis (FPCA) and is adapted to rough curves. The second is built from the partial least square (PLS) approach and is relevant when one consider multivariate response. The last, based on derivatives, is well adapted in the presence of smooth curves.

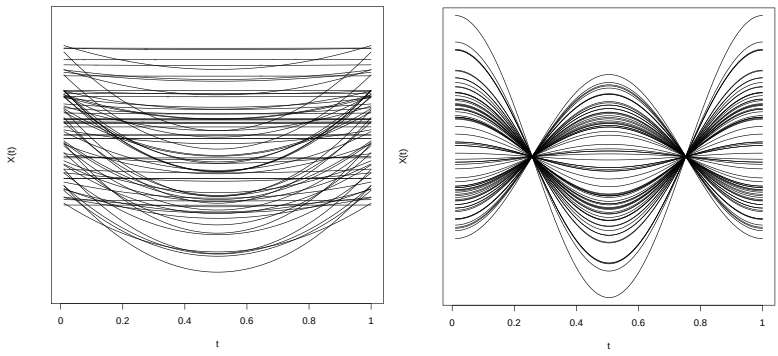


Figure 4.1 – Some simulated curves of Case 1 (left) and Case 2 (right)

Specifically, it approximates L_2 metric between derivatives of the curves based on their $B-$spline representation. Note that other semi-metrics are encountered in the literature. However, according to Delsol (2008), the theoretical justification of the usefulness of a particular semi-metric is still an open problem. In this work, we consider a semi-metric between curves based on their first $q = 2$ derivatives because of the smoothness of the curves. This semi-metric (between X_i and X_j) is defined by

$$\sqrt{\int \left(X_i^{(q)}(t) - X_j^{(q)}(t)\right)^2 dt}, \qquad q = 0, 1, 2, \ldots$$

where, for any q-times differentiable real function X, $X^{(q)}$ denotes the qth derivative of X (we refer, for example, to Ferraty and Vieu (2006) for the theoretical setting about semi-metrics used for functional nonparametric investigations). To confirm our semi-metric choice, we tested, in addition to the semi-metrics based on their first derivatives, two other semi-metrics (based on PCA and on Fourier's decomposition) and different parameters such as the number of derivatives, principal components, basis, etc. It turns out that the results are similar or worse than those obtained considering a semi-metric between curves based on their first $q = 2$ derivatives.

Recall that, in the work of Dabo-Niang et al. (2011b), a theoretical estimator of the spatial regression function for functional data is proposed. This estimator does not directly take into account the spatial locations. However, in the application section, the authors explained how this can be done using the k-nearest neighbors method. Then, in the simulation study, they proposed a procedure of estimation based on nearest neighbors. This combination looks like to the estimator $r_{\mathbf{n}}(x_{\mathbf{j}})$ introduced in this chapter. The difference is that in Dabo-Niang et al. (2011b) the k-nearest neighbors of a point $\mathbf{j}$ are considered in the pointwise regression estimation whereas, with our methodology, all the points in the ball of radius $\rho_{\mathbf{n},opt}$ and center $\mathbf{j}$ are considered.

To evaluate the performance of the proposed regression estimator, now denoted by $r_{\mathbf{n}}^{\sharp}(\cdot)$ and to compare it with the one that does not directly take into account the distance between locations and denoted $r_{\mathbf{n}}^{\star}(\cdot)$ (the theoretical estimator introduced in Dabo-Niang et al. (2011b)), each studied model is replicated 30 times. Recall that $r_{\mathbf{n}}^{\sharp}(\cdot)$ and $r_{\mathbf{n}}^{\star}(\cdot)$ are defined by

$$r_{\mathbf{n}}^{\sharp}(X_{\mathbf{j}}) = \frac{\sum_{\substack{\mathbf{i}\in\mathcal{I}_{\mathbf{n}},\\ \mathbf{i}\neq\mathbf{j}}} Y_{\mathbf{i}} K_1\left(\frac{d(X_{\mathbf{i}},X_{\mathbf{j}})}{b_{\mathbf{n}}^{\sharp}}\right) K_2\left(\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{i}-\mathbf{j}}{\mathbf{n}}\right\|\right)}{\sum_{\substack{\mathbf{i}\in\mathcal{I}_{\mathbf{n}},\\ \mathbf{i}\neq\mathbf{j}}} K_1\left(\frac{d(X_{\mathbf{i}},X_{\mathbf{j}})}{b_{\mathbf{n}}^{\sharp}}\right) K_2\left(\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{i}-\mathbf{j}}{\mathbf{n}}\right\|\right)}$$

and

$$r^{\star}_{\mathbf{n}}(X_{\mathbf{j}}) = \frac{\sum_{\substack{\mathbf{i}\in\mathcal{I}_{\mathbf{n}},\\ \mathbf{i}\neq\mathbf{j}}} Y_{\mathbf{i}} K_1\left(\frac{d(X_{\mathbf{i}},X_{\mathbf{j}})}{b^{\star}_{\mathbf{n}}}\right)}{\sum_{\substack{\mathbf{i}\in\mathcal{I}_{\mathbf{n}},\\ \mathbf{i}\neq\mathbf{j}}} K_1\left(\frac{d(X_{\mathbf{i}},X_{\mathbf{j}})}{b^{\star}_{\mathbf{n}}}\right)}.$$

Note that if $\sum_{\substack{\mathbf{i}\in\mathcal{I}_{\mathbf{n}},\\ \mathbf{i}\neq\mathbf{j}}} K_1\left(\frac{d(X_{\mathbf{i}},X_{\mathbf{j}})}{b^{\sharp}_{\mathbf{n}}}\right) K_2\left(\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{i}-\mathbf{j}}{\mathbf{n}}\right\|\right) = 0$ then $r^{\sharp}_{\mathbf{n}}(X_{\mathbf{j}}) = \frac{1}{\widehat{\mathbf{n}}-1}\sum_{\substack{\mathbf{i}\in\mathcal{I}_{\mathbf{n}},\\ \mathbf{i}\neq\mathbf{j}}} Y_{\mathbf{i}}$. In the same way, if $\sum_{\substack{\mathbf{i}\in\mathcal{I}_{\mathbf{n}},\\ \mathbf{i}\neq\mathbf{j}}} K_1\left(\frac{d(X_{\mathbf{i}},X_{\mathbf{j}})}{b^{\star}_{\mathbf{n}}}\right) = 0$ then $r^{\star}_{\mathbf{n}}(X_{\mathbf{j}}) = \frac{1}{\widehat{\mathbf{n}}-1}\sum_{\substack{\mathbf{i}\in\mathcal{I}_{\mathbf{n}},\\ \mathbf{i}\neq\mathbf{j}}} Y_{\mathbf{i}}$.
At each replication k, we compute the mean squared error over the $\widehat{\mathbf{n}}$ sites. The bandwidths used, different at each replication, are those obtained using the previous procedure 4.3.1. For the k^{th} replication, we define the mean squared error ($MSE^{(k)}$) by

$$MSE^{(k)} = \frac{1}{\widehat{\mathbf{n}}}\sum_{\mathbf{j}\in\mathcal{I}_{\mathbf{n}}} (r^{\dagger}_{\mathbf{n},opt}(X_{\mathbf{j}}) - Y_{\mathbf{j}}))^2, \text{ with } r^{\dagger}_{\mathbf{n}} = r^{\sharp}_{\mathbf{n}} \text{ or } r^{\star}_{\mathbf{n}}. \tag{4.1}$$

The obtained results are summarized in Table 4.1. For each situation (Model, Case and value of a), this table provides the average MSE over the 30 replications of Equation (4.1) and the corresponding standard deviation. The AMSE$^{\sharp}$'s (average mean squared error) column makes reference to the proposed estimator $r^{\sharp}_{\mathbf{n}}$ whereas the AMSE*'s column corresponds to the estimator $r^{\star}_{\mathbf{n}}$ which takes no account of location. Besides, we use a statistical hypothesis test rather than directly compare the average MSE accuracy. The column entitled "p-value" gives, for each considered situation, the p-value of a paired t-test performing in order to determine if MSE$^{\sharp}$ is significantly less than MSE* (the alternative hypothesis is then H_1: AMSE$^{\sharp}$ < AMSE*). The two last columns give the average of the coefficients of determination over the 30 replications. Recall that a value of R^2 close to 1 means that the quality of estimation is reliable.

Model	Case	a	$AMSE^{\sharp}$	$AMSE^{\star}$	p-value	$AR^{2\sharp}$	$AR^{2\star}$
A	1	5	0.034 (0.014)	0.095 (0.030)	3.92×10^{-14}	0.652	0.057
		20	0.041 (0.013)	0.097 (0.024)	3.30×10^{-17}	0.956	0.896
		50	0.060 (0.014)	0.100 (0.022)	3.66×10^{-13}	0.981	0.969
	2	5	0.007 (0.003)	0.093 (0.030)	3.94×10^{-16}	0.925	0.054
		20	0.036 (0.006)	0.097 (0.031)	6.84×10^{-13}	0.960	0.895
		50	0.058 (0.011)	0.100 (0.031)	1.12×10^{-09}	0.982	0.970
B	1	5	0.012 (0.004)	0.092 (0.029)	6.86×10^{-16}	0.914	0.361
		20	0.049 (0.008)	0.100 (0.029)	3.52×10^{-12}	0.994	0.988
		50	0.071 (0.014)	0.100 (0.025)	1.56×10^{-10}	0.998	0.997
	2	5	0.010 (0.001)	0.093 (0.030)	7.65×10^{-16}	0.926	0.356
		20	0.060 (0.010)	0.100 (0.031)	4.58×10^{-10}	0.993	0.988
		50	0.086 (0.017)	0.108 (0.031)	7.23×10^{-07}	0.997	0.996

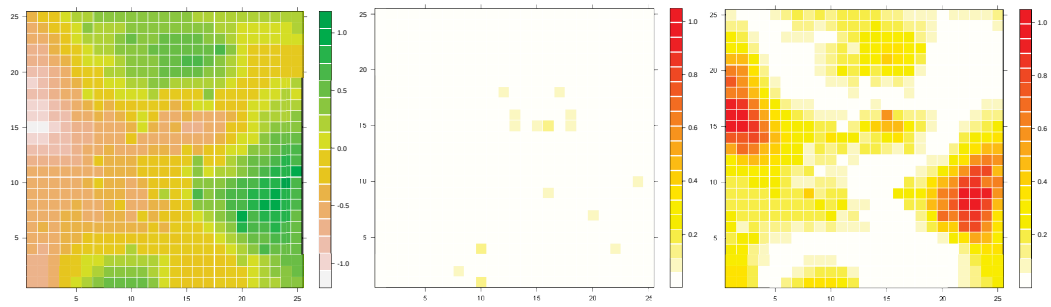
Table 4.1 – Simulation results according to the models A and B, the cases 1 and 2 and the value of $a = 5$, 20 and 50

The table gives the average mean squared errors (AMSE) for each situation and in brackets the corresponding standard deviation. The column entitled "p-value" gives the p-value of a paired t-test performing in order to determine whether $AMSE^{\sharp}$ is significantly less than $AMSE^{\star}$. The two last columns display the average coefficients of determination (AR^2).

The first general point to make about this study is that, when $a = 5$, regardless the considered kind of model or case, the estimator $r^{\sharp}_{\mathbf{n}}$ leads to better results since the mean squared errors are significantly lower than with $r^{\star}_{\mathbf{n}}$. For instance, for Model A and Case 2, the average of the mean squared errors is 0.007 using the estimator $r^{\sharp}_{\mathbf{n}}$ and 0.093 with $r^{\star}_{\mathbf{n}}$. Moreover, it can be seen that the standard deviations are greater with $r^{\star}_{\mathbf{n}}$ than

with $r_{\mathbf{n}}^{\sharp}$. Secondly, we note that when the value of a increases, $AMSE^{\sharp}$ is still higher than $AMSE^{\star}$ but the difference becomes narrower. Consequently, the higher the value of a (less spatial dependency), the lower the difference between the results of the two estimators is. In other words, our estimator $r_{\mathbf{n}}^{\sharp}$ outperforms $r_{\mathbf{n}}^{\star}$ when the spatial dependence is important. However, the two estimators tend to give similar performance in case of spatially independent fields. The low p-values (less than 7.23×10^{-07}) confirm that $r_{\mathbf{n}}^{\sharp}$ produces less errors than $r_{\mathbf{n}}^{\star}$. Nevertheless, the probability of erroneously rejecting the null hypothesis is highest when the value of a is equal to 20 or 50 rather than 5 (without one exception) since the p-value increases with the value of a. Finally, we may note that the R^2 criterion strengthens the previous comments. In fact, the values $AR^{2\sharp}$ are higher than $AR^{2\star}$ and the difference between them decreases as the value of a increases.

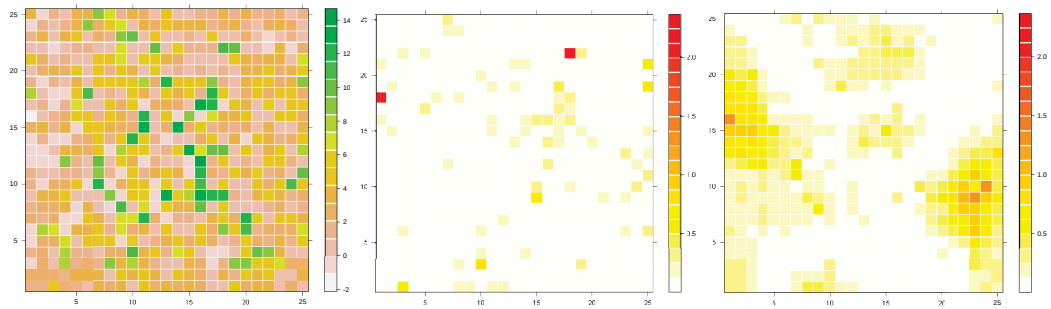
Insight into the performance of the two regression estimators can also be viewed from graphical outputs. In fact, Figures 4.2, 4.3 and 4.4 illustrate different situations. The first deals with spatially dependent data ($a = 5$) simulated from Model A and Case 2 of which a representation of Y is depicted in Figure 4.2a. Figures 4.2b and 4.2c show squared errors (more precisely, at each site $\mathbf{j}$, $[r_{\mathbf{n}}(X_{\mathbf{j}}) - Y_{\mathbf{j}}]^2$ is represented) obtained using functions $r_{\mathbf{n}}^{\sharp}$ and $r_{\mathbf{n}}^{\star}$, respectively. These two figures confirm that our methodology generates less errors than using the regression function $r_{\mathbf{n}}^{\star}$ since the more colorful the representation is, the greater the error is. Figure 4.3 considers lower spatial dependence ($a = 20$) simulated from Model B and Case 1 for which the field Y is represented in Figure 4.3a. Figure 4.3b displays slightly less errors than in Figure 4.3c. Finally, Figure 4.4 gives summarized results of Model B and Case 2, with almost independent spatial data ($a = 50$). The two estimators seem to provide similar errors according to Figures 4.4b and 4.4c. It is not surprising to note that when a is high the two estimators produce similar results. In fact, in this situation, the bandwidths $\rho_{\mathbf{n}}$ are large and could take the maximal distance between observations. In short, the two estimators work in an almost identical manner in lack of spatial dependence.



(a) (b) (c)

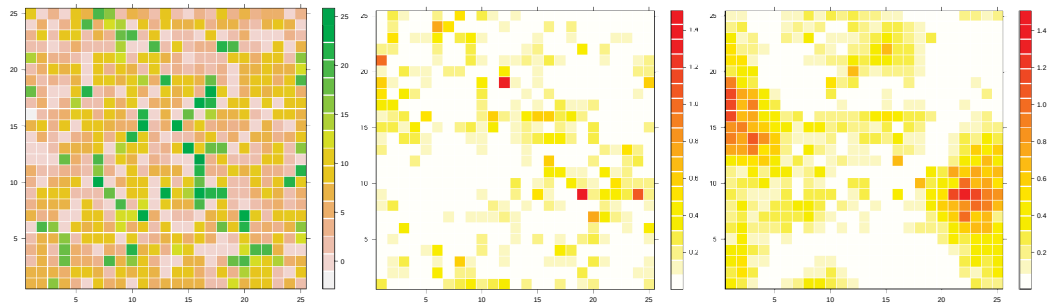
Figure 4.2 – Illustration of the results in the first considered situation
A simulated field considering Model A, Case 2 and $a = 5$ with (a) an image of the field Y; (b) the squared errors using $r_{\mathbf{n}}^{\sharp}$; (c) the squared errors using $r_{\mathbf{n}}^{\star}$

Regarding the bandwidths selection, we carried out a cross-validation procedure. This selection is made differently, according to the situation, $r_{\mathbf{n}}^{\sharp}$ and $r_{\mathbf{n}}^{\star}$. Firstly, with spatially dependent data ($a = 5$) the selected bandwidths $\rho_{\mathbf{n}}$ have the smallest values. This result was expected because when the spatial dependence is high, sites that are close together tend to be more related than sites that are far apart. From Model A and Case 2, the bandwidths $\rho_{\mathbf{n},opt}$, dealing with the kernel on the locations, are between 0.126 and 0.322. For the bandwidth linked to the distance between the observations (according to K_1),



(a) (b) (c)

Figure 4.3 – Illustration of the results in the second considered situation
A simulated field considering Model B, Case 1 and $a = 20$ with (a) an image of the field Y; (b) the squared errors using $r^{\sharp}_{\mathbf{n}}$; (c) the squared errors using $r^{\star}_{\mathbf{n}}$

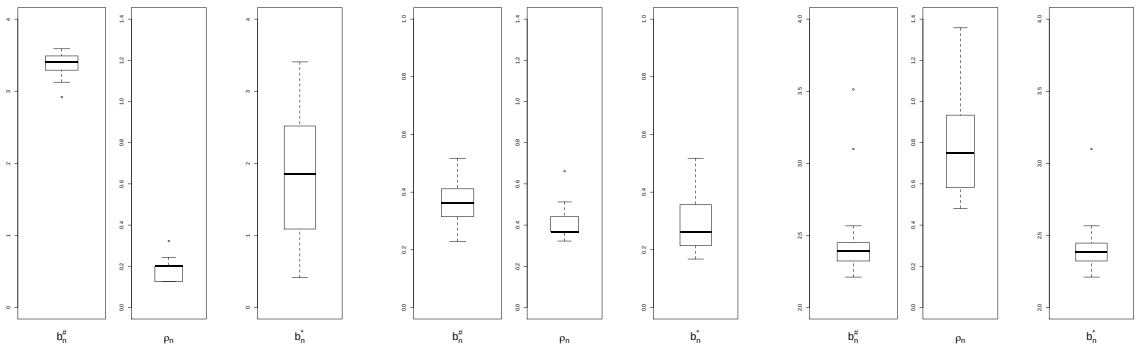


(a) (b) (c)

Figure 4.4 – Illustration of the results in the third considered situation
A simulated field considering Model B, Case 2 and $a = 50$ with (a) an image of the field Y; (b) the squared errors using $r^{\sharp}_{\mathbf{n}}$; (c) the squared errors using $r^{\star}_{\mathbf{n}}$

the selection differs according to the considered estimator. In fact, the values of $b_{\mathbf{n},opt}$ are widely lower considering $r^{\star}_{\mathbf{n}}$ rather than $r^{\sharp}_{\mathbf{n}}$. For more details on the values of the optimal bandwidths, through the 30 replications, Figure 4.5a displays the corresponding boxplots. Secondly, when $a = 20$, considering Model B and Case 1, the values of $\rho_{\mathbf{n},opt}$ are slightly higher than when $a = 5$ with values comprised between 0.322 and 0.662 (see Figure 4.5b). Finally, considering $a = 50$ with Model B and Case 2, the values of $\rho_{\mathbf{n},opt}$ are more scattered and higher than with $a = 5$ or 20 since it varies between 0.482 and 1.358 at each run (see Figure 4.5c). Moreover, for $a = 20$ and $a = 50$ the bandwidth selection of $b_{\mathbf{n},opt}$ is equivalent using $r^{\sharp}_{\mathbf{n}}$ or $r^{\star}_{\mathbf{n}}$ (see Figures 4.5b and 4.5c). In these situations, the value of $\rho_{\mathbf{n},opt}$ varies at each run while the locations do not change. In fact, contrary to the condition $a = 5$, the values of $X_{i,j}(t)$ are more scattered and then imply a change in the value of $\rho_{\mathbf{n},opt}$.

The previous study highlights the reliable performance of our estimator, particularly in presence of spatial dependence. But a disadvantage may be that the cross-validation procedure on the two parameters $b_{\mathbf{n}}$ and $\rho_{\mathbf{n}}$ is very time-consuming. To this end, we tried to deal with simulations considering a fixed bandwidth $\rho_{\mathbf{n}}$ as in Kelejian and Prucha (2007) where it is advised to take $d_{\mathbf{n}} = n\rho_{\mathbf{n}} = \lfloor \widehat{\mathbf{n}}^{1/4} \rfloor$ with $\lfloor \cdot \rfloor$ denotes the integer part. In our case, with $\widehat{\mathbf{n}} = 625$ sites, the corresponding bandwidths would be $\rho_{\mathbf{n}} \approx 0.20$. It allows to save time and obtain results that are quite satisfactory when the spatial dependence



(a) (b) (c)

Figure 4.5 – Boxplots of $b^{\sharp}_{\mathbf{n},opt}$, $\rho_{\mathbf{n},opt}$ and $b^{\star}_{\mathbf{n},opt}$ respectively
The boxplots correspond to the 30 replications of the three following situations:
(a) Model A, Case 2 and $a = 5$; (b) Model B, Case 1 and $a = 20$; (c) Model B, Case 2 and $a = 50$

is high. More precisely, when $a = 5$ the results are similar or slightly worse than those obtained by the cross-validation procedure on the two parameters: it is explained by the fact that the cross-validation procedure chooses a value of $\rho_{\mathbf{n}}$ close to 0.20 (different at each replication). Nevertheless, the fixed bandwidth $\rho_{\mathbf{n}} = 0.20$ produces better results than using the estimator $r^{*}_{\mathbf{n}}$. Note that the results depend largely on the spatial dependence structure considered. However, the results are worse with weaker spatial dependence ($a =$ 20 or 50). In fact, in some cases (depending on the spatial dependency) the performance obtained by fixing $\rho_{\mathbf{n}}$ (according to the sample size $\widehat{\mathbf{n}}$ as above) is poorer than those obtained using the estimator $r^{*}_{\mathbf{n}}$. In this case, the cross-validation procedure on the two parameters remains necessary.

4.4 Conclusion

This work proposes a new method to model spatial regression function for functional random fields. Our main theoretical contribution was to derive the convergence in mean square. One can see the proposed methodology as a good alternative to the classical kernel approach for functional spatial data. More precisely, it is apparent that the proposed approach is particularly well adapted to the spatial regression estimation, for functional data, in presence of spatial dependence. This good behavior is observed both from an asymptotic point of view and from a simulation study. However, in case of low spatial dependence, the two estimators, herein called $r^{\sharp}_{\mathbf{n}}$ and $r^{\star}_{\mathbf{n}}$, produce similar results.

In addition, this work offers very interesting perspectives of investigation. First of all, a future work will be tied up to the uniform convergence of our estimator. Then, we could improve the choice of $b_{\mathbf{n}}$ and $\rho_{\mathbf{n}}$ which is outside the scope of this chapter. For further study, we could investigate this new approach using local linear spatial regression (see, for example, Hallin et al. (2004b)). Also, an adaptation of this method to issues such as the spatial conditional mode or quantile regression estimation could be developed. Application of the proposed regression estimator to real data, and more particularly to data collected by the French Research Institute for Exploitation of the Sea (Ifremer) during the campaign IBTS (International Bottom Trawl Survey), will be investigated. Moreover, an other perspective is the study of regression estimation for continuous indexed spatial functional fields $\{Z_{\mathbf{i}},\, \mathbf{i} \in \mathbb{R}^N\}$ that can be applied to spatial prediction.

4.5 Appendix

Some preliminary results for the proofs

In the following, we will often use the notation

$$K_{1\mathbf{i}} = K_1\left(\frac{d(x_{\mathbf{j}}, X_{\mathbf{i}})}{b_{\mathbf{n}}}\right) \qquad \text{and} \qquad K_{2\mathbf{i}} = K_{2,\rho_{\mathbf{n}}}(\|\mathbf{j}-\mathbf{i}\|).$$

Let $W_{\mathbf{ni}} = \dfrac{K_{1\mathbf{i}}K_{2\mathbf{i}}}{\sum_{\mathbf{k}\in\mathcal{I}_{\mathbf{n}}} K_{1\mathbf{k}}K_{2\mathbf{k}}}$ with the convention $0/0=0$, then $\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}} = 0$ or 1. Thus, we have

$$r_{\mathbf{n}}(x_{\mathbf{j}}) = \begin{cases} \displaystyle\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}}Y_{\mathbf{i}} & \text{if} \quad \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}} = 1; \\ \dfrac{1}{\widehat{\mathbf{n}}}\displaystyle\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} & \text{otherwise.} \end{cases}$$

Lemma 4.5. Under hypothesis **H2**, we have

$$\mathbb{E}^{1/2}\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}}\mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}}) - r(x_{\mathbf{j}})\right]^2 = O(b_{\mathbf{n}}).$$

Lemma 4.6. Under hypotheses **H1**, **H3**, **H4** or **H5** and **H6**, we have

$$\mathbb{E}^{1/2}\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}}(Y_{\mathbf{i}} - \mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}}))\right]^2 = O\left(\frac{1}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}})}\right)^{1/2}.$$

Sketch of the proof for Lemma 4.6: The expression $W_{\mathbf{ni}}(Y_{\mathbf{i}}-\mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}}))$ is decomposed in the sum of two terms, for which it is sufficient to show that:

1. $\|e_{\mathbf{n}}(x_{\mathbf{j}})\|_2 = \left\|\frac{1}{a_{\mathbf{n},\mathbf{j}}\mathbb{E}[K_{1\mathbf{i}}]}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{1\mathbf{i}}K_{2\mathbf{i}}[Y_{\mathbf{i}} - \mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}})]\right\|_2 = O(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}}))^{-1/2}$. To obtain this result, we let $\xi_{\mathbf{i}} = K_{1\mathbf{i}}K_{2\mathbf{i}}[Y_{\mathbf{i}}-\mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}})]$ and study $\mathbb{E}\left(\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}\xi_{\mathbf{i}}\right)^2 = \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}\mathbb{E}\left[\xi_{\mathbf{i}}^2\right] + \sum_{\mathbf{i},\mathbf{k}\in S}\mathbb{E}\left[\xi_{\mathbf{i}}\xi_{\mathbf{k}}\right] + \sum_{\mathbf{i},\mathbf{k}\in S^c}\mathbb{E}\left[\xi_{\mathbf{i}}\xi_{\mathbf{k}}\right]$ with $S = \{\mathbf{i},\mathbf{k}\in\mathcal{I}_{\mathbf{n}}, \|\mathbf{i}-\mathbf{k}\|\le D_{\mathbf{n}}\}$ and denote by S^c the complementary of S. Moreover $D_{\mathbf{n}}$ is a sequence of real numbers tending to ∞ as $\widehat{\mathbf{n}}\to\infty$.
2. $\mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{1\mathbf{i}}K_{2\mathbf{i}} \le \frac{a_{\mathbf{n},\mathbf{j}}u}{2}\right] = O(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}}))^{-1/2}$ using the well-known spatial block decomposition (Tran (1990)), Markov and Bernstein inequalities and Lemmas A.2 and 4.8, with $u = \mathbb{E}[K_{1\mathbf{i}}]$.

Lemma 4.7. Under the hypotheses of Lemma 4.6, we have

$$\mathbb{E}^{1/2}\left[\frac{1}{\widehat{\mathbf{n}}}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} - r(x_{\mathbf{j}})\right]^2 = O\left(\frac{1}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}})}\right)^{1/2}.$$

Lemma 4.8. Under the hypotheses **H1** and **H3**, we have

$$I_{\mathbf{n}}(x_{\mathbf{j}}) + R_{\mathbf{n}}(x_{\mathbf{j}}) = O\left(\frac{1}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}})}\right).$$

where

$$I_{\mathbf{n}}(x_{\mathbf{j}}) = \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}\mathbb{E}\left[\left(\Lambda_{\mathbf{i}}(x_{\mathbf{j}})\right)^2\right] \qquad \text{and} \qquad R_{\mathbf{n}}(x_{\mathbf{j}}) = \sum_{\mathbf{i},\mathbf{k}\in\mathcal{I}_{\mathbf{n}}}\sum_{\mathbf{i}\neq\mathbf{k}}\left|\mathbb{E}\left[\Lambda_{\mathbf{i}}(x_{\mathbf{j}})\Lambda_{\mathbf{k}}(x_{\mathbf{j}})\right]\right|$$

with $\Lambda_{\mathbf{i}}(x_{\mathbf{j}}) = \dfrac{1}{a_{\mathbf{n},\mathbf{j}}\mathbb{E}(K_{1\mathbf{i}})}\left[K_{1\mathbf{i}}K_{2\mathbf{i}} - \mathbb{E}(K_{1\mathbf{i}}K_{2\mathbf{i}})\right]$.

Proofs

Proof of Theorem 4.3

We study the expression $\|r_{\mathbf{n}}(x_{\mathbf{j}}) - r(x_{\mathbf{j}})\|_2 = (\mathbb{E}[|r_{\mathbf{n}}(x_{\mathbf{j}}) - r(x_{\mathbf{j}})|^2])^{1/2}$.

$$
\begin{aligned}
r_{\mathbf{n}}(x_{\mathbf{j}}) - r(x_{\mathbf{j}}) = & \left(\sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}} \mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}}) - r(x_{\mathbf{j}}) \right) \mathbf{1}_{\left[\sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}} = 1\right]} \\
& + \left(\sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}} (Y_{\mathbf{i}} - \mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}})) \right) \mathbf{1}_{\left[\sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}} = 1\right]} \\
& + \left(\frac{1}{\widehat{\mathbf{n}}} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} - r(x_{\mathbf{j}}) \right) \mathbf{1}_{\left[\sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}} = 0\right]}
\end{aligned}
$$

$$
\|r_{\mathbf{n}}(x_{\mathbf{j}}) - r(x_{\mathbf{j}})\|_2 \leq \mathbb{E}^{1/2}[\mathbf{A}]^2 + \mathbb{E}^{1/2}[\mathbf{B}]^2 + \mathbb{E}^{1/2}[\mathbf{C}]^2
$$

applying Minkowski's inequality. The terms on the right-hand-side of the previous equation are dealt in the Lemmas 4.5, 4.6 and 4.7 respectively. ■

Proof of Lemma 4.5

$$
\begin{aligned}
\mathbb{E}^{1/2}[\mathbf{A}]^2 &\leq \mathbb{E}^{1/2} \left[\left(\sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}} |r(X_{\mathbf{i}}) - r(x_{\mathbf{j}})| \right) \mathbf{1}_{\left[\sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}} = 1\right]} \right]^2 \\
&\leq \mathbb{E}^{1/2} \left[\left(\sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}} (C_3 \times d(X_{\mathbf{i}}, x_{\mathbf{j}})) \right) \mathbf{1}_{\left[\sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}} = 1\right]} \right]^2 \\
&\leq \mathbb{E}^{1/2} \left[C_3 \times \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}} b_{\mathbf{n}} \right]^2 \\
&\leq \mathbb{E}^{1/2} \left[C_3 \times b_{\mathbf{n}} \right]^2 \\
&= O(b_{\mathbf{n}}),
\end{aligned}
$$

by assumptions **H1** and **H2** (Lipschitz condition). ■

Proof of Lemma 4.6

Let

$$
G = \left(\sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}} [Y_{\mathbf{i}} - \mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}})] \right) \mathbf{1}_{\left[\sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}} = 1\right]} = \left(\frac{e_{\mathbf{n}}(x_{\mathbf{j}})}{f_{\mathbf{n}}(x_{\mathbf{j}})} \right) \mathbf{1}_{\left[\sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}} = 1\right]}
$$

with

$$
e_{\mathbf{n}}(x_{\mathbf{j}}) = \frac{1}{a_{\mathbf{n},\mathbf{j}} \mathbb{E}\left[K_{1\mathbf{i}}\right]} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K_{1\mathbf{i}} K_{2\mathbf{i}} [Y_{\mathbf{i}} - \mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}})]
$$

$$
f_{\mathbf{n}}(x_{\mathbf{j}}) = \frac{1}{a_{\mathbf{n},\mathbf{j}} \mathbb{E}\left[K_{1\mathbf{i}}\right]} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K_{1\mathbf{i}} K_{2\mathbf{i}}.
$$

Note that $\forall \mathbf{i}$: $0 \leq |Y_{\mathbf{i}} - \mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}})| \leq 2M$, then, $|G| \leq \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}} 2M \leq 2M$.

$$
\begin{aligned}
|G| &= |G|\mathbf{1}_{\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{1\mathbf{i}}K_{2\mathbf{i}}>c\right]} + |G|\mathbf{1}_{\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{1\mathbf{i}}K_{2\mathbf{i}}\leq c\right]} \\
&\leq \frac{|e_{\mathbf{n}}(x_{\mathbf{j}})|}{f_{\mathbf{n}}(x_{\mathbf{j}})}\mathbf{1}_{\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{1\mathbf{i}}K_{2\mathbf{i}}>c\right]} + 2M \times \mathbf{1}_{\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{1\mathbf{i}}K_{2\mathbf{i}}\leq c\right]}
\end{aligned}
$$

where c is a given constant. Let us take $c = \frac{a_{\mathbf{n},\mathbf{j}}u}{2}$ with $u = \mathbb{E}[K_{1\mathbf{i}}] \leq C \times \varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}})$ since by assumption **H1**, we have

$$C_1\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}}) \leq \mathbb{E}\left[K_{1\mathbf{i}}\right] \leq C_2\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}}).$$

If $\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{1\mathbf{i}}K_{2\mathbf{i}} > c = \frac{a_{\mathbf{n},\mathbf{j}}u}{2}$ then $f_{\mathbf{n}}(x_{\mathbf{j}}) > \dfrac{a_{\mathbf{n},\mathbf{j}}u}{2a_{\mathbf{n},\mathbf{j}}\mathbb{E}\left[K_{1\mathbf{i}}\right]} > \dfrac{1}{2}$. Consequently,

$$\|G\|_2 \leq 2\|e_{\mathbf{n}}(x_{\mathbf{j}})\|_2 + 2M\left(\mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{1\mathbf{i}}K_{2\mathbf{i}} \leq \frac{a_{\mathbf{n},\mathbf{j}}u}{2}\right]\right)^{1/2}.$$

$$\diamondsuit \quad \|e_{\mathbf{n}}(x_{\mathbf{j}})\|_2 = \frac{1}{a_{\mathbf{n},\mathbf{j}}\mathbb{E}\left[K_{1\mathbf{i}}\right]}\left[\mathbb{E}\left(\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}\xi_{\mathbf{i}}\right)^2\right]^{1/2}$$

where $\xi_{\mathbf{i}} = K_{1\mathbf{i}}K_{2\mathbf{i}}\theta_{\mathbf{i}}$ with $\theta_{\mathbf{i}} = Y_{\mathbf{i}} - \mathbb{E}(Y_{\mathbf{i}}|X_{\mathbf{i}})$.
Let $D_{\mathbf{n}}$ be a sequence of real numbers tending to ∞ as $\widehat{\mathbf{n}} \to \infty$. Let $S = \{\mathbf{i},\mathbf{k} \in \mathcal{I}_{\mathbf{n}}, \|\mathbf{i} - \mathbf{k}\| \leq D_{\mathbf{n}}\}$ and denote by S^c the complementary of S. Let $V_{\mathbf{j}} = \left\{\mathbf{i}, \left\|\frac{\mathbf{i}-\mathbf{j}}{\mathbf{n}}\right\| \leq \rho_{\mathbf{n}}\right\}$ with $\mathrm{Card}(V_{\mathbf{j}}) = k_{\mathbf{n}} = C_N\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N + O((\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N)^{\beta})$, see the definition of $r_{\mathbf{n}}(x_{\mathbf{j}})$. First, we are interested in

$$\mathbb{E}\left(\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}\xi_{\mathbf{i}}\right)^2 = \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}\mathbb{E}\left[\xi_{\mathbf{i}}^2\right] + \sum_{\mathbf{i},\mathbf{k}\in S}\mathbb{E}\left[\xi_{\mathbf{i}}\xi_{\mathbf{k}}\right] + \sum_{\mathbf{i},\mathbf{k}\in S^c}\mathbb{E}\left[\xi_{\mathbf{i}}\xi_{\mathbf{k}}\right]$$

- $$
\begin{aligned}
\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}\mathbb{E}\left[\xi_{\mathbf{i}}^2\right] \leq \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}\mathbb{E}\left[(K_{1\mathbf{i}}K_{2\mathbf{i}}|\theta_{\mathbf{i}}|)^2\right] &\leq 4M^2\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}K_{2\mathbf{i}}^2\mathbb{E}\left[(K_{1\mathbf{i}})^2\right] \\
&\leq 4M^2 \times k_{\mathbf{n}}\mathbb{E}[(K_{1\mathbf{i}})^2] \\
&\leq 4M^2 \times C_2^2 \times k_{\mathbf{n}}\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}}) \\
&= O(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}})),
\end{aligned}
$$

since $C_1^2\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}}) \leq \mathbb{E}\left[K_{1\mathbf{i}}^2\right] \leq C_2^2\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}})$.

- $$
\begin{aligned}
\sum_{\mathbf{i},\mathbf{k}\in S}\mathbb{E}\left[\xi_{\mathbf{i}}\xi_{\mathbf{k}}\right] &\leq 4M^2\sum_{\mathbf{i},\mathbf{k}\in S}K_{2\mathbf{i}}K_{2\mathbf{k}}\mathbb{E}\left[K_{1\mathbf{i}}K_{1\mathbf{k}}\right] \\
&\leq 4M^2\sum_{\mathbf{i},\mathbf{k}\in S}K_{2\mathbf{i}}K_{2\mathbf{k}}\mathbb{P}\left[(X_{\mathbf{i}},X_{\mathbf{k}}) \in \mathcal{B}(x_{\mathbf{j}},b_{\mathbf{n}}) \times \mathcal{B}(x_{\mathbf{j}},b_{\mathbf{n}})\right]
\end{aligned}
$$

By assumption **H3**, we have

$$\begin{aligned}\sum_{\mathbf{i},\mathbf{k}\in S}\mathbb{E}\left[\xi_{\mathbf{i}}\xi_{\mathbf{k}}\right] &\leq 4M^2C_4\sum_{\mathbf{i},\mathbf{k}\in S}\mathbf{1}_{[0,1]}\left(\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{j}-\mathbf{i}}{\mathbf{n}}\right\|\right)\mathbf{1}_{[0,1]}\left(\rho_{\mathbf{n}}^{-1}\left\|\frac{\mathbf{j}-\mathbf{k}}{\mathbf{n}}\right\|\right)(\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}}))^{1+\epsilon_1}\\ &\leq 4M^2C_4\sum_{\mathbf{i},\mathbf{k}\in V_{\mathbf{j}}}\mathbf{1}_{[0,1]}\left(\frac{\|\mathbf{i}-\mathbf{k}\|}{D_{\mathbf{n}}}\right)(\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}}))^{1+\epsilon_1}\\ &\leq 4M^2C_4\sum_{\mathbf{i}\in V_{\mathbf{j}}}\sum_{\mathbf{i}-\mathbf{u}\in V_{\mathbf{j}}}\mathbf{1}_{\{\mathbf{u};\|\mathbf{u}\|\leq D_{\mathbf{n}}\}}(\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}}))^{1+\epsilon_1}\\ &\leq 4M^2C_4k_{\mathbf{n}}D_{\mathbf{n}}^N(\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}}))^{1+\epsilon_1}\end{aligned}$$

- $$\sum_{\mathbf{i},\mathbf{k}\in S^c}\mathbb{E}\left[\xi_{\mathbf{i}}\xi_{\mathbf{k}}\right] \leq \sum_{\mathbf{i},\mathbf{k}\in S^c}\left|\mathbb{E}\left(K_{1\mathbf{i}}K_{2\mathbf{i}}K_{1\mathbf{k}}K_{2\mathbf{k}}Y_{\mathbf{i}}Y_{\mathbf{k}} - K_{1\mathbf{i}}K_{2\mathbf{i}}\mathbb{E}\left[Y_{\mathbf{i}}|X_{\mathbf{i}}\right]K_{1\mathbf{k}}K_{2\mathbf{k}}\mathbb{E}[Y_{\mathbf{k}}|X_{\mathbf{k}}]\right)\right|$$

and, since the function K_1 and K_2 are bounded, we get by applying Lemma A.1

$$\begin{aligned}\sum_{\mathbf{i},\mathbf{k}\in S^c}\mathbb{E}\left[\xi_{\mathbf{i}}\xi_{\mathbf{k}}\right] \leq C\sum_{\mathbf{i},\mathbf{k}\in S^c}\{\psi(1,1)\gamma(\|\mathbf{i}-\mathbf{k}\|)\} &\leq C\sum_{\mathbf{i},\mathbf{k}\in S^c\cap V_{\mathbf{j}}}\gamma(\|\mathbf{i}-\mathbf{k}\|)\\ &\leq C2^N\sum_{\mathbf{k}\in V_{\mathbf{j}}}\sum_{\substack{\mathbf{k}-\mathbf{u}\in V_{\mathbf{j}},\\ \|\mathbf{u}\|>D_{\mathbf{n}}}}\gamma(\|\mathbf{u}\|)\\ &\leq Ck_{\mathbf{n}}\sum_{\|\mathbf{i}\|>D_{\mathbf{n}}}\gamma(\|\mathbf{i}\|).\end{aligned}$$

Since $\sum_{\|\mathbf{i}\|>D_{\mathbf{n}}}\gamma(\|\mathbf{i}\|) \leq C\sum_{\|\mathbf{i}\|>D_{\mathbf{n}}}\|\mathbf{i}\|^{-\theta} \leq C\sum_{\|\mathbf{i}\|>D_{\mathbf{n}}}\|\mathbf{i}\|^{-\theta}\|\mathbf{i}\|^{-N}\|\mathbf{i}\|^{N}$ and $\|\mathbf{i}\| > D_{\mathbf{n}}$, $\|\mathbf{i}\|^{-N} \leq (D_{\mathbf{n}})^{-N}$, we have

$$\begin{aligned}C\sum_{\|\mathbf{i}\|>D_{\mathbf{n}}}\|\mathbf{i}\|^{-\theta}\|\mathbf{i}\|^{-N-\epsilon_1}\|\mathbf{i}\|^{N+\epsilon_1} &\leq C(D_{\mathbf{n}})^{-N-\epsilon_1}\sum_{\|\mathbf{i}\|>D_{\mathbf{n}}}\|\mathbf{i}\|^{-\theta}\|\mathbf{i}\|^{N+\epsilon_1}\\ &\leq CD_{\mathbf{n}}^{-N-\epsilon_1}\sum_{\|\mathbf{i}\|>D_{\mathbf{n}}}\|\mathbf{i}\|^{N+\epsilon_1-\theta}\end{aligned}$$

and then $\sum_{\mathbf{i},\mathbf{k}\in S^c}\mathbb{E}\left[\xi_{\mathbf{i}}\xi_{\mathbf{k}}\right] \leq Ck_{\mathbf{n}}D_{\mathbf{n}}^{-N-\epsilon_1}\sum_{\|\mathbf{i}\|>D_{\mathbf{n}}}\|\mathbf{i}\|^{N+\epsilon_1-\theta}$. The fact that $\theta > 4N > N+1$ leads to choose $D_{\mathbf{n}} = (\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}}))^{\frac{-\epsilon_1}{N}+a}$ with $a > 0$ and such that $Na \leq \epsilon_1 - \frac{N}{N+\epsilon_1}$. In fact, these conditions lead to

$$\begin{aligned}D_{\mathbf{n}}^{-(N+\epsilon_1)} &= \varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}})(\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}}))^{\frac{-(N+\epsilon_1)(Na-\epsilon_1)-N}{N}}\\ &= O\left(\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}})\right)\end{aligned}$$

since $\frac{-(N+\epsilon_1)(Na-\epsilon_1)-N}{N} > 0$. Moreover, this choice of $D_{\mathbf{n}}$ implies that

$$\begin{aligned}\sum_{\mathbf{i},\mathbf{k}\in S}\mathbb{E}\left[\xi_{\mathbf{i}}\xi_{\mathbf{k}}\right] &\leq 4M^2C_4k_{\mathbf{n}}D_{\mathbf{n}}^N(\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}}))^{1+\epsilon_1}\\ &\leq 4M^2C_4k_{\mathbf{n}}(\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}}))^{1+Na}\\ &= O(\widehat{\mathbf{n}}\rho_{\mathbf{n}}^N\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}}))\end{aligned}$$

Consequently, $\left[\mathbb{E}\left(\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}}\xi_\mathbf{i}\right)^2\right]^{1/2} = O(\widehat{\mathbf{n}}\rho_\mathbf{n}^N\varphi_{x_\mathbf{j}}(b_\mathbf{n}))^{1/2}$ and $\|e_\mathbf{n}(x_\mathbf{j})\|_2 = O\left(\widehat{\mathbf{n}}\rho_\mathbf{n}^N\varphi_{x_\mathbf{j}}(b_\mathbf{n})\right)^{-1/2}$. Second, we deal with

$$\begin{aligned}\diamondsuit \quad P = \mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} K_{1\mathbf{i}}K_{2\mathbf{i}} \le \frac{a_{\mathbf{n},\mathbf{j}}u}{2}\right] &= \mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} (K_{1\mathbf{i}}K_{2\mathbf{i}} - \mathbb{E}(K_{1\mathbf{i}}K_{2\mathbf{i}})) \le \frac{-a_{\mathbf{n},\mathbf{j}}u}{2}\right] \\ &\le \mathbb{P}\left[\frac{1}{a_{\mathbf{n},\mathbf{j}}u}\left|\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} (K_{1\mathbf{i}}K_{2\mathbf{i}} - \mathbb{E}(K_{1\mathbf{i}}K_{2\mathbf{i}}))\right| \ge \frac{1}{2}\right] \\ &\le \mathbb{P}\left[\left|S_\mathbf{n}(x_\mathbf{j})\right| \ge \epsilon\right]\end{aligned}$$

with $S_\mathbf{n}(x_\mathbf{j}) = \sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}}\Lambda_\mathbf{i}(x_\mathbf{j}) = \sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}}\frac{1}{a_{\mathbf{n},\mathbf{j}}u}(K_{1\mathbf{i}}K_{2\mathbf{i}} - \mathbb{E}(K_{1\mathbf{i}}K_{2\mathbf{i}}))$. We will now introduce the spatial blocks decomposition introduced by Tran (1990) which will be useful afterwards. Without loss of generality, we suppose that $n_k = 2bq_k$, for $1 \le k \le N$. The random variables $\Lambda_\mathbf{i}(x_\mathbf{j})$ can be grouped into $2^N q_1 \dots q_N$ cubic blocks of size b. Let,

$$U(1,\mathbf{n},x_\mathbf{j},\mathbf{j}) = \sum_{\substack{i_k=2j_kb+1,\\ k=1,\dots,N.}}^{(2j_k+1)b}\Lambda_\mathbf{i}(x_\mathbf{j}),$$

$$U(2,\mathbf{n},x_\mathbf{j},\mathbf{j}) = \sum_{\substack{i_k=2j_kb+1,\\ k=1,\dots,N-1.}}^{(2j_k+1)b}\ \sum_{i_N=(2j_N+1)b+1}^{2(j_N+1)b}\Lambda_\mathbf{i}(x_\mathbf{j}),$$

$$U(3,\mathbf{n},x_\mathbf{j},\mathbf{j}) = \sum_{\substack{i_k=2j_kb+1,\\ k=1,\dots,N-2.}}^{(2j_k+1)b}\ \sum_{i_{N-1}=(2j_{N-1}+1)b+1}^{2(j_{N-1}+1)b}\ \sum_{i_N=2j_Nb+1}^{(2j_N+1)b}\Lambda_\mathbf{i}(x_\mathbf{j}),$$

$$U(4,\mathbf{n},x_\mathbf{j},\mathbf{j}) = \sum_{\substack{i_k=2j_kb+1,\\ k=1,\dots,N-2.}}^{(2j_k+1)b}\ \sum_{i_{N-1}=(2j_{N-1}+1)b+1}^{2(j_{N-1}+1)b}\ \sum_{i_N=(2j_N+1)b+1}^{(2j_N+1)b}\Lambda_\mathbf{i}(x_\mathbf{j})$$

and so on. Noticing that

$$U(2^{N-1},\mathbf{n},x_\mathbf{j},\mathbf{j}) = \sum_{\substack{i_k=(2j_k+1)b+1,\\ k=1,\dots,N-1.}}^{2(j_k+1)b}\ \sum_{i_N=2j_Nb+1}^{(2j_N+1)b}\Lambda_\mathbf{i}(x_\mathbf{j})$$

$$U(2^N,\mathbf{n},x_\mathbf{j},\mathbf{j}) = \sum_{\substack{i_k=(2j_k+1)b+1,\\ k=1,\dots,N.}}^{2(j_k+1)b}\Lambda_\mathbf{i}(x_\mathbf{j})$$

for each integer $1 \le l \le 2^N$, we define $T(\mathbf{n},x_\mathbf{j},l) = \sum_{\substack{j_k=0\\ k=1,\dots,N.}}^{q_k-1} U(l,\mathbf{n},x_\mathbf{j},\mathbf{j})$. We obtain $S_\mathbf{n}(x_\mathbf{j}) = \sum_{l=1}^{2^N} T(\mathbf{n},x_\mathbf{j},l)$. For $\epsilon > 0$,

$$P \le \mathbb{P}\left(\left|\sum_{l=1}^{2^N} T(\mathbf{n},x_\mathbf{j},l)\right| > \epsilon\right) \le 2^N\mathbb{P}\left(|T(\mathbf{n},x_\mathbf{j},1)| > \frac{\epsilon}{2^N}\right).$$

We enumerate in arbitrary manner the $\widehat{q} = q_1 \times \dots \times q_N$ terms $U(1,\mathbf{n},x_\mathbf{j},\mathbf{j})$ of the sum $T(\mathbf{n},x_\mathbf{j},1)$, and refer to them as $W_1,\dots,W_{\widehat{q}}$. Note that $U(1,\mathbf{n},x_\mathbf{j},\mathbf{j})$ is a measurable σ-algebra generated by $X_\mathbf{i}$, with $\mathbf{i}$ such that $2j_kb+1 \le i_k \le (2j_k+1)b$, $k = 1,\dots,N$.

For all $l = 1, \dots, \widehat{q}$, the sets of the sites in W_l are separated by a distance of at least equal to b. In addition, since K_2 and K_1 are bounded, we can write $|W_l| \leq C\frac{b^N}{a_{\mathbf{n},\mathbf{j}}u}$ with $C = \|K_1\|_\infty \|K_2\|_\infty$ (where $\|\cdot\|_\infty$ is the sup norm). Lemma A.2 insures the existence of some random variables $W_1^*, W_2^*, \dots, W_{\widehat{q}}^*$ such that

$$\begin{aligned}\sum_{l=1}^{\widehat{q}} \mathbb{E}|W_l - W_l^*| &\leq 2\widehat{q}C\frac{b^N}{a_{\mathbf{n},\mathbf{j}}u}\psi((\widehat{q}-1)b^N, b^N)\gamma(b) \\ &\leq 2C\frac{\widehat{\mathbf{n}}}{2^N b^N}\frac{b^N}{a_{\mathbf{n},\mathbf{j}}u}\psi(\widehat{\mathbf{n}}, b^N)\gamma(b).\end{aligned}$$

Markov inequality allows us to write

$$\mathbb{P}\left(\sum_{l=1}^{\widehat{q}} |W_l - W_l^*| > \frac{\epsilon}{2^{N+1}}\right) \leq 2C\frac{\widehat{\mathbf{n}}}{2^N b^N}\frac{b^N}{a_{\mathbf{n},\mathbf{j}}u}\psi(\widehat{\mathbf{n}}, b^N)\gamma(b)\frac{2^{N+1}}{\epsilon},$$

and by Bernstein inequality, we have

$$\mathbb{P}\left(\sum_{l=1}^{\widehat{q}} |W_l^*| > \frac{\epsilon}{2^{N+1}}\right) \leq 2\exp\left\{\frac{-\epsilon^2/(2^{N+1})^2}{4\sum_{l=1}^{\widehat{q}}\mathbb{E}(W_l^{*2}) + \frac{2\epsilon}{2^{N+1}}\frac{b^N}{a_{\mathbf{n},\mathbf{j}}u}C}\right\}$$

which leads to

$$P \leq 2^{N+1}\exp\left\{\frac{-\epsilon^2/(2^{N+1})^2}{4\sum_{l=1}^{\widehat{q}}\mathbb{E}(W_l^{*2}) + 2^{-N}C\epsilon\frac{b^N}{a_{\mathbf{n},\mathbf{j}}u}}\right\} + 2^{N+1}C\frac{\widehat{\mathbf{n}}}{2^N b^N}\frac{b^N}{a_{\mathbf{n},\mathbf{j}}u}\psi(\widehat{\mathbf{n}}, b^N)\gamma(b)\frac{2^{N+1}}{\epsilon}$$

Let $\delta > 0$,

$$\epsilon = \epsilon_{\mathbf{n}} = \delta\left(\frac{\log\widehat{\mathbf{n}}}{\widehat{\mathbf{n}}\varphi_{x_\mathbf{j}}(b_\mathbf{n})\rho_\mathbf{n}^N}\right)^{1/2} \qquad \text{and} \qquad b = \left\lfloor\left(\frac{\widehat{\mathbf{n}}\varphi_{x_\mathbf{j}}(b_\mathbf{n})\rho_\mathbf{n}^N}{\log\widehat{\mathbf{n}}}\right)^{\frac{1}{2N}}\right\rfloor.$$

Since the variables W_l and W_l^* have the same distributions, we have

$$\begin{aligned}\sum_{l=1}^{\widehat{q}}\mathbb{E}W_l^{*2} = \sum_{l=1}^{\widehat{q}} Var(W_l^*) &= \sum_{l=1}^{\widehat{q}} Var(W_l) \\ &\leq I_\mathbf{n}(x_\mathbf{j}) + R_\mathbf{n}(x_\mathbf{j}),\end{aligned}$$

and according to Lemma 4.8, we have $\sum_{l=1}^{\widehat{q}}\mathbb{E}W_l^{*2} \leq O\left([\widehat{\mathbf{n}}\rho_\mathbf{n}^N\varphi_{x_\mathbf{j}}(b_\mathbf{n})]^{-1}\right)$. Then,

$$P \leq 2^{N+1}\exp\left\{\frac{-\epsilon^2}{2^{2N+2}\left(4\frac{C}{\widehat{\mathbf{n}}\rho_\mathbf{n}^N\varphi_{x_\mathbf{j}}(b_\mathbf{n})} + C2^{-N}\epsilon\frac{b^N}{a_{\mathbf{n},\mathbf{j}}u}\right)}\right\} + 2^{N+2}C\frac{\widehat{\mathbf{n}}}{a_{\mathbf{n},\mathbf{j}}u}\psi(\widehat{\mathbf{n}}, b^N)b^{-\theta}\epsilon^{-1}$$

Since $C_1 k_{\mathbf{n}} \leq a_{\mathbf{n},\mathbf{j}} \leq C_2 k_{\mathbf{n}}$ and $k_{\mathbf{n}} = C_N d_{\mathbf{n}}^N + O(d_{\mathbf{n}}^\beta)$, $\beta < N$, we have

$$\begin{aligned} P &\leq 2^{N+1} \exp\left\{ \frac{-\delta^2 \frac{\log \widehat{\mathbf{n}}}{\widehat{\mathbf{n}}\varphi_{x_\mathbf{j}}(b_\mathbf{n})\rho_\mathbf{n}^N}}{\frac{2^{2N+4}C}{\widehat{\mathbf{n}}\rho_\mathbf{n}^N \varphi_{x_\mathbf{j}}(b_\mathbf{n})} + \frac{C2^{N+2}\delta}{\widehat{\mathbf{n}}\rho_\mathbf{n}^N \varphi_{x_\mathbf{j}}(b_\mathbf{n})}} \right\} + 2^{N+2} C \frac{\widehat{\mathbf{n}}}{a_{\mathbf{n},\mathbf{j}} u} \psi(\widehat{\mathbf{n}}, b^N) b^{-\theta} \delta^{-1} \left(\frac{\widehat{\mathbf{n}}\varphi_{x_\mathbf{j}}(b_\mathbf{n})\rho_\mathbf{n}^N}{\log \widehat{\mathbf{n}}} \right)^{1/2} \\ &\leq 2^{N+1} \exp\left\{ \log \widehat{\mathbf{n}}^{-a} \right\} + 2^{N+2} C\delta^{-1} \frac{\widehat{\mathbf{n}}}{a_{\mathbf{n},\mathbf{j}} u} \psi(\widehat{\mathbf{n}}, b^N) \left(\frac{\widehat{\mathbf{n}}\varphi_{x_\mathbf{j}}(b_\mathbf{n})\rho_\mathbf{n}^N}{\log \widehat{\mathbf{n}}} \right)^{\frac{N-\theta}{2N}} \\ &\leq C\widehat{\mathbf{n}}^{-a} + 2^{N+2} C\delta^{-1} \frac{\widehat{\mathbf{n}}}{a_{\mathbf{n},\mathbf{j}} \varphi_{x_\mathbf{j}}(b_\mathbf{n})} \psi(\widehat{\mathbf{n}}, b^N) \left(\frac{\widehat{\mathbf{n}}\varphi_{x_\mathbf{j}}(b_\mathbf{n})\rho_\mathbf{n}^N}{\log \widehat{\mathbf{n}}} \right)^{\frac{N-\theta}{2N}} \end{aligned}$$

with $a = \dfrac{\delta^2}{2^{2N+4}C + C_N 2^{N+2}\delta} > 0$. Note that $\widehat{\mathbf{n}}^{1-a}\varphi_{x_\mathbf{j}}(b_\mathbf{n})\rho_\mathbf{n}^N$ tends to 0 for $a > 1$ and then $C\widehat{\mathbf{n}}^{-a} = o\left([\widehat{\mathbf{n}}\varphi_{x_\mathbf{j}}(b_\mathbf{n})\rho_\mathbf{n}^N]^{-1}\right)$. Moreover $a > 1$ if and only if $\delta > 2^{N+1}C(1+\sqrt{4C}) > 2^{N+1}C$ (with $\delta > 0$). Now, we treat the second term. From assumptions on $\psi(n,m)$, two cases arise.

First case: $\psi(n,m) \leq C\min(n,m)$, $\forall n, m \in \mathbb{N}$

$$\begin{aligned} &\widehat{\mathbf{n}}\varphi_{x_\mathbf{j}}(b_\mathbf{n})\rho_\mathbf{n}^N C2^{N+2}\delta^{-1} \frac{\widehat{\mathbf{n}}}{a_{\mathbf{n},\mathbf{j}}\varphi_{x_\mathbf{j}}(b_\mathbf{n})} b^N \left(\frac{\widehat{\mathbf{n}}\varphi_{x_\mathbf{j}}(b_\mathbf{n})\rho_\mathbf{n}^N}{\log \widehat{\mathbf{n}}} \right)^{\frac{N-\theta}{2N}} \\ &\qquad \leq \widehat{\mathbf{n}}\rho_\mathbf{n}^N C2^{N+2}\delta^{-1} \frac{\widehat{\mathbf{n}}}{a_{\mathbf{n},\mathbf{j}}} \left(\frac{\widehat{\mathbf{n}}\varphi_{x_\mathbf{j}}(b_\mathbf{n})\rho_\mathbf{n}^N}{\log \widehat{\mathbf{n}}} \right)^{\frac{2N-\theta}{2N}} \\ &\qquad \leq \widehat{\mathbf{n}}\rho_\mathbf{n}^N C_N 2^{N+2}\delta^{-1} \frac{1}{\rho_\mathbf{n}^N} \left(\frac{\widehat{\mathbf{n}}\varphi_{x_\mathbf{j}}(b_\mathbf{n})\rho_\mathbf{n}^N}{\log \widehat{\mathbf{n}}} \right)^{\frac{2N-\theta}{2N}} \\ &\qquad \leq C_N \left[\widehat{\mathbf{n}}\varphi_{x_\mathbf{j}}(b_\mathbf{n})^{\frac{2N-\theta}{4N-\theta}} \rho_\mathbf{n}^{\frac{N(2N-\theta)}{4N-\theta}} \log \widehat{\mathbf{n}}^{\frac{\theta-2N}{4N-\theta}} \right]^{\frac{4N-\theta}{2N}} \end{aligned}$$

which tends to 0 according to hypothesis **H4**.

Second case: $\psi(n,m) \leq C(n+m+1)^{\tilde{\beta}}$, $\forall n, m \in \mathbb{N}$. Note that $\psi(\widehat{\mathbf{n}}, b^N) \leq C(\widehat{\mathbf{n}}+b^N+1)^{\tilde{\beta}} \leq C\widehat{\mathbf{n}}^{\tilde{\beta}}$.

$$\begin{aligned} &\widehat{\mathbf{n}}\varphi_{x_\mathbf{j}}(b_\mathbf{n})\rho_\mathbf{n}^N \frac{C2^{N+2}\delta^{-1}\widehat{\mathbf{n}}}{a_{\mathbf{n},\mathbf{j}}\varphi_{x_\mathbf{j}}(b_\mathbf{n})} \widehat{\mathbf{n}}^{\tilde{\beta}} \left(\frac{\widehat{\mathbf{n}}\varphi_{x_\mathbf{j}}(b_\mathbf{n})\rho_\mathbf{n}^N}{\log \widehat{\mathbf{n}}} \right)^{\frac{N-\theta}{2N}} \\ &\qquad \leq \widehat{\mathbf{n}}\rho_\mathbf{n}^N C_N 2^{N+2}\delta^{-1} \frac{1}{\rho_\mathbf{n}^N} \widehat{\mathbf{n}}^{\tilde{\beta}} \left(\frac{\widehat{\mathbf{n}}\varphi_{x_\mathbf{j}}(b_\mathbf{n})\rho_\mathbf{n}^N}{\log \widehat{\mathbf{n}}} \right)^{\frac{N-\theta}{2N}} \\ &\qquad \leq C_N \left[\widehat{\mathbf{n}}\varphi_{x_\mathbf{j}}(b_\mathbf{n})^{\frac{N-\theta}{N(3+2\tilde{\beta})-\theta}} \rho_\mathbf{n}^{\frac{N(N-\theta)}{N(3+2\tilde{\beta})-\theta}} \log \widehat{\mathbf{n}}^{\frac{\theta-N}{N(3+2\tilde{\beta})-\theta}} \right]^{\frac{N(3+2\tilde{\beta})-\theta}{2N}} \end{aligned}$$

which tends to 0 according to hypothesis **H5**. Therefore, in the two cases, we have

$$\mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{I}_\mathbf{n}} K_{1\mathbf{i}} K_{2\mathbf{i}} \leq \frac{a_{\mathbf{n},\mathbf{j}} u}{2} \right] = O\left(\widehat{\mathbf{n}}\rho_\mathbf{n}^N \varphi_{x_\mathbf{j}}(b_\mathbf{n}) \right)^{-1}$$

Consequently, $\|G\|_2 = O\left(\widehat{\mathbf{n}}\rho_\mathbf{n}^N \varphi_{x_\mathbf{j}}(b_\mathbf{n}) \right)^{-1/2}$ ■

Proof of Lemma 4.7

Since $Y_{\mathbf{i}}$ and $r(\cdot)$ are bounded, we have

$$\begin{aligned}
\mathbb{E}^{1/2}[\mathbf{C}] &\leq \mathbb{E}^{1/2}\left[\left|\frac{1}{\widehat{\mathbf{n}}}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}} - r(x_{\mathbf{j}})\right| \mathbf{1}_{\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}}=0\right]}\right] \\
&\leq 2M\mathbb{E}^{1/2}\left[\mathbf{1}_{\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} W_{\mathbf{ni}}=0\right]}\right] \\
&\leq 2M\left(\mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{1\mathbf{i}}K_{2\mathbf{i}} = 0\right]\right)^{1/2} \\
&\leq 2M\left(\mathbb{P}\left[\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{1\mathbf{i}}K_{2\mathbf{i}} \leq \frac{a_{\mathbf{n},\mathbf{j}}u}{2}\right]\right)^{1/2} \\
&= O\left(\frac{1}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{N}\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}})}\right)^{1/2},
\end{aligned}$$

using Lemma 4.6. ■

Proof of Lemma 4.8

Firstly, we deal with $I_{\mathbf{n}}(x_{\mathbf{j}}) = \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \mathbb{E}\left[\left(\frac{1}{a_{\mathbf{n},\mathbf{j}}u}K_{1\mathbf{i}}K_{2\mathbf{i}}\right)^2\right] - \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}}\left(\frac{1}{a_{\mathbf{n},\mathbf{j}}u}\mathbb{E}(K_{1\mathbf{i}}K_{2\mathbf{i}})\right)^2$.

$$\begin{aligned}
\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} \mathbb{E}\left[\left(\frac{1}{a_{\mathbf{n},\mathbf{j}}u}K_{1\mathbf{i}}K_{2\mathbf{i}}\right)^2\right] &\leq C\frac{1}{a_{\mathbf{n},\mathbf{j}}^2u^2}\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_{2\mathbf{i}}^2\mathbb{E}\left[K_{1\mathbf{i}}^2\right] \\
&\leq C\frac{k_{\mathbf{n}}}{a_{\mathbf{n},\mathbf{j}}^2u^2}\mathbb{E}\left[K_{1\mathbf{i}}^2\right] \\
&\leq \frac{C}{k_{\mathbf{n}}\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}})} \\
&= O\left([\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{N}\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}})]^{-1}\right)
\end{aligned}$$

Then, we have $I_{\mathbf{n}}(x_{\mathbf{j}}) = O\left([\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{N}\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}})]^{-1}\right)$. We now treat the term $R_{\mathbf{n}}(x_{\mathbf{j}})$. Since the functions $K_1(\cdot)$ and $K_2(\cdot)$ are bounded, we get by applying Lemma A.1

$$|\mathbb{E}\left[\Lambda_{\mathbf{i}}(x_{\mathbf{j}})\Lambda_{\mathbf{k}}(x_{\mathbf{j}})\right]| \leq C\frac{K_{2\mathbf{i}}K_{2\mathbf{k}}}{a_{\mathbf{n},\mathbf{j}}^2u^2}\psi(1,1)\gamma(\|\mathbf{i}-\mathbf{k}\|).$$

Let $E_{\mathbf{n}}$ be a sequence of real numbers tending to ∞ as $\widehat{\mathbf{n}}\to\infty$. Let $T = \{\mathbf{i},\mathbf{k}\in\mathcal{I}_{\mathbf{n}}, \|\mathbf{i}-\mathbf{k}\| \leq E_{\mathbf{n}}\}$ and denote by T^c the complementary of T. Let $R_{\mathbf{n}}^{(1)} = \sum_{\mathbf{i},\mathbf{k}\in T}\left|\mathbb{E}\left[\Lambda_{\mathbf{i}}(x_{\mathbf{j}})\Lambda_{\mathbf{k}}(x_{\mathbf{j}})\right]\right|$ and $R_{\mathbf{n}}^{(2)} = \sum_{\mathbf{i},\mathbf{k}\in T^c}\left|\mathbb{E}\left[\Lambda_{\mathbf{i}}(x_{\mathbf{j}})\Lambda_{\mathbf{k}}(x_{\mathbf{j}})\right]\right|$. Hence, $R_{\mathbf{n}}(x_{\mathbf{j}}) \leq R_{\mathbf{n}}^{(1)} + R_{\mathbf{n}}^{(2)}$. Moreover, using the same arguments as in the proof of Lemma 4.6, we have

$$I_{\mathbf{n}}(x_{\mathbf{j}}) + R_{\mathbf{n}}(x_{\mathbf{j}}) = O\left(\frac{1}{\widehat{\mathbf{n}}\rho_{\mathbf{n}}^{N}\varphi_{x_{\mathbf{j}}}(b_{\mathbf{n}})}\right).$$

■

Chapter 5

Efficiency in functional nonparametric models with autoregressive errors

Contents

Résumé en français

Dans les chapitres 2, 3 et 4 qui précèdent, nous avons proposé des estimateurs qui tiennent compte de la dépendance spatiale, et qui sont déduits d'une généralisation de travaux en séries temporelles. Cependant, nous n'avons pas spécifié la structure de la dépendance des erreurs. Dans ce chapitre, nous proposons une procédure permettant de tenir compte de la dépendance tout en spécifiant la structure de dépendance des erreurs dans un cadre de séries temporelles fonctionnelles. Cette procédure est une généralisation d'un travail existant dans le cadre réel. Plus précisément, ce chapitre concerne l'étude d'un modèle de régression $Y_t = r(X_t) + u_t$, $t = 1, \ldots, T$, en séries temporelles lorsque les variables explicatives X_t sont fonctionnelles, les variables réponses Y_t sont réelles et les termes d'erreurs u_t sont autocorrélés. Plus précisément, les variables explicatives X_t appartiennent à l'espace fonctionnel $(\mathcal{E}, d)$, muni de la semi-métrique d. La particularité de ce chapitre est de proposer une approche basée sur l'estimateur à noyau qui permet de tenir compte de l'information contenue dans le terme d'erreur. En effet, l'estimateur à noyau classique (1.1), adapté au cadre des données fonctionnelles dans Ferraty et Vieu (2000), ignore la structure de corrélation, induisant une perte d'information. Nous montrons que l'information contenue dans le terme d'erreur u_t permet d'améliorer l'estimation de la fonction de régression. Notons que l'approche proposée existe dans la littérature pour

données à valeurs réelles (voir Xiao et al. (2003)). L'idée principale est de transformer le modèle de régression original de sorte que le terme d'erreur du modèle transformé devienne non corrélé. Dans ce travail, nous supposons que le processus u_t admette une représentation autorégressive d'ordre 1, $u_t = \epsilon_t - a_1\epsilon_{t-1}$ où ϵ_t est un processus i.i.d. de moyenne nulle mais la méthode peut être généralisée à des ordres supérieurs. Le modèle transformé s'écrit

$$\underline{Y}_t = r(X_t) + \epsilon_t$$

où $\underline{Y}_t = Y_t - a_1\left(Y_{t-1} - r(X_{t-1})\right)$ est la série filtrée. Nous proposons d'estimer la fonction de régression $r(\cdot)$ par

$$\overline{r}(x) = \frac{\sum_{t=1}^{T} \underline{Y}_t K_0\left(\frac{d(x,X_t)}{h_0}\right)}{\sum_{s=1}^{T} K_0\left(\frac{d(x,X_s)}{h_0}\right)}, \qquad x \in (\mathcal{E}, d)$$

où K_0 est un noyau et h_0 la fenêtre correspondante. Les propriétés asymptotiques de cet estimateur sont étudiées. Sous certaines conditions, nous montrons la convergence en moyenne d'ordre q dont les vitesses de convergence sont les suivantes

$$\|\overline{r}(x) - r(x)\|_q = O(h_0^{\beta}) + O\left(\left(\frac{1}{T\phi(h_0)}\right)^{1/2}\right).$$

Nous obtenons également la normalité asymptotique de $\overline{r}(x)$.

Cependant, en pratique, $\underline{Y}_t$ est inconnu ce qui ne permet pas de calculer $\overline{r}(x)$. Pour contourner cette difficulté, nous proposons d'approximer $\underline{Y}_t$ par $\widehat{\underline{Y}}_t$ basé sur l'estimation de a_1. La procédure d'approximation est la suivante

1. Obtenir un estimateur consistant de r par la régression de Y_t sur X_t, noté $\widehat{r}$, et calculer les résidus estimés $\widehat{u}_t = Y_t - \widehat{r}(X_t)$
2. Estimer le coefficient a_1 de l'autorégression de $\widehat{u}_t$: $\widehat{u}_t = \widehat{a}_1\widehat{u}_{t-1} + \eta$ avec η un bruit i.i.d.
3. Approximer $\underline{Y}_t$, $t = 2, \ldots, T$, c'est à dire $\widehat{\underline{Y}}_t = Y_t - \widehat{a}_1(Y_{t-1} - \widehat{r}(X_{t-1}))$.

Ainsi, l'estimateur de la fonction de régression r que nous proposons est défini par

$$\widetilde{r}(x) = \frac{\sum_{t=2}^{T} \widehat{\underline{Y}}_t K_1\left(\frac{d(x,X_1)}{h_1}\right)}{\sum_{s=2}^{T} K_1\left(\frac{d(x,X_s)}{h_1}\right)}$$

où K_1 est un noyau et h_1 est la fenêtre correspondante. Après avoir énoncé les hypothèses nécessaires, nous montrons la normalité asymptotique de $\widetilde{r}(x)$.

Le comportement pratique de cet estimateur est également étudié. Nous l'appliquons à des données simulées ainsi qu'à des données de concentration en ozone dans l'air. En ce qui concerne l'application aux données réelles, nous avons adapté l'estimateur de la régression à la prédiction. L'objectif étant de prédire la concentration en ozone à une certaine date non observée à partir du passé. Lorsque le processus des erreurs présente une forte corrélation, nous constatons que notre procédure permet d'améliorer les résultats obtenus avec l'estimateur classique.

L'étude présentée dans ce chapitre est issue d'un travail en collaboration avec Sophie Dabo-Niang (Université Charles de Gaulle) et Serge Guillas (University College London).

5.1 Introduction

The use of functional random variables is spreading in statistical analyses due to the availability of high frequency data and of new mathematical strategies to deal with such statistical objects. The field is known as Functional Data Analysis (FDA). Applications of FDA are growing across fields as diverse as energy studies (Antoniadis et al. (2014)), linguistics (Aston et al. (2010)), atmospheric chemistry (Park et al. (2013)), and human vision (Ogden and Greene (2010)). The functional variables are mainly curves, but surfaces and manifolds are nowadays considered (e.g. Guillas and Lai (2010); Sangalli et al. (2013)). For an introduction to this field as well as illustrations and applications, see Ramsay and Silverman (2005). Besides, Ferraty and Vieu (2006) present nonparametric methods, suited to such functional regression, with a more mathematical flavor. More recently, Cuevas (2014) provides an updated survey of the state of the art in FDA theory.

Among the nonparametric functional regression methods, the kernel estimator is often used to estimate the regression operator. It yields almost sure consistency in the case of an independent sample (Ferraty and Vieu, 2002) or an $\alpha-$mixing sample (Ferraty et al., 2002a,b), but also asymptotic normality in the independent case (Ferraty et al., 2007) with exact computation of all the constants for its precise use in practice. Masry (2005) established the asymptotic normality of the nonparametric regression estimator for strongly mixing processes albeit with abstract expressions of the constants so this is more challenging to use in practice. Delsol (2007a, 2009) generalized the results of Ferraty et al. (2007) to the case of an α-mixing dataset.

In this chapter, we consider the regression of a scalar random variable onto a functional random variable. The estimation of the regression function is tackled by means of a nonparametric kernel approach. Our regression model is:

$$Y_t = r(X_t) + u_t, \qquad t = 1, \ldots, T, \tag{5.1}$$

where the explanatory variable is functional (that is, X_t takes values in some possibly infinite-dimensional space). Moreover, the stationary residual process u_t is autocorrelated and independent of X_t. We also assume some smoothness conditions on the unknown functional operator $r(\cdot)$. Although, for the kernel methods proposed in the literature, it is generally better to ignore the correlation structure entirely (the so-called "working independence estimator", e.g. Ruckstuhl et al. (2000), Lin and Carroll (2000)), we show here that taking into account the autocorrelation of the error process helps improve the estimation of the regression function.

We extend the kernel-based procedure proposed by Xiao et al. (2003) for estimating $r(x)$ in the time series regression model for multivariate explanatory variables x to a functional setting. Xiao et al. (2003) showed that their procedure is more efficient than the conventional local polynomial method. The main idea is to transform the original regression model, so that this transformed regression has a residual term that is uncorrelated. This transformation depends on the function $r(\cdot)$ and on the parameters of the autoregressive representation of u, since the regression function is nonlinear. The error correlation structure is assumed to be an autoregressive process of order 1 for simplicity (but could be extended to higher orders). Firstly, the parameters of the autoregressive representation are estimated. In a second step, a transformation $\widehat{\underline{Y}}_t$ of the dependent variable Y_t is constructed by plugging in the estimated autocorrelation parameter. Finally, the estimation of r is carried out on this filtered series $\widehat{\underline{Y}}_t$.

The remainder of this chapter is organized as follows. In Section 5.2, we introduce the estimation method as well as some assumptions. We then provide asymptotic results for the

estimator proposed. Section 5.3 is devoted to a simulation case study and an illustration of our method for ozone levels over the US. The conclusion is done in Section 5.4 while the proofs are given in the Appendix.

5.2 Assumptions and main results

Suppose that we have a sample $\{(X_1, Y_1), \ldots, (X_T, Y_T)\}$, where $X_t \in (\mathcal{E}, d)$, $t = 1, \ldots, T$, is a random variable taking its values in a semi-metric space $(\mathcal{E}, d)$ of infinite dimension and $Y_t \in \mathbb{R}$ is the response from the nonparametric regression (5.1). We assume that the residual process u_t is stationary, has mean 0 with autocovariance function γ_u and has the following invertible linear process representation (with bounded coefficients):

$$u_t = c_0\epsilon_t + c_1\epsilon_{t-1},$$

where the ϵ_t are i.i.d. with mean 0, variance σ_ϵ^2 and $\mathbb{E}[|\epsilon_t|] < \infty$ but which could be generalized to the following form

$$u_t = \sum_{j=0}^{\infty} c_j\epsilon_{t-j}$$

Without loss of generality, let $c_0 = 1$. The coefficient c_1 and the regression function $r(\cdot)$ are unknown, except for the fact that $r(\cdot)$ is a smooth function.

5.2.1 Known autocorrelation parameter

In this section, we motivate the construction of the final estimate by considering the unrealistic case where the autocorrelation is known. Let $c(L) = \sum_{j=0}^{\infty} c_j L^j$ where L is the usual lag operator. Inverting $c(L)$, we obtain an autoregressive representation of u_t. Let $c(L)^{-1} = a(L) = a_0 - \sum_{j=1}^{\infty} a_j L^j$ be the inverse operator, so we have $a(L)u_t = \epsilon_t$. Here, we consider a truncated version of $a(L)$ at order 1, that is $a(L) = a_0 - a_1 L$. Applying $a(L)$ to the regression in Equation (5.1), we obtain

$$a(L)Y_t = a(L)r(X_t) + \epsilon_t,$$

the error term in this transformed model is now uncorrelated. We write

$$\underline{Y}_t = r(X_t) + \epsilon_t, \tag{5.2}$$

where $\underline{Y}_t$ is the filtered series

$$\underline{Y}_t = Y_t - a_1 \left(Y_{t-1} - r(X_{t-1})\right).$$

If $\underline{Y}_t$ were known then a nonparametric kernel regression of $\underline{Y}_t$ on X_t would be more efficient than the conventional kernel estimation. In this work, we employ a Nadaraya-Watson estimator as introduced in Ferraty and Vieu (2004), Masry (2005), Dabo-Niang and Rhomari (2009) where for any sample $\{V_t, X_t\}$, the estimation of the regression of V_t onto X_t is

$$\frac{\sum_{t=1}^{T} V_t K\left(\frac{d(x,X_t)}{h}\right)}{\sum_{s=1}^{T} K\left(\frac{d(x,X_s)}{h}\right)}, \qquad x \in \mathcal{E}$$

where $K(\cdot)$ is a function over $[0, +\infty[$ called kernel, $h > 0$ is the bandwidth parameter and $d(\cdot, \cdot)$ is a semi-metric. For $x \in \mathcal{E}$ fixed, let $\widehat{r}(x)$ be the corresponding estimator with

$V_t = Y_t$ and let $\overline{r}(x)$ be the corresponding estimator when $V_t = \underline{Y}_t$. Let K_0 and K_1 two kernels over $[0; +\infty[$, h_0 and h_1 the two corresponding bandwidths. We write

$$\overline{r}(x) = \frac{\dfrac{1}{T\mathbb{E}\left[K_0\left(\frac{d(x,X_1)}{h_0}\right)\right]}\displaystyle\sum_{t=1}^{T}\underline{Y}_t K_0\left(\frac{d(x,X_t)}{h_0}\right)}{\dfrac{1}{T\mathbb{E}\left[K_0\left(\frac{d(x,X_1)}{h_0}\right)\right]}\displaystyle\sum_{s=1}^{T}K_0\left(\frac{d(x,X_s)}{h_0}\right)} = \frac{\overline{r}_2(x)}{\overline{r}_1(x)}$$

The two following theorems provide asymptotic results for the estimator $\overline{r}(x)$, and the proofs are given in the Appendix. More precisely, the first deals with the mean of order q convergence of this estimate while the second is about its asymptotic normality. The necessary assumptions for having such results will be written down in Section 5.2.2.

Theorem 5.1. Assume **H1**-**H4** (or, more precisely **H1**(1), **H2**(1), **H3**(1), **H4**, where in **H4**, $\delta > \max\{q/2 - 1, 1 - 2/\nu\}$, $q > 1$) and assume that Y_t is bounded. Let $T\phi(h_0) \to \infty$ as $T \to \infty$. Then, $\overline{r}(x)$ converges in mean of order q to $r(x)$, more precisely, we have

$$\|\overline{r}(x) - r(x)\|_q = O(h_0^{\beta}) + O\left(\left(\frac{1}{T\phi(h_0)}\right)^{1/2}\right).$$

When the response variable Y_t is unbounded, one can establish a convergence in probability of $\overline{r}$, avoiding the restriction of boundedness of the response (even though the bound assumed in the previous result can be arbitrarily large) using the result of Corollary 1 of Masry (2005) which states that:

Proposition 5.2. Under assumptions **H1**-**H5** (or, more precisely **H1**, **H2**, **H3**(1), **H3**(2), **H4** and **H5**)

$$\lim_{T\to\infty}\overline{r}(x) = r(x) \qquad \text{in probability.}$$

The proof of this result is given in Masry (2005) (Corollary 1).
We now write

$$B_T(x) = \mathbb{E}[\overline{r}(x)] - r(x)$$

and let $\Delta_t^{(i)}(x) = K_i\left(\dfrac{d(x,X_t)}{h_i}\right)$, $Z_t^{(i)}(x) = [\underline{Y}_t - r(x)]\Delta_t^{(i)}(x) - E\left[(\underline{Y}_t - r(x))\Delta_t^{(i)}(x)\right]$, for $i = 0, 1$ (see below). In the following, $\xrightarrow{d}$ denotes the convergence in distribution.

Theorem 5.3. Under assumptions **H1**-**H5** (or more precisely **H1**, **H2**, **H3**(1), **H3**(2), **H4**, **H5**)

$$(T\phi(h_0))^{1/2}\left[\overline{r}(x) - r(x) - B_T(x)\right] \xrightarrow{d} \mathcal{N}(0, \sigma^2(x))$$

with $\sigma^2(x) = \dfrac{C_2 g_2(x)}{C_1^2 f_1(x)} = \lim\limits_{T\to\infty}\dfrac{\phi(h_0)Var(Z_T^{(0)}(x))}{\mathbb{E}^2(\Delta_1^{(0)}(x))}, x \in (\mathcal{E}, d)$ whenever $f_1(x) > 0$.

Theorem 5.3 comes from Theorem 5 in Masry (2005). The functions $g_2(x)$, $f_1(x)$ and $\phi(h_0)$ are given in assumptions **H1** and **H3**.

Remark 5.4. As stated in Masry (2005), if in addition to the assumptions of Therorem 5.3 we have $T\phi(h_0)h_0^{2\beta} \to 0$ (it is a stronger assumption on the bandwidth parameter) then one can remove the bias term from Therorem 5.3 that is $(T\phi(h_0))^{1/2}\left[\overline{r}(x) - r(x)\right] \xrightarrow{d} \mathcal{N}(0, \sigma^2(x))$.

5.2.2 Unknown correlation structure, assumptions and main results

In practice, the coefficient c_1 is unknown and therefore $\underline{Y}_t$ is not computable, so the regression $\underline{Y}_t = r(X_t) + \epsilon_t$ and $\overline{r}(x)$ are unworkable. A feasible estimator is obtained by replacing the left side of Equation (5.2) by an approximation of $\underline{Y}_t$ based on the estimate of a_1. The proposed estimation procedure extends Xiao et al. (2003):

1. Obtain a preliminary consistent estimate of r by the regression of Y_t on X_t with corresponding kernel K_0 and bandwidth h_0 assuming i.i.d. errors. Denote the preliminary estimate as $\widehat{r}(X_t)$ and calculate the estimated residuals
$$\widehat{u}_t = Y_t - \widehat{r}(X_t).$$
2. Conduct an estimation of the AR(1) coefficients in the autoregression of $\widehat{u}_t$: $\widehat{u}_t = \widehat{a}_1 \widehat{u}_{t-1} + \eta$, where η is an i.i.d. noise.
3. Construct an approximation of $\underline{Y}_t$, $t = 2, \ldots, T$ that is $\widehat{\underline{Y}}_t = Y_t - \widehat{a}_1 \left(Y_{t-1} - \widehat{r}(X_{t-1})\right)$. The proposed estimator of $r(x)$ is then obtained from the regression of $\widehat{\underline{Y}}_t$ on X_t with corresponding kernel K_1 and bandwidth h_1, resulting in the estimator $\widetilde{r}(x)$:
$$\widetilde{r}(x) = \frac{\dfrac{1}{T\mathbb{E}\left[K_1\left(\frac{d(x,X_1)}{h_1}\right)\right]} \displaystyle\sum_{t=2}^{T} \widehat{\underline{Y}}_t K_1\left(\frac{d(x,X_1)}{h_1}\right)}{\dfrac{1}{T\mathbb{E}\left[K_1\left(\frac{d(x,X_1)}{h_1}\right)\right]} \displaystyle\sum_{s=2}^{T} K_1\left(\frac{d(x,X_s)}{h_1}\right)}$$

Note that one can iterate this process, in case the initial estimate of the autocorrelation is not accurate enough as the bias in this estimate will propagate to the filtered series and hence to the estimation of $r(x)$. In our numerical studies, we present both the initial estimate and the estimate based on an additional iteration of the steps above.

Let us now explain in details the theoretical set-up that enables us to prove the asymptotic results of this work. We first assume that the error process $\{u_t\}$ is independent of the process $\{X_t\}$ and that $\mathbb{E}[\epsilon_t | X_t] = 0$. Moreover, we consider that $\{X_t\}$ is an α-mixing process, the most general case of weakly dependent variables. In other words, we assume that when $|t - s|$ tends to infinity, X_t and X_s become roughly independent. Let $\mathcal{F}_a^b$ be the σ-algebra of events generated by the random variables $\{X_t : a \leq t \leq b\}$. Recall that a stationary process $\{X_t\}$ is called strongly mixing (Rosenblatt (1956a)) if
$$\sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_k^\infty} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| = \alpha(k) \underset{k \to \infty}{\longrightarrow} 0.$$

Our assumptions are listed below:

H1 *(Smoothness)*

(1) $r(\cdot)$ is a bounded Lipschitz function: $|r(u) - r(v)| \leq c_3 d(u,v)^\beta$ for all $u, v \in (\mathcal{E}, d)$ for some $\beta > 0$.

(2) Let $g_2(u) = Var[Y_t | X_t = u]$, $u \in (\mathcal{E}, d)$.
$g_2(u)$ is independent of t and is continuous in some neighborhood of x
$$\sup_{\{u : d(x,u) \leq h\}} |g_2(u) - g_2(x)| = o(1) \quad \text{as } h \to 0$$
Assume $\mathbb{E}|Y_t|^\nu < \infty$ and $\mathbb{E}[|\epsilon_t|^\nu] < \infty$, for some $\nu > 2$. Assume
$$g_\nu(u) = \mathbb{E}[|Y_t - r(x)|^\nu | X_t = u], u \in (\mathcal{E}, d)$$
is continuous in some neighborhood of x.

(3) Define

$$g(u,v;x) = \mathbb{E}[(Y_t - r(x))(Y_s - r(x))|X_t = u, X_s = v], \quad t \neq s \text{ and } u,v \in (\mathcal{E},d)$$

Assume that $g(u,v;x)$ does not depend on t,s and is continuous in some neighborhood of (x,x).

H2 *(Kernel)* The kernels K_i, $i = 0$ or 1, are symmetric nonnegative bounded kernels with compact support $[0,1]$ satisfying

(1) $\int K_i(u)du = 1$ and $c_{1,i}\mathbf{1}_{[0,1]} < K_i < c_{2,i}\mathbf{1}_{[0,1]}$, $c_{1,i}$ and $c_{2,i}$ are two finite constants.

(2) For $j = 1, 2$, we have $I_j(h) \to C_j$ as $h \to 0$, for some positive constant C_j, with

$$I_j(h) = \frac{1}{\phi(h)/h}\int_0^1 K_i^j(u)\phi'(hv)dv$$

where $\phi(\cdot)$ is defined below.

Let $\mathcal{B}(x,h)$ be a ball centered at $x \in (\mathcal{E},d)$ with radius h and let f_k, $k = 1$, 2 and 3, be finite nonnegative functionals. Finally, we introduce the following notations, where $F_x(h)$ corresponds to the well-known notion of small ball probabilities (see e.g. Dabo-Niang (2004), Ferraty and Vieu (2006)):

$$\begin{aligned}
F_x(h) &= \mathbb{P}[X_t \in \mathcal{B}(x,h)] && := \mathbb{P}[d(X_t,x) \leq h] \\
F_{x,x}^{s,t}(h) &= \mathbb{P}[(X_t,X_s) \in \mathcal{B}(x,h) \times \mathcal{B}(x,h)] && := \mathbb{P}[d(X_t,x) \leq h, d(X_s,x) \leq h] \\
F_{x,y}^{s,t}(h) &= \mathbb{P}[(X_t,X_s) \in \mathcal{B}(x,h) \times \mathcal{B}(y,h)] && := \mathbb{P}[d(X_t,x) \leq h, d(X_s,y) \leq h]
\end{aligned}$$

H3 *(Distributions)*

(1) $F_x(h) = \phi(h)f_1(x)$ as $h \to 0$, where $\phi(0) = 0$ and $\phi(h)$ is absolutely continuous in a neighborhood of the origin and $f_1(X_t)$ is uniformly bounded and bounded away from zero.

(2) $\sup_{t\neq s} F_{x,x}^{s,t}(h) \leq \psi_1(h)f_2(x)$ as $h \to 0$, where $\psi_1(h) \to 0$ as $h \to 0$ and $f_2(X_t) < \infty$ is uniformly bounded and bounded away from zero.
Assume that the ratio $\psi_1(h)/\phi^2(h)$ is bounded. It is also supposed that $\exists\zeta_1 \in (0,1)$, $0 < F_{x,x}(h) = O(\phi(h)^{1+\zeta_1})$.

(3) $\sup_{t\neq s} F_{x,y}^{s,t}(h) \leq \psi_2(h)f_3(x,y)$ as $h \to 0$, where $\psi_2(h) \to 0$ as $h \to 0$ and $f_3(X_t,X_s) < \infty$ is uniformly bounded and bounded away from zero.
Assume that the ratio $\psi_2(h)/\phi^2(h)$ is bounded.

H4 *(Mixing)*

$$\sum_{l=1}^{\infty} l^{\delta}[\alpha(l)]^{1-2/\nu} < \infty$$

for some $\nu > 2$ and $\delta > 1 - 2/\nu$. Note that ν is the order of the moment in **H1**(2).

H5 Let $h_i \to 0$, $h_0/h_1 \to 0$ and $\frac{\log T}{T^{1/2}\phi(h_0)} \to 0$ as $T \to \infty$. Let $\{v_T\}$ be a sequence of positive integers satisfying $v_T \to \infty$ such that $v_T = o((T\phi(h_0))^{1/2})$ and $(T/\phi(h_0))^{1/2}\alpha(v_T) \to 0$, $Th_0^{2\beta} \to 0$ as $T \to \infty$.

Remark 5.5.

- Hypotheses **H1**(1) is a mild smoothness assumption for kernel functions in nonparametric estimation whereas hypotheses **H1**(2) and **H1**(3) are continuity assumptions on certain second-order moments.

- Hypothesis **H2**(1) on the kernel is standard. From hypothesis **H2**(2), if the kernel K_i satisfies $0 < c_1 \leq K_i(t) \leq c_2 < \infty$, then $c_1 \leq I_j(h) \leq c_2$. In fact, this assumption yields an expression of the asymptotic variance (rather than upper and lower bounds).
- Hypotheses of type **H3** were proposed in Masry (2005) and have been motivated by the work of Gasser et al. (1998). These hypotheses are linked to the volume of an n-ball. When $X \in \mathbb{R}^d$, $f_1(x)$ refers to the probability density of the random variable X and $\phi(h)$ is the volume of the unit ball in $\mathbb{R}^d$. Assumptions **H3**(2) and **H3**(3) concern the behavior of joint distribution.
- Hypothesis **H4** is a standard assumption on the decay of the strongly mixing coefficient $\alpha(l)$ and hypothesis **H5** concerns the rate of the decay of the mixing coefficient.

The following theorem gives the asymptotic normality of the estimator $\widetilde{r}(x)$ based on the transformation of the dependent variable.

Theorem 5.6. Under assumptions **H1-H5**, we have

$$(T\phi(h_1))^{1/2}[\widetilde{r}(x) - r(x) - B_T(x)] \xrightarrow{d} \mathcal{N}(0, \sigma^2(x))$$

with $\sigma^2(x) = \dfrac{C_2 g_2(x)}{C_1^2 f_1(x)} = \lim\limits_{T\to\infty} \dfrac{\phi(h_1) Var(Z_T^{(1)}(x))}{\mathbb{E}^2(\Delta_1^{(1)}(x))}$, $x \in (\mathcal{E}, d)$ whenever $f_1(x) > 0$.

The following theorem gives a consistency result of the estimator $\widetilde{r}(x)$.

Theorem 5.7. Under assumptions **H1-H5**,

$$\lim_{T\to\infty} \widetilde{r}(x) = r(x) \quad \text{in probability.}$$

Remark 5.8. One can establish a convergence in probability of $\widetilde{r}(x)$ with rate (for instance assuming for simplicity the boundedness of the response, even though the bound can be arbitrarily large) and state that:

$$|\widetilde{r}(x) - r(x)| = O_p(h_0^\beta) + o_p\left(\sqrt{\frac{1}{T\phi(h_1)}}\right)$$

under conditions of Theorem 5.1 in addition to **H3**(2), **H3**(3) and **H5**.

The proofs of these theorems are postponed in the Appendix section.

5.3 Numerical results

5.3.1 Simulation study

We investigate the proposed estimator on simulated data. The functional observations X_t (with $t = 1, \ldots, T$) are defined by $X_t(w) = 1 + 10e_{0,t} + 3e_{1,t}w^2 + 4e_{2,t}(1-w)^3$, $w \in [0, 1]$ where $e_{0,t}$, $e_{1,t}$ and $e_{2,t}$ are i.i.d. $\mathcal{N}(0, 1)$. Example of some simulated curves are illustrated on Figure 5.1. We take $r(x) = \sqrt{|0.5 \int_0^1 x^4 dx|}$. The error process u_t is an $AR(1)$ process, that is $u_t = \epsilon_t + \rho\epsilon_{t-1}$. Various values of ρ are considered. The number of replications is 200. We report the relative efficiency (denoted as RE) calculated based on the ratio of squared errors. Table 5.1 describes summary statistics of the relative efficiency for $T = 200$ whereas Table 5.2 gives the average of the relative efficiency for different values of T. $RE1$

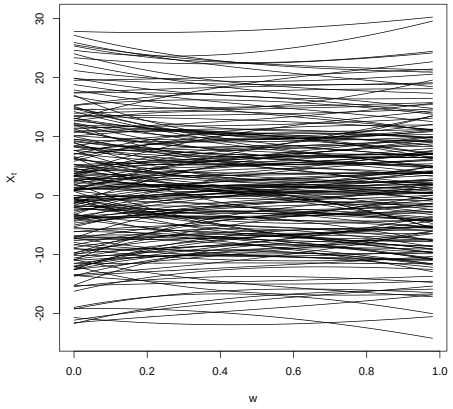


Figure 5.1 – Example of 200 simulated curves X_t

reports the relative efficiency of the proposed efficient estimator $\widetilde{r}(x)$ over the conventional estimator $\widehat{r}(x)$, and $RE2$ concerns the relative efficiency of the iterated estimator, over $\widehat{r}(x)$. We did not implement the efficient estimator of Xiao et al. (2003) as we only consider here for simplicity the case of one lag, but the efficient estimator could be used in our context with larger lags than one. Instead we here report results about the iterated estimates. The semi-metric $d(\cdot,\cdot)$ for computing proximities between curves X_t plays a major role and depends on the specified statistical problem and dataset. After trying some semi-metrics which can select most of the pertinent information of the curves, we choose $d(\cdot,\cdot)$ inside the family of principal component semi-metrics (see Ferraty and Vieu (2006)) which is defined by $d_q^{PCA}(X_i,X) = \sqrt{\sum_{k=1}^{q}\left(\int[X_i(t)-X(t)]v_k(t)\right)^2}$ where $v_1, v_2, \ldots$ are the orthonormal eigenfunctions of the covariance operator and q is a tuning parameter whose the value is 4 in this simulation study. Regarding the implementation of the estimators, we use the quadratic kernel (Epanechnikov) (defined by $K(x) = \frac{3}{4}\left(1-x^2\right)\mathbf{1}_{[-1;+1]}(x)$). Another choice to make is the bandwidth parameters. It is well known that the performance of the kernel estimate depends on the choice of the window parameter. The bound in Remark 5.8 allows us to choose the window parameters that minimize this bound. This choice of the bandwidths leads to be optimal in the finite dimensional case. In practice, a useful bandwidth choice method is cross-validation as follows.

1. We consider the preliminary estimate $\widehat{r}$ of r by the regression of Y_t on X_t with quadratic kernel K, the semi-metric d_4^{PCA} and data driven bandwidth h_0^{opt} assuming i.i.d. errors (see Ferraty and Vieu (2006) for more details):

$$h_0^{opt} = \arg\min_h \sum_{t=1}^{T}(Y_t - \widehat{r}_{-t}(X_t))^2$$

where

$$\widehat{r}_{-t}(x) = \frac{\sum_{u=1,u\neq t}^{T} Y_u K\left(\frac{d(x,X_u)}{h_0}\right)}{\sum_{s=1,s\neq t}^{T} K\left(\frac{d(x,X_s)}{h_0}\right)}$$

We calculate the estimated residuals $\widehat{u}_t = Y_t - \widehat{r}(X_t)$.

2. We conduct an estimation of the AR(1) coefficients in the autoregression of $\widehat{u}_t$: $\widehat{u}_t = \widehat{a}_1\widehat{u}_{t-1} + \eta$, as in Section 5.2.2. We construct $\widehat{\underline{Y}}_t = Y_t - \widehat{a}_1\left(Y_{t-1} - \widehat{r}(X_{t-1})\right)$, $t = 2,\ldots,T$ and the estimate $\widetilde{r}(x)$ from the regression of $\widehat{\underline{Y}}_t$ on X_t with quadratic

kernel K, the semi-metric d_4^{PCA} and optimal data driven bandwidth h_1^{opt} in the same way as above, replacing $\widehat{r}(x)$ by $\widetilde{r}(x)$ resulting in:

$$h_1^{opt} = arg \min_h \sum_{t=2}^{T} (Y_t - \widetilde{r}_{-t}(X_t))^2.$$

The results in Table 5.1 show that there is great variability in the improvements across replications. The inter-quartile ranges of the relative efficiencies are nevertheless tight: typically within 0.1-0.2, except when the improvements are large (e.g. for $\rho = 0.9$). The mean improvements for the estimator is always positive except when $\rho = 0.1$, a very small level of autocorrelation. The iterated estimator is much less efficient than the initial estimator. It seems that the additional steps are adding several layers of noise in the procedure and therefore degrade the estimation. Table 5.2 allows us to see the effect of sample size on the mean relative efficiency. It seems that such benefit is stronger whenever the autocorrelation is higher (as expected to be able to capture it properly).

ρ	RE	Min	Q_1	Med	$Mean$	Q_3	Max
0.99	1	0.1114	0.9162	0.9788	0.9046	1.0010	1.1290
	2	0.1392	0.9279	0.9806	0.9398	1.0350	1.6170
0.95	1	0.1359	0.6961	0.8795	0.8149	0.9717	1.1530
	2	0.2483	0.7643	0.9453	0.9177	1.0850	1.7290
0.90	1	0.1436	0.5672	0.7786	**0.7276**	0.9132	1.5420
	2	0.1182	0.7430	0.8951	0.9052	1.0790	2.5070
0.80	1	0.2637	0.6776	0.7976	0.8050	0.9366	1.5400
	2	0.3677	0.8257	0.9993	1.0400	1.2040	2.5230
0.60	1	0.3909	0.7671	0.8894	0.8929	1.0160	1.7520
	2	0.5705	0.9222	1.0900	1.1440	1.2790	3.7060
0.50	1	0.4097	0.8226	0.9410	0.9340	1.0290	1.4030
	2	0.6234	0.9793	1.1050	1.1310	1.2430	2.8680
0.25	1	0.6937	0.9467	0.9959	0.9979	1.0410	1.3910
	2	0.6331	0.9716	1.0250	1.0520	1.1000	1.7120
0.10	1	0.8446	0.9820	1.0030	1.0120	1.0280	1.5180
	2	0.8606	0.9795	1.0110	1.0330	1.0570	1.5310

Table 5.1 – Elementary statistics of the relative efficiency for T=200

T	100		200		500	
ρ	$RE1$	$RE2$	$RE1$	$RE2$	$RE1$	$RE2$
0.99	0.922	0.975	0.905	0.940	0.848	0.862
0.95	0.836	0.927	0.815	0.918	0.779	0.874
0.90	0.831	1.015	0.728	0.905	0.739	0.900
0.80	0.837	1.058	0.805	1.040	0.775	1.008
0.60	0.925	1.189	0.893	1.144	0.884	1.116
0.50	0.968	1.193	0.934	1.131	0.936	1.144
0.25	1.016	1.084	0.998	1.052	0.993	1.071
0.10	1.011	1.041	1.012	1.033	1.007	1.025

Table 5.2 – Mean of the relative efficiency for $T = 100$, 200 and 500

The efficiencies for functional data seem better than for univariate time series (Xiao et al. (2003)), although Xiao et al. (2003) considered an ARMA(1,1) case - and an AR(2) pre-whitening - in their simulations that is more challenging (but in dimension one, not in infinite dimension as here). Indeed in Xiao et al. (2003), the relative improvement was never below 0.85. Here, we can reach average reductions below 0.75 for high correlation and long enough time series to capture this high level of correlation accurately. According to Ferraty

and Vieu (2006), the curse of dimensionality, a well-known concept in nonparametric inference, does not affect functional data with high correlation. This, combined with an appropriate choice of the semi-metric, can explain the fact that the efficiencies seem better in functional context than univariate on. One illustration is given on Figure 5.2: for one replication, considering $T = 200$ and a value of $\rho = 0.9$, one can see, first, a zoom of the curves and then all the series. The black curve displays the true function $r(\cdot)$, the dotted blue curve corresponds to the standard kernel estimation whereas the red and green curves correspond to the proposed estimator with one or two iterations respectively. Note that in this case the common estimate of the curve is far from the true curve. On the contrary, the curves obtained considering our methodology not only have the same shape as the true curve but are very close to the truth. In this case, the information of the autocorrelation function of the error process clearly improves the quality of the regression estimation. However, when the autoregressive parameter is smaller, as expected, our methodology does not improve the results obtained through the standard kernel procedure that does not account for correlation in the errors. For example, Figure 5.3 shows the curves obtained considering $\rho = 0.25$ for one replication. We cannot see large differences between the displayed curves. The three estimation curves are closed to the curve representing the true function.

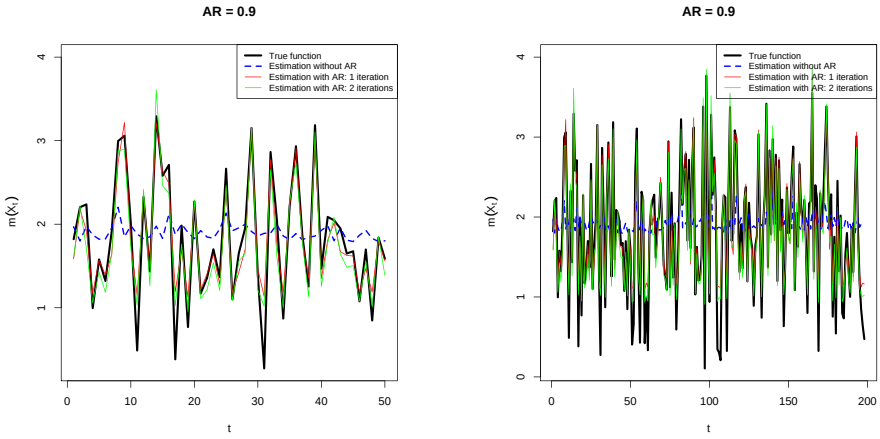


Figure 5.2 – Illustrations of the results for $T = 200$ and $\rho = 0.9$
The black curve displays the true function $r(\cdot)$, the dotted blue curve corresponds to the standard kernel estimation whereas the red and green curves correspond to the proposed estimator with one or two iterations respectively. On the left: a zoom of the curves. On the right: all the series.

5.3.2 Real data application

Here, we illustrate our methodology for the ozone concentration forecasting problem and compare our predictions with the ones obtained using the classical kernel regression model for functional data. The goal is to forecast ground-level ozone concentrations using observations from monitoring stations within the south-eastern US region, over a span of 3 months in the summer of 2005. These forecasts may contribute to better public health: for example, hourly forecasts made one day ahead of this harmful pollutant allow people avoid outdoor activities likely to damage their health (Ettinger et al. (2012)).

We are given the ozone concentration for different stations for every hour from June 2 to August 31, 2005 (that is 91 days). Since some of the stations had missing values,

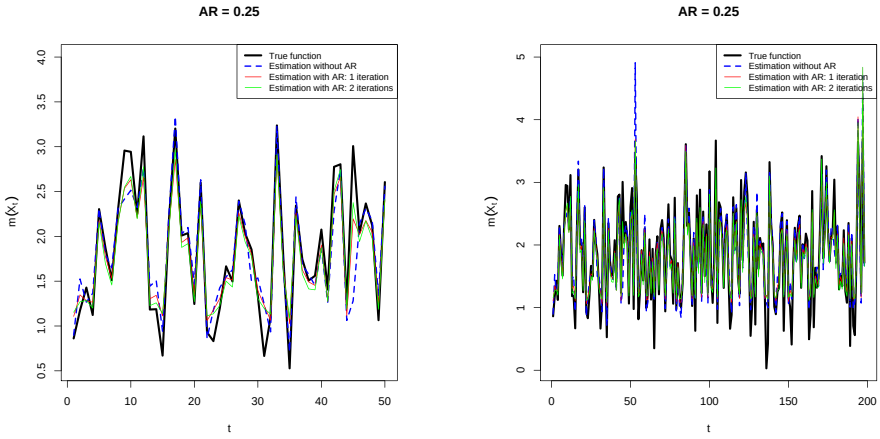


Figure 5.3 – Illustrations of the results for $T = 200$ and $\rho = 0.25$
The black curve displays the true function $r(\cdot)$, the dotted blue curve corresponds to the standard kernel estimation whereas the red and green curves correspond to the proposed estimator with one or two iterations respectively. On the left: a zoom of the curves. On the right: all the series.

we use linear interpolation to estimate the missing values. We are interested in 1-day ozone forecasting (specifically, r-hours ahead ozone forecasting, for $r = 1, \ldots, 24$). We denote the ozone concentration at time t by $Z(t)$ where t refers to the day and the hour of observation. We suppose that $Z(t)$ is observed for $t \in [1; 2160)$ (24 hours $\times$ 90 days) and we are interested in predicting $Z(2160 + r)$ for $r = 1, 2, \ldots, 24$. In order to apply the functional methodology, we cut the original time series into a set of daily functional data. Here, we have decided to predict future ozone concentration by using the concentration data for the whole last day (24 hours). Then, in order to illustrate our purpose, we will not use the 91th day and we will predict it by means of the data corresponding to the 90 previous ones. Then, as presented in Ferraty and Vieu (2006), for fixed r, the data will be reorganized into a functional explanatory sample $\{X_i, i = 1, \ldots, 89\}$ which will be loaded in the following 89×24 matrix:

$Z(1)$	$Z(2)$	$\cdots$	$Z(24)$
$Z(25)$	$Z(26)$	$\cdots$	$Z(48)$
$\vdots$			
$Z(2113)$	$Z(2114)$	$\cdots$	$Z(2136)$

and a response real sample $\{Y_i^{(r)}, i = 1, \ldots, 89\}$, which will be loaded in the following 89-dimensional vector:

$Z(24 + r)$	$Z(48 + r)$	$\cdots$	$Z(2136 + r)$

For a fixed horizon r, we will predict the value of $Z(2160 + r)$. In the following, the predictions have been achieved for any value of $r = 1, 2, \ldots, 24$. Note that in our procedure several parameters need to be selected. For the kernel, we use the quadratic one. Cross-validation methods, expressed in terms of k-nearest neighbours, are used for (local) smoothing parameter selection (see Chapter 7 in Ferraty and Vieu (2006)). Moreover, we use a semi-metric based on the first functional principal components of the data curves. Finally, we proceed in the following way:

1. Select the horizon prediction r and organize the data as it is explained previously;
2. Predict $Y_{90}^{(r)}$, at fixed horizon r, by the classical kernel regression approach with a local choice of the number k of neighbors:

$$\widehat{Y}_{90}^{(r)} = \widehat{r}(X_{90}) = \frac{\sum_{i=1}^{89} Y_i^{(r)} K\left(\frac{d(X_i,X_{90})}{h_{k_{opt}(X_{i_0})}}\right)}{\sum_{i=1}^{89} K\left(\frac{d(X_i,X_{90})}{h_{k_{opt}(X_{i_0})}}\right)} \tag{5.3}$$

where $i_0 = \arg\min_{i=1,\ldots,89} d(X_{90}, X_i)$ and $h_{k_{opt}(X_{i_0})}$ is the bandwidth corresponding to the optimal number of neighbors at X_{i_0} obtained by

$$k_{opt}(X_{i_0}) = \arg\min_k \left| Y_{i_0} - \frac{\sum_{i=1, i\neq i_0}^{n} Y_i K\left(\frac{d(X_i,X_{i_0})}{h_{k(X_{i_0})}}\right)}{\sum_{i=1, i\neq i_0}^{n} K\left(\frac{d(X_i,X_{i_0})}{h_{k(X_{i_0})}}\right)} \right| \tag{5.4}$$

3. At step (2), during the learning step, the 89 response variables are estimated, denoted e_i, $i = 1, \ldots, 89$. Then, we construct the residual terms $\{\widehat{u}_i\}$, where for $i = 1, \ldots, 89$, $\widehat{u}_i = \widehat{Y}_i - e_i$. We estimate the $AR(1)$ coefficient, a_1, in the autoregression of $\widehat{u}_t$.
4. Construct $\underline{\widehat{Y}}_i$, $i = 2, \ldots, 89$, as explained in Section 5.2.2, and do Step (2)

$$\widetilde{Y}_{90}^{(r)} = \widetilde{r}(X_{90}) = \frac{\sum_{i=2}^{89} \underline{\widehat{Y}}_i^{(r)} K\left(\frac{d(X_i,X_{90})}{h_{k_{opt}(X_{i_0})}}\right)}{\sum_{i=2}^{89} K\left(\frac{d(X_i,X_{90})}{h_{k_{opt}(X_{i_0})}}\right)} \tag{5.5}$$

We apply the previous procedure on Station 17 to predict ozone on August 31st, the 91st day. For this station, it is better to consider the squared root of the data in order to keep distributions roughly normal. The series of square root of observations are represented in Figure 5.4. On the middle panel of this figure, the square roots of daily curves are plotted and the red curve represents the curve we want to forecast. The results obtained at Step (2) (by the classical kernel method) are displayed in blue on the right panel of Figure 5.4 whereas those of Step (4) (from our procedure presented in Section 5.2.2) are displayed in red. From this figure, one can observe that our method improves upon the results obtained with classical method, in particular for the second half of the day. In fact, the estimated coefficients in the autoregression of $\widehat{u}_i$ are higher for that part of the day, see Table 5.3. In addition, we compute the mean squared errors (MSE) to compare the results obtained by the different methods. For Station 17, the MSE from the classical approach is 4.4 whereas with our approach it is reduced to 3.3. Again we note that the fact of taking into account the autocorrelation in the error process allows to improves ozone forecasting.

r	1	2	3	4	5	6	7	8	9	10	11	12
$\widehat{a}_1$	0.11	0.04	0.03	−0.05	−0.15	−0.11	0.03	0.12	−0.02	0.01	−0.02	0.13
r	13	14	15	16	17	18	19	20	21	22	23	24
$\widehat{a}_1$	0.22	0.20	0.23	0.35	0.28	0.33	0.30	0.20	0.18	0.12	0.13	0.16

Table 5.3 – Station 17: for horizon prediction r, estimated autoregressive coefficient $\widehat{a}_1$

Now, we present results from three other stations and/or situations. Firstly, we consider the situation where we want to predict ozone concentrations on August 30, that is the 90^{th} day, from the 89 previous day, at Station 86. We consider here the raw data (not the square

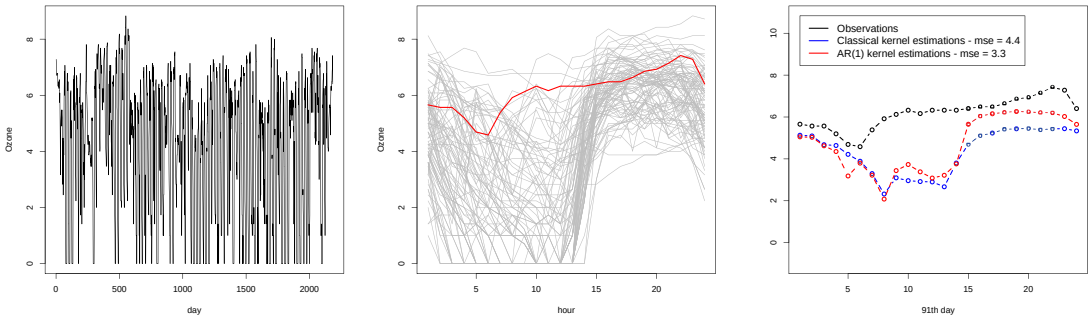


Figure 5.4 – Ozone concentrations at Station 17
Left: square root of all the series. Middle: considered curves. Right: predictions.

roots, as the analyses for this station works without a transformation), plotted in Figure 5.5. On the middle panel of this figure daily curves are plotted and the red curve represents the curve we want to forecast. The predictions are presented on the right panel of this figure and the estimated $AR(1)$ coefficients are given in Table 5.4. The overall results are similar to the previous case.

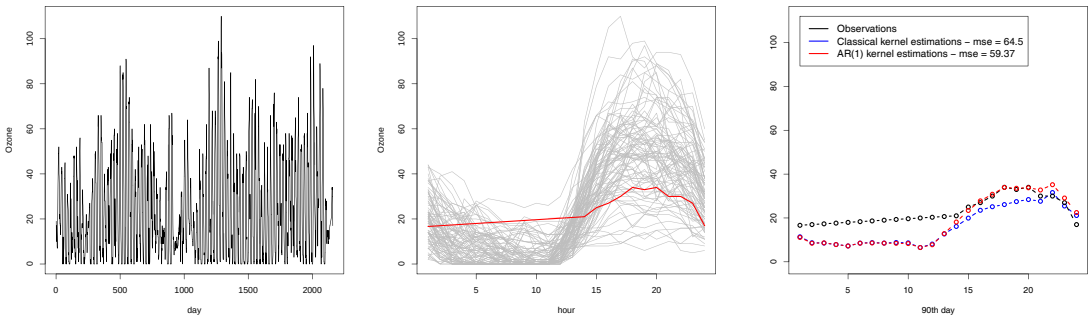


Figure 5.5 – Ozone concentrations at Station 86
Left: entire series. Middle: considered curves. Right: prediction.

r	1	2	3	4	5	6	7	8	9	10	11	12
$\widehat{a}_1$	−0.10	−0.16	−0.11	−0.01	0.09	0.06	0.13	0.03	0.24	0.09	0.01	0.14
r	13	14	15	16	17	18	19	20	21	22	23	24
$\widehat{a}_1$	0.21	0.22	0.26	0.26	0.29	0.31	0.31	0.28	0.26	0.25	0.25	0.13

Table 5.4 – Station 86: for horizon prediction r, estimated autoregressive coefficient $\widehat{a}_1$

For the two following situations, the considered curves are constructed differently. Firstly, instead of considering the day from 00h to 23h, we choose that each day begins at 5am. This choice is based on the possibility of releasing warnings in the morning for the population to decide on its daily activities. Then we predict ozone concentrations in August 31 from 5am to 11pm. The corresponding illustrations and results are given on Figure 5.6 and Table 5.5.

Finally, for Station 93, we define the day as times from 8am to 8pm. The goal here is to predict ozone concentrations in August 31, from 8am to 8pm. The corresponding

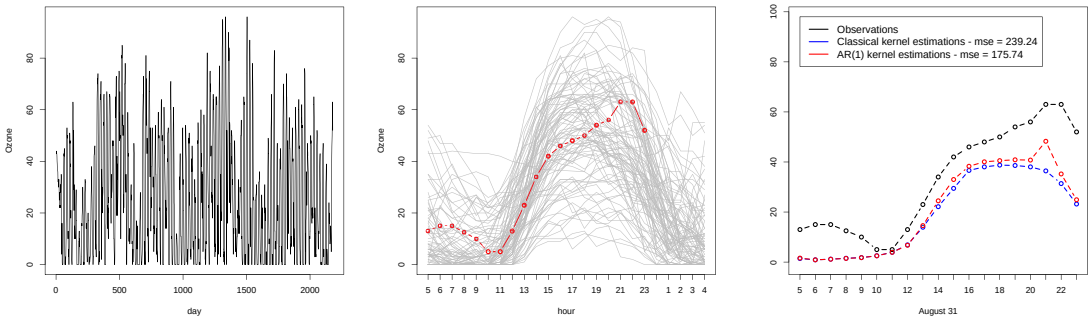


Figure 5.6 – Ozone concentration at Station 22
Left: entire series. Middle: considered curves. Right: predictions.

r	1	2	3	4	5	6	7	8	9	10	11	12
$\widehat{a}_1$	-0.5	-0.03	-0.05	-0.06	-0.03	-0.04	-0.01	-0.06	0.11	0.21	0.24	0.22
r	13	14	15	16	17	18	19					
$\widehat{a}_1$	0.19	0.16	0.22	0.27	0.31	0.28	0.21					

Table 5.5 – Station 22: for horizon prediction r, estimated autoregressive coefficient $\widehat{a}_1$

illustrations and results are given on Figure 5.7 and Table 5.6.

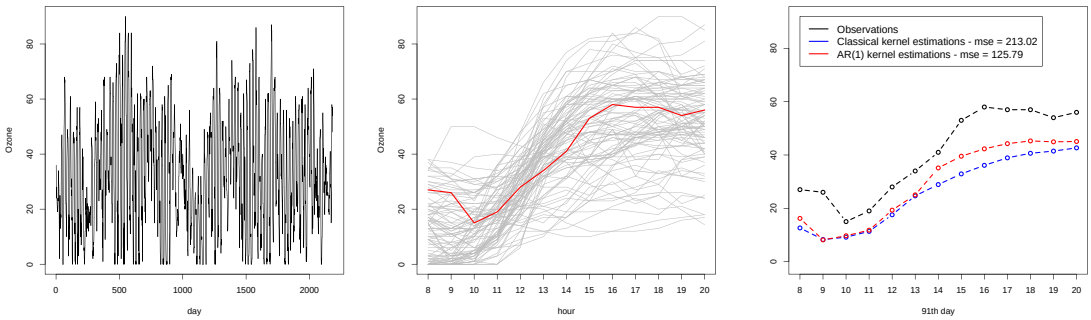


Figure 5.7 – Ozone concentrations at Station 93
Left: entire series. Middle: considered curves. Right: predictions.

r	1	2	3	4	5	6	7	8	9	10	11	12	13
$\widehat{a}_1$	0.16	0.32	0.25	0.28	0.13	0.02	0.09	0.16	0.14	0.18	0.24	0.18	0.25

Table 5.6 – Station 93: for horizon prediction r, estimated autoregressive coefficients $\widehat{a}_1$

To conclude, the three last studied cases also show that in using our methodology can improve the ozone concentration predictions obtained with the classical kernel regression estimate.

5.4 Conclusion

We have developed a two-stage procedure in order to estimate a nonlinear functional regression where the explanatory variable is functional and the residual process is stationary and autocorrelated. We have used the information of the autocorrelation function of the error process to improve the regression function kernel-based estimation. The asymptotic normality of our estimator is proved under some conditions. Some numerical results from a simulation case study and an application on real data illustrate the benefit of using this approach. Our methodology improves the standard kernel estimator in presence of highly autocorrelated data.

Potential improvements relate to the optimal implementation of our method. Indeed, the numerical illustrations indicate that there is a "sweet spot" where the number of time points provide enough information relative to the autocorrelation level to allow an optimal reduction of the prediction error. Another aspect is that for small autocorrelation levels, our approach deteriorates slightly the prediction errors compared to the use of independence-based kernel methods. An improved methods should account for that fact and revert back to the basic independence-based kernel methods in these regimes.

Finally, our approach could be extended to other time series of functional data. Aue et al. (2014) recently provided a dimension reduction technique with functional principal components (FPC) analysis that enables the use of vector-valued time series of FPC scores. However, this model did not allow of autocorrelation in the residuals that can still be present as we show in our ozone application above. A combination of the two approaches would have the potential to further improve the quality of predictions.

5.5 Appendix

Preliminary result for the proof

We use $\|x\|_q = (\mathbb{E}(x^q))^{1/q}$ to denote the norm L^q of x and C to signify a generic positive constant whose exact value may vary from case to case. The following lemma, used in the proof of Theorem 5.1, gives bounds for even moments of sum of strongly mixing random variables whose proof is given in Cox and Kim (1995).

Lemma 5.9. From Cox and Kim (1995)
Let $\xi(t)$ be a strong mixing process. Let r be a positive integer and assume $\mathbb{E}\xi(t) = 0$ and that for some $\nu > 2$,

$$M_{\nu r} = \sup_t\{\|\xi(t)\|_{\nu r}\} = \sup_t\{(\mathbb{E}|\xi(t)|^{\nu r})^{1/(\nu r)}\} \leq 1.$$

Suppose further that there is a constant V not depending on t such that

$$\mathbb{E}[|\xi(t)|^k] \leq V, \qquad 2 \leq k \leq 2r.$$

Finally, assume that the mixing coefficients satisfy

$$\sum_{t=1}^{\infty} t^{r-1}\alpha(t)^{1-2/\nu} \leq \infty.$$

Then there exists a constant C depending on r but not depending on the distribution of $\xi(t)$ nor on V, T, nor P such that

$$\mathbb{E}\left[\left(\sum_{t=1}^{T}\xi(t)\right)^{2r}\right] \leq C\{T^r M_{\nu r}^{2r}\sum_{t=P}^{\infty} t^{r-1}\alpha(t)^{1-2/\nu} + \sum_{j=1}^{r} T^j P^{2r-j} V^j\}$$

for any integer T and P with $0 < P < T$.

Lemma 5.10. Under assumptions **H1**-**H5**,

$$\begin{aligned} Q_{T_1} &= O_p\left(\sqrt{\frac{1}{T\phi(h_1)}}\right) + O_p\left(\frac{\log T}{T\phi(h_0)\phi(h_1)^{1/2}}\right) \\ &= o_p\left(\sqrt{\frac{1}{T\phi(h_1)}}\right) \end{aligned}$$

Lemma 5.11. Under assumptions **H1**-**H4**,

$$\begin{aligned} Q_{T_2} &= O_p\left(h_0^\beta + \frac{\log T}{T\phi(h_0)}\right) \\ &= O_p\left(h_0^\beta\right) + o_p\left(\sqrt{\frac{1}{T\phi(h_1)}}\right) \end{aligned}$$

Lemma 5.12. Under assumptions **H1**-**H5**,

$$\begin{aligned} Q_{T_3} &= O_p\left(h_0^\beta + T^{-1/2} + \frac{\log T}{T\phi(h_0)}\right) \\ &= O_p\left(h_0^\beta\right) + o_p\left(\sqrt{\frac{1}{T\phi(h_1)}}\right) \end{aligned}$$

Proofs

Proof of Theorem 5.1

Let $W_t = \frac{\Delta_t^{(0)}(x)}{\sum_{t=1}^T \Delta_t^{(0)}(x)}$ and adopt the convention $\frac{0}{0} = 0$. Then, we can write

$$\overline{r}(x) = \begin{cases} \sum_{t=1}^T W_t \underline{Y}_t, & \text{if } \sum_{t=1}^T W_t = 1, \\ \frac{1}{T}\sum_{t=1}^T \underline{Y}_t, & \text{otherwise.} \end{cases}$$

$$\begin{aligned} \overline{r}(x) - r(x) = &\left(\sum_{t=1}^T W_t \left[\mathbb{E}(\underline{Y}_t|X_t) - r(x)\right]\right) \mathbf{1}_{\left[\sum_{t=1}^T W_t = 1\right]} \\ &+ \left(\sum_{t=1}^T W_t \left[\underline{Y}_t - \mathbb{E}(\underline{Y}_t|X_t)\right]\right) \mathbf{1}_{\left[\sum_{t=1}^T W_t = 1\right]} + \left(\frac{1}{T}\sum_{t=1}^T \underline{Y}_t - r(x)\right) \mathbf{1}_{\left[\sum_{t=1}^T W_t = 0\right]} \end{aligned}$$

$\|\overline{r}(x) - r(x)\|_q \le \|\mathbf{A}\|_q + \|\mathbf{B}\|_q + \|\mathbf{C}\|_q$ applying Minkowski's inequality.

- $$\begin{aligned} \|\mathbf{A}\|_q &= \mathbb{E}^{1/q}\left[\sum_{t=1}^T W_t \left[r(X_t) - r(x)\right]\right]^q \\ &\le \mathbb{E}^{1/q}\left[C\sum_{t=1}^T W_t d(X_t, x)^\beta\right]^q \\ &\le \mathbb{E}^{1/q}\left[C\sum_{t=1}^T W_t h_0^\beta\right]^q \\ &\le \mathbb{E}^{1/q}\left[Ch_0^\beta\right]^q \\ &= O(h_0^\beta) \qquad \text{using assumption } \mathbf{H1}(1). \end{aligned}$$

- $\|\mathbf{B}\|_q = \|D\|_q \mathbf{1}_{[\sum_{t=1}^T W_t=1]}$

where

$$D = \sum_{t=1}^{T} W_t \left[\underline{Y}_t - \mathbb{E}(\underline{Y}_t|X_t)\right] = \frac{\overline{r}_2^*(x)}{\overline{r}_1(x)}$$

with $\overline{r}_2^*(x) = \dfrac{1}{T\mathbb{E}\left[\Delta_1^{(0)}(x)\right]} \displaystyle\sum_{t=1}^{T} \Delta_t^{(0)}(x) \left[\underline{Y}_t - \mathbb{E}(\underline{Y}_t|X_t)\right]$. We note that $\forall t : 0 \leq |\underline{Y}_t - \mathbb{E}(\underline{Y}_t|X_t)| \leq C$ since Y_t is bounded, then, $|D| \leq \displaystyle\sum_{t=1}^{T} W_t C \leq C$.

$$\begin{aligned}
|D| &= |D|\mathbf{1}_{\left[\sum_{t=1}^T \Delta_t^{(0)}(x)>c\right]} + |D|\mathbf{1}_{\left[\sum_{t=1}^T \Delta_t^{(0)}(x)\leq c\right]} \\
&\leq \frac{|\overline{r}_2^*(x)|}{\overline{r}_1(x)}\mathbf{1}_{\left[\sum_{t=1}^T \Delta_t^{(0)}(x)>c\right]} + C\mathbf{1}_{\left[\sum_{t=1}^T \Delta_t^{(0)}(x)\leq c\right]}
\end{aligned}$$

where c is a given real constant. We take $c = \dfrac{Tu}{2}$ with $u = \mathbb{E}\left[\Delta_t^{(0)}(x)\right]$ and by assumption **H2**(1) we have

$$\begin{aligned}
c_{1,0}\mathbb{P}[X_t \in \mathcal{B}(x, h_0)] \leq \mathbb{E}\left[\Delta_t^{(0)}(x)\right] \leq c_{2,0}\mathbb{P}[X_t \in \mathcal{B}(x, h_0)] \\
c_{1,0}F_x(h_0) \leq \mathbb{E}\left[\Delta_t^{(0)}(x)\right] \leq c_{2,0}F_x(h_0)
\end{aligned}$$

Since $F_x(h_0)$ is a probability, $0 \leq F_x(h_0) \leq 1$ then $0 \leq \mathbb{E}\left[\Delta_t^{(0)}(x)\right] \leq c_{2,0}$. If $\sum_{t=1}^T \Delta_t^{(0)}(x) > \frac{Tu}{2}$ then, $\overline{r}_1(x) > \frac{Tu}{2T\mathbb{E}\left[\Delta_1^{(0)}(x)\right]} > \frac{Tu}{2Tu} > \frac{1}{2}$ and therefore $\frac{|\overline{r}_2^*(x)|}{\overline{r}_1(x)} < 2|\overline{r}_2^*(x)|$, for T large enough. Consequently,

$$\begin{aligned}
|D| &\leq C|\overline{r}_2^*(x)| + C\mathbf{1}_{\left[\sum_{t=1}^T \Delta_t^{(0)}(x)\leq Tu/2\right]} \\
\|D\|_q &\leq C\|\overline{r}_2^*(x)\|_q + C\left(\mathbb{P}\left[\sum_{t=1}^{T} \Delta_t^{(0)}(x) \leq Tu/2\right]\right)^{1/q}
\end{aligned}$$

▷ $\|\overline{r}_2^*(x)\|_q = \dfrac{1}{T\mathbb{E}\left[\Delta_1^{(0)}(x)\right]} \left[\mathbb{E}\left(\displaystyle\sum_{t=1}^{T} \xi_t\right)^q\right]^{1/q}$,

where $\xi_t = \Delta_t^{(0)}(x)\left[\underline{Y}_t - \mathbb{E}(\underline{Y}_t|X_t)\right]$. We are interested in the term $\mathbb{E}\left(\sum_{t=1}^T \xi_t\right)^q$. In the following, we assume that $q = 2r$ where r is a positive integer ($r > 1$ to satisfy mixing conditions). If q takes values different than $2r$, just apply Hölder's inequality. Note that ξ_t is a strong mixing process such that $\mathbb{E}(\xi_t) = 0$.
For some $\nu > 2$,

$$M_{\nu r} = \sup_t\{(\mathbb{E}|\xi_t|^{\nu r})^{1/(\nu r)}\} = O(\phi(h_0)^{1/(\nu r)}) \leq 1.$$

for T large enough. Then, by assumptions **H2**(1) and **H3**(1), there is a constant V not depending on t such that

$$\mathbb{E}[|\xi_t|^k] \leq V = O(\phi(h_0)), \qquad 2 \leq k \leq 2r.$$

Considering assumption **H4** and Lemma 5.9 (Theorem 1 of Cox and Kim (1995)), there exists a constant C depending on r but not depending on the distribution of ξ_t nor on V, T, nor P such that

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{t=1}^{T}\xi_t\right)^{2r}\right] &\leq C\{T^r M_{\nu r}^{2r}\sum_{t=P}^{\infty} t^{r-1}\alpha(t)^{1-2/\nu} + \sum_{j=1}^{r} T^j P^{2r-j} V^j\} \\ &\leq C\{T^r \phi(h_0)^{2/\nu}\sum_{t=P}^{\infty} t^{r-1}\alpha(t)^{1-2/\nu} + \sum_{j=1}^{r} T^j P^{2r-j} V^j\}\end{aligned}$$

for any integer T and P with $0 < P < T$. Let $P = \lfloor \phi(h_0)^{-(r-2/\nu)/(\delta-r+1)} \rfloor$, where $\lfloor\cdot\rfloor$ denotes the integer part, then

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{t=1}^{T}\xi_t\right)^{2r}\right] &\leq C\left\{T^r \phi(h_0)^{2/\nu}\sum_{t=P}^{\infty} t^{r-1}\alpha(t)^{1-2/\nu} + \sum_{j=1}^{r} T^j \phi(h_0)^j\right\} \\ &\leq C\{T^r \phi(h_0)^r\} \qquad \text{using assumption } \mathbf{H4} \\ &= O\left((T\phi(h_0))^r\right)\end{aligned}$$

$$\left(\mathbb{E}\left(\sum_{t=1}^{T}\xi_t\right)^{2r}\right)^{1/(2r)} = O\left((T\phi(h_0))^{1/2}\right)$$

In summary, we have $\|\overline{r}_2^*(x)\|_q = O\left((T\phi(h_0))^{-1/2}\right)$.

$$\begin{aligned}\triangleright \quad \mathbb{P}\left[\sum_{t=1}^{T}\Delta_t^{(0)}(x) \leq \frac{Tu}{2}\right] &= \mathbb{P}\left[\sum_{t=1}^{T}\left(\Delta_t^{(0)}(x) - \mathbb{E}\left[\Delta_t^{(0)}(x)\right]\right) \leq \frac{Tu}{2} - Tu\right] \\ &\leq \mathbb{P}\left[\left|\sum_{t=1}^{T}\left(\Delta_t^{(0)}(x) - \mathbb{E}\left[\Delta_t^{(0)}(x)\right]\right)\right| \geq \frac{Tu}{2}\right] \\ &\leq \mathbb{P}\left[\frac{1}{Tu}\left|\sum_{t=1}^{T}\left(\Delta_t^{(0)}(x) - \mathbb{E}\left[\Delta_t^{(0)}(x)\right]\right)\right| \geq \frac{1}{2}\right] \\ &\leq \mathbb{P}\left[\frac{1}{T}\left|\sum_{t=1}^{T}\frac{1}{u}\left(\Delta_t^{(0)}(x) - \mathbb{E}\left[\Delta_t^{(0)}(x)\right]\right)\right| \geq \varepsilon\right]\end{aligned}$$

with $\varepsilon = \mu\left(\frac{\log T}{T\phi(h_0)}\right)^{1/2} \to 0$, where $\mu > 0$ is a constant to be chosen later.
Let $A_t = \frac{1}{u}\left(\Delta_t^{(0)}(x) - \mathbb{E}\left[\Delta_t^{(0)}(x)\right]\right)$. We see that A_t is a zero-mean real-valued process and

$$\sup_{1\leq t\leq T}\|A_t\|_\infty = \sup_{1\leq t\leq T}\left\|\frac{1}{u}\left(\Delta_t^{(0)}(x) - \mathbb{E}\left[\Delta_t^{(0)}(x)\right]\right)\right\|_\infty \leq b.$$

where b is a constant. Let $S_t = \sum_{t=1}^{T} A_t$ and $s \in \left[1, \frac{T}{2}\right]$. Then, $\mathbb{P}\left[\sum_{t=1}^{T}\Delta_t^{(0)}(x) \leq \frac{Tu}{2}\right] \leq \mathbb{P}\left[\frac{1}{T}|S_t| \geq \varepsilon\right]$. Applying the Theorem 1.3 page 27 in Bosq (1998), we have

$$\begin{aligned}\mathbb{P}\left[\sum_{t=1}^{T}\Delta_t^{(0)}(x) \leq \frac{Tu}{2}\right] \leq \mathbb{P}\left[\frac{1}{T}|S_t| \geq \varepsilon\right] &\leq 4\exp\left(\frac{-\varepsilon_T^2}{8b^2}s\right) + 22\left(1+\frac{4b}{\varepsilon}\right)^{1/2} s\alpha\left(\left\lfloor\frac{T}{2s}\right\rfloor\right) \\ &\leq 4\exp\left(\frac{-\varepsilon_T^2}{8b^2}s\right) + 22\left(1+\frac{4b}{\varepsilon}\right)^{1/2} sC\left(\frac{T}{2s}\right)^{\frac{-\delta\nu}{(\nu-2)}}\end{aligned}$$

since from assumption **H4**,

$$
\begin{aligned}
l^{\delta}\alpha(l)^{1-(2/\nu)} &\longrightarrow 0 \\
\alpha(l)^{1-(2/\nu)} &= o(l^{-\delta}) \\
\alpha(l) &= O\left(l^{\frac{-\delta\nu}{\nu-2}}\right)
\end{aligned}
$$

Note that since $\nu > 2$ and $\delta > 1 - \frac{2}{\nu}$ then we have $\frac{\delta\nu}{\nu-2} > 1$. Moreover, since $s \in \left[1, \frac{T}{2}\right]$, we have $\frac{2}{T} < \frac{1}{s} < 1$ and $1 < \frac{T}{2s} < \frac{T}{2}$.

$$
\begin{aligned}
\mathbb{P}\left[\sum_{t=1}^{T}\Delta_t^{(0)}(x) \leq \frac{Tu}{2}\right] &\leq 4\exp\left(\frac{-\mu^2\log T}{8b^2T\phi(h_0)}\frac{T}{2}\right) + 22\left(1+\frac{4b}{\varepsilon_T}\right)^{1/2}\frac{T}{2}\left(\frac{T}{2s}\right)^{\frac{-\delta\nu}{(\nu-2)}} \\
&\leq 4\exp\left(\frac{-\mu^2}{\phi(h_0)16b^2}\log T\right) + 11C\left(1+\frac{4b}{\mu\sqrt{\frac{\log T}{T\phi(h_0)}}}\right)^{1/2} T\left(\frac{T}{2s}\right)^{\frac{-\delta\nu}{(\nu-2)}} \\
&\leq 4\exp\left(\frac{-c}{\phi(h_0)}\log T\right) + 11C\left(1+\frac{4b}{\mu}\sqrt{\frac{T\phi(h_0)}{\log T}}\right)^{1/2} T^{1-\frac{\delta\nu}{\nu-2}}(2s)^{\frac{\delta\nu}{(\nu-2)}} \\
&\leq CT^{\frac{-c}{\phi(h_0)}} + C\left(\frac{T\phi(h_0)}{\log T}\right)^{1/4} T^{1-\frac{\delta\nu}{\nu-2}} \\
&\leq CT^{-\gamma} + C\left(\frac{T\phi(h_0)}{\log T}\right)^{1/4} T^{1-\frac{\delta\nu}{\nu-2}}
\end{aligned}
$$

$\gamma = \frac{c}{\phi(h_0)} > 0$, $c = \mu^2/(16b^2)$.

$$
\begin{aligned}
T\phi(h_0)\mathbb{P}\left[\sum_{t=1}^{T}\Delta_t^{(0)}(x) \leq \frac{Tu}{2}\right] &\leq T^{1-\gamma}\phi(h_0) + T^{5/4}\phi(h_0)^{5/4}(\log T)^{-1/4}T^{1-\frac{\delta\nu}{\nu-2}} \\
&\leq T^{1-\gamma}\phi(h_0) + \left(\frac{T\phi(h_0)}{\log T}\right)^{5/4}\frac{\log T}{T^{\frac{\delta\nu}{\nu-2}-1}} \\
&\leq T^{1-\gamma}\phi(h_0) + T\phi(h_0)\left(\frac{T\phi(h_0)}{\log T}\right)^{1/4} T^{1-\frac{\delta\nu}{\nu-2}}
\end{aligned}
$$

$T^{1-\gamma}\phi(h_0) \to 0$ if $\gamma > 1$ therefore, it suffices to choose μ such that $\gamma > 1$. Since $\delta > 1 - 2/\nu$, we have $1 - \frac{\delta\nu}{\nu-2} < 0$. We have then $\left(\frac{T\phi(h_0)}{\log T}\right)^{1/4} T^{1-\frac{\delta\nu}{\nu-2}} = o\left(\frac{1}{T\phi(h_0)}\right)$. Therefore, $\left(\mathbb{P}\left[\sum_{t=1}^{T}\Delta_t^{(0)}(x) \leq \frac{Tu}{2}\right]\right)^{1/q} = o\left(\left(\frac{1}{T\phi(h_0)}\right)^{1/p}\right)$. In conclusion,

$$
\begin{aligned}
\mathbb{E}^{1/q}\left[\mathbf{B}\right]^q &= \|D\|_q \mathbf{1}_{\left[\sum_{t=1}^{T} W_t = 1\right]} \\
&= O\left(\left(\frac{1}{T\phi(h_0)}\right)^{1/2}\right) + o\left(\left(\frac{1}{T\phi(h_0)}\right)^{1/q}\right)
\end{aligned}
$$

- $$\|\mathbf{C}\|_q \leq \mathbb{E}^{1/q}\left[\left|\frac{1}{T}\sum_{t=1}^{T}\underline{Y}_t - r(x)\right|\mathbf{1}_{\left[\sum_{t=1}^{T}W_t=0\right]}\right]^q \leq C\mathbb{E}^{1/q}\left[\mathbf{1}_{\left[\sum_{t=1}^{T}W_t=0\right]}\right]^q$$
$$\leq C\left[\mathbb{P}\left(\sum_{t=1}^{T}W_t=0\right)\right]^{1/q} \leq C\left[\mathbb{P}\left(\sum_{t=1}^{T}\Delta_t^{(0)}(x)=0\right)\right]^{1/q}$$
$$\leq C\left[\mathbb{P}\left(\sum_{t=1}^{T}\Delta_t^{(0)}(x)\leq\frac{Tu}{2}\right)\right]^{1/q}$$
$$= o\left(\left(\frac{1}{T\phi(h_0)}\right)^{1/q}\right)$$

Finally,

$$\|\overline{r}(x)-r(x)\|_q = O(h_0^\beta) + O\left(\left(\frac{1}{T\phi(h_0)}\right)^{1/2}\right) + o\left(\left(\frac{1}{T\phi(h_0)}\right)^{1/q}\right)$$

■

Proof of Theorem 5.3

The proof follows work of Masry (2005). Note that $\underline{Y}_t = r(X_t) + \epsilon_t$; we obtain $\overline{r}(x) = r(x) + B_T(x) + V_T(x)$ where $B_T(x)$ is the bias term and $V_T(x)$ is the variance effect defined by

$$B_T(x) = \frac{\mathbb{E}[\overline{r}_2(x)] - r(x)\mathbb{E}[\overline{r}_1(x)]}{\mathbb{E}[\overline{r}_1(x)]} \qquad V_T(x) = \frac{Q_T(x) - B_T(x)(\overline{r}_1(x) - \mathbb{E}[\overline{r}_1(x)])}{\overline{r}_1(x)}$$

with $Q_T(x) = (\overline{r}_2(x) - \mathbb{E}[\overline{r}_2(x)]) - r(x)(\overline{r}_1(x) - \mathbb{E}[\overline{r}_1(x)])$. By the result of Masry (2005), $B_T(x) = o(h_0^\beta)$ and

$$[T\phi(h_0)]^{1/2}[\overline{r}(x) - r(x) - B_T(x)] \xrightarrow{d} \mathcal{N}(0, \sigma^2(x))$$

where $\sigma^2(x) = \frac{Cg_2(x)}{f_1(x)}$. ■

Proof of Theorem 5.6

$$\begin{aligned}\widehat{\underline{Y}}_t &= Y_t - \widehat{a}_1\left(Y_{t-1} - \widehat{r}(X_{t-1})\right)\\ &= Y_t - \widehat{a}_1\left(r(X_{t-1}) + u_{t-1} - \widehat{r}(X_{t-1})\right)\\ &\quad + a_1\left(r(X_{t-1}) + u_{t-1} - \widehat{r}(X_{t-1})\right) - a_1\left(r(X_{t-1}) + u_{t-1} - \widehat{r}(X_{t-1})\right)\\ &= Y_t - a_1u_{t-1} - (\widehat{a}_1 - a_1)u_{t-1} + a_1\left(\widehat{r}(X_{t-1}) - r(X_{t-1})\right) + (\widehat{a}_1 - a_1)\left(\widehat{r}(X_{t-1}) - r(X_{t-1})\right)\end{aligned}$$

We have

$$\widetilde{r}(x) = \frac{\frac{1}{T\mathbb{E}\left[\Delta_1^{(1)}(x)\right]}\sum_{t=1}^{T}\widehat{\underline{Y}}_t\Delta_t^{(1)}(x)}{\frac{1}{T\mathbb{E}\left[\Delta_1^{(1)}(x)\right]}\sum_{t=1}^{T}\Delta_t^{(1)}(x)} = \frac{\widetilde{r}_2(x)}{\widetilde{r}_1(x)}$$

$$\sum_{t=1}^{T}\widehat{\underline{Y}}_t\Delta_t^{(1)}(x) = \sum_{t=1}^{T}\underline{Y}_t\Delta_t^{(1)}(x) - \sum_{t=1}^{T}(\widehat{a}_1 - a_1)u_{t-1}\Delta_t^{(1)}(x) + \sum_{t=1}^{T}a_1\left(\widehat{r}(X_{t-1}) - r(X_{t-1})\right)\Delta_t^{(1)}(x)$$
$$+ \sum_{t=1}^{T}(\widehat{a}_1 - a_1)\left(\widehat{r}(X_{t-1}) - r(X_{t-1})\right)\Delta_t^{(1)}(x)$$

Then $\widetilde{r}_2(x) = \overline{r}_2(x)^\sharp - \widetilde{r}_{21}(x) + \widetilde{r}_{22}(x) + \widetilde{r}_{23}(x)$ with

$$\overline{r}_2(x)^\sharp = \frac{1}{T\mathbb{E}\left[\Delta_t^{(1)}(x)\right]} \sum_{t=1}^{T} \underline{Y}_t \Delta_t^{(1)}(x)$$

$$\widetilde{r}_{21}(x) = \frac{1}{T\mathbb{E}\left[\Delta_t^{(1)}(x)\right]} \sum_{t=1}^{T} (\widehat{a}_1 - a_1) u_{t-1} \Delta_t^{(1)}(x)$$

$$\widetilde{r}_{22}(x) = \frac{1}{T\mathbb{E}\left[\Delta_t^{(1)}(x)\right]} \sum_{t=1}^{T} a_1 \left(\widehat{r}(X_{t-1}) - r(X_{t-1})\right) \Delta_t^{(1)}(x)$$

$$\widetilde{r}_{23}(x) = \frac{1}{T\mathbb{E}\left[\Delta_t^{(1)}(x)\right]} \sum_{t=1}^{T} (\widehat{a}_1 - a_1) \left(\widehat{r}(X_{t-1}) - r(X_{t-1})\right) \Delta_t^{(1)}(x)$$

Note that $\overline{r}_2(x)^\sharp = \overline{r}_2(x)$ with K_0 and h_0 are replaced by K_1 and h_1 respectively. Since $\widetilde{r}(x) = \dfrac{\widetilde{r}_2(x)}{\widetilde{r}_1(x)}$ we have $\widetilde{r}(x) = \overline{r}(x) - Q_{T_1} + Q_{T_2} + Q_{T_3}$ with $Q_{T_l} = \frac{\widetilde{r}_{2l}(x)}{\widetilde{r}_1(x)}$, for $l = 1, 2, 3$. We analyze the asymptotic properties of $Q_{T_l}, l = 1, 2, 3$, in Lemmas 5.10, 5.11 and 5.12, which are key results for proof of this theorem.

■

Proof of Lemma 5.10

Let $\widehat{a}_1 - a_1 = \widehat{a}_1 - \overline{a}_1 + \overline{a}_1 - a_1$ with $\widehat{a}_1 = \dfrac{\sum_{i=1}^{T-1} \widehat{u}_{i+1}\widehat{u}_i}{\sum_{i=1}^{T-1} \widehat{u}_i^2}$ and $\overline{a}_1 = \dfrac{\sum_{i=1}^{T-1} u_{i+1}u_i}{\sum_{i=1}^{T-1} u_i^2}$. We decompose $\widetilde{r}_{21}$ such as

$$\begin{aligned}\widetilde{r}_{21}(x) &= \frac{1}{T\mathbb{E}\left[\Delta_t^{(1)}(x)\right]} \sum_{t=1}^{T} (\overline{a}_1 - a_1) u_{t-1} \Delta_t^{(1)}(x) + \frac{1}{T\mathbb{E}\left[\Delta_t^{(1)}(x)\right]} \sum_{t=1}^{T} (\widehat{a}_1 - \overline{a}_1) u_{t-1} \Delta_t^{(1)}(x) \\ &= \widetilde{r}_{21A}(x) + \widetilde{r}_{21B}(x)\end{aligned}$$

$\widehat{a}_1$ is found through an autoregression of $\widehat{u}_t = Y_t - \widehat{r}(X_t)$ on $\widehat{u}_{t-1}$, that is $\begin{pmatrix} \widehat{u}_t \\ \vdots \\ \widehat{u}_T \end{pmatrix} = \widehat{a}_1 \begin{pmatrix} \widehat{u}_{t-1} \\ \vdots \\ \widehat{u}_{T-1} \end{pmatrix}$ where $\widehat{a}_1 = \left(\widehat{U}_1'\widehat{U}_1\right)^{-1} \widehat{U}_1'\widehat{u}$ with $\widehat{u} = \begin{pmatrix} \widehat{u}_2 \\ \vdots \\ \widehat{u}_T \end{pmatrix}$ and $\widehat{U}_1 = \begin{pmatrix} \widehat{u}_1 \\ \vdots \\ \widehat{u}_{T-1} \end{pmatrix}$.

Part 1- Lemma 5.10 First, we deal with $\widetilde{r}_{21A}(x)$ that is the situation where the linear regression's coefficient based on $Y_t - r(X_t) = u_t$ is close to the true one. Define

$$G_1 = \frac{1}{T}\sum_{t=2}^{T} u_{t-1}^2 = \frac{1}{T}U_1'U_1 \qquad \text{and} \qquad \Gamma_1 = \frac{1}{T}\sum_{t=2}^{T} \mathbb{E}(u_{t-1}^2) = \frac{1}{T}\mathbb{E}(U_1'U_1) = \frac{T-1}{T}\gamma_u(0),$$

where γ_u is the covariance function of the process $\{u_t\}$; $\gamma_u(j) = Cov(u_t, u_{t-j})$. We use Theorem 5.3.2. in Deistler and Hannan (1988), page 167 and write that

$$|G_1 - \Gamma_1| = O_p\left(\sqrt{\frac{\log\log T}{T}}\right) \tag{5.6}$$

Note that

$$\overline{a}_1 - a_1 = \frac{\sum_{t=1}^{T-1} u_{t+1}u_t}{\sum_{t=1}^{T-1} u_t^2} - a_1 = \frac{\sum_{t=1}^{T-1} u_t(u_{t+1} - a_1 u_t)}{\sum_{t=1}^{T-1} u_t^2} = \frac{\sum_{t=2}^{T} u_{t-1}(u_t - a_1 u_{t-1})}{\sum_{t=2}^{T} u_{t-1}^2}$$
$$= \frac{\sum_{t=2}^{T} u_{t-1}\epsilon_t}{\sum_{t=2}^{T} u_{t-1}^2} = \frac{\frac{1}{T}\sum_{t=2}^{T} u_{t-1}\epsilon_t}{G_1}$$

So check magnitude of $\frac{1}{T}\sum_{t=2}^{T} u_{t-1}\epsilon_t$:

$$\mathbb{E}\left(\frac{1}{T}\sum_{t=2}^{T} u_{t-1}\epsilon_t\right)^2 = \frac{1}{T^2}\sum_{t=2}^{T}\mathbb{E}(u_{t-1})^2\mathbb{E}(\epsilon_t)^2 = \frac{T-1}{T^2}\gamma_u(0)\sigma_\epsilon^2 = O\left(T^{-1}\right)$$

with $\sigma_\epsilon^2 = Var(\epsilon_t)$. Thus, $\frac{1}{T}\sum_{t=2}^{T} u_{t-1}\epsilon_t$ is of order $O_p\left(T^{-1/2}\right)$. Combining (5.6) and this last result, we have

$$|\overline{a}_1 - a_1| = O_p\left(T^{-1/2}\right) \tag{5.7}$$

$$\|\widetilde{r}_{21A}(x)\|_2 = \frac{1}{T\mathbb{E}\left[\Delta_t^{(1)}(x)\right]}\left\|\sum_{t=1}^{T}(\overline{a}_1 - a_1)u_{t-1}\Delta_t^{(1)}(x)\right\|_2$$
$$\leq \frac{C_1}{TF_x(h_1)}\sum_{t=1}^{T}\left\|(\overline{a}_1 - a)u_{t-1}\Delta_t^{(1)}(x)\right\|_2 \leq \frac{1}{TF_x(h_1)}\sum_{t=1}^{T}|\overline{a}_1 - a_1|\left\|u_{t-1}\Delta_t^{(1)}(x)\right\|_2$$

Notice that $\{u_t\}$ is independent of $\{X_t\}$, then $\left\|u_{t-1}\Delta_t^{(1)}(x)\right\|_2 \leq C\left[\gamma_u(0)F_x(h_1)\right]^{1/2}$. Combining this last and (5.7), we have $\widetilde{r}_{21A}(x) = O_p\left(\sqrt{\frac{1}{T\phi(h_1)}}\right)$ by assumption **H3**(1).

Part 2 - Lemma 5.10 Let us treat $\widetilde{r}_{21B}(x)$. We have

$$\widehat{a}_1 = \frac{\frac{1}{T}\sum_{t=2}^{T}\widehat{u}_t\widehat{u}_{t-1}}{\frac{1}{T}\sum_{t=2}^{T}\widehat{u}_{t-1}^2} = \frac{\widehat{a}_{12}}{\widehat{a}_{11}} \quad \text{and} \quad \overline{a}_1 = \frac{\frac{1}{T}\sum_{t=2}^{T} u_t u_{t-1}}{\frac{1}{T}\sum_{t=2}^{T} u_{t-1}^2} = \frac{\overline{a}_{12}}{\overline{a}_{11}} \quad \text{then}$$

$$\widehat{a}_1 - \overline{a}_1 = \frac{\widehat{a}_{12}\overline{a}_{11} - \overline{a}_{12}\widehat{a}_{11}}{\overline{a}_{11}\widehat{a}_{11}}$$
$$= \frac{\widehat{a}_{12} - \overline{a}_{12}}{\widehat{a}_{11}} + \frac{\overline{a}_{12}}{\widehat{a}_{11}} - \frac{\overline{a}_{12}}{\overline{a}_{11}}$$
$$= \frac{\widehat{a}_{12} - \overline{a}_{12}}{\widehat{a}_{11}} + \overline{a}_{12}\left(\frac{\overline{a}_{11} - \widehat{a}_{11}}{\widehat{a}_{11}\overline{a}_{11}}\right)$$

Let us study $\widehat{a}_{12} - \overline{a}_{12} = \frac{1}{T}\sum_{t=2}^{T}(\widehat{u}_{t-1}\widehat{u}_t - u_{t-1}u_t)$. Let

$$\widehat{u}_t = r(X_t) + u_t - \frac{1}{T}\sum_{i=1}^{T} W_i(X_t)\{r(X_i) + u_i\}$$
$$= u_t + \frac{1}{T}\sum_{i=1}^{T} W_i(X_t)\{r(X_t) - r(X_i)\} - \frac{1}{T}\sum_{i=1}^{T} W_i(X_t)u_i$$
$$= u_t - \widehat{B}_t - \widehat{V}_t$$

with $W_i(X_t) = \frac{\Delta_i^{(0)}(X_t)}{\frac{1}{T}\sum_{i=1}^{T}\Delta_i^{(0)}(X_t)} \geq 0$.

$$\widehat{u}_{t-1}\widehat{u}_t - u_{t-1}u_t = (u_{t-1} - \widehat{B}_{t-1} - \widehat{V}_{t-1})(u_t - \widehat{B}_t - \widehat{V}_t) - u_{t-1}u_t$$
$$= (-u_{t-1}\widehat{B}_t - u_{t-1}\widehat{V}_t - \widehat{B}_{t-1}u_t - \widehat{V}_{t-1}u_t)$$
$$+ (\widehat{B}_{t-1}\widehat{B}_t + \widehat{B}_{t-1}\widehat{V}_t + \widehat{V}_{t-1}\widehat{B}_t + \widehat{V}_{t-1}\widehat{V}_t)$$

Part 2a

$$\left|\frac{1}{T}\sum_{t=2}^{T}\left(\widehat{V}_{t-1}\widehat{V}_t+\widehat{B}_{t-1}\widehat{B}_t+\widehat{B}_{t-1}\widehat{V}_t\ +\widehat{V}_{t-1}\widehat{B}_t\right)\right|$$
$$\leq \frac{1}{T}\sum_{t=2}^{T}\left[(\sup_s|\widehat{V}_s|)^2+(\sup_s|\widehat{B}_s|)^2+2\sup_s|\widehat{B}_s|\sup_s|\widehat{V}_s|\right]$$

- $\left|\widehat{B}_t\right| = \left|\frac{1}{T}\sum_{i=1}^{T}W_i(X_t)\{r(X_t)-r(X_i)\}\right| \leq \frac{1}{T}\sum_{i=1}^{T}W_i(X_t)\left|r(X_t)-r(X_i)\right|$
$$\leq \frac{1}{T}\sum_{i=1}^{T}W_i(X_t)(cd(X_i,X_t)^\beta)$$
$$\leq \frac{1}{T}\sum_{i=1}^{T}W_i(X_t)(ch_0^\beta)$$
$$\leq \left(\frac{c_{2,0}}{c_{1,0}}ch_0^\beta\right)$$
$$= O(h_0^\beta)$$

using assumption **H1**(1)

- $\left|\widehat{V}_t\right| = \frac{1}{T}\left|\sum_{i=1}^{T}W_i(X_t)u_i\right|$

Recall that ϵ_i and u_i are of mean 0 and respectively variances σ_ϵ^2 and $\sigma_u^2=\frac{\sigma_\epsilon^2}{1-a_1^2}$. Moreover, recall that $\varepsilon=\mu\left(\frac{\log T}{T\phi(h_0)}\right)^{1/2}\to 0$, where $\mu>0$ is a constant. We will show that $\left\|\widehat{V}_t\right\|_2 = O\left(\frac{1}{T\phi(h_0)}\right)^{1/2}$. We can write $\widehat{V}_t=\frac{h(X_t)}{\hat{f}(X_t)}$ with

$$h(X_t)=\frac{1}{T\mathbb{E}\left[\Delta_1^{(0)}(X_t)\right]}\sum_{i=1}^{T}\Delta_i^{(0)}(X_t)u_i \qquad \widehat{f}(X_t)=\frac{1}{T\mathbb{E}\left[\Delta_1^{(0)}(X_t)\right]}\sum_{i=1}^{T}\Delta_i^{(0)}(X_t)\longrightarrow 1$$

Since $h(X_t)\to\frac{\mathbb{E}[\Delta_i^{(0)}(X_t)u_i]}{\mathbb{E}[\Delta_i^{(0)}(X_t)]}=0$, then $h(X_t)$ is asymptotically bounded by a constant H. Then $|\widehat{V}_t|\leq H$ for T large enough and we have, for a given real constant c:

$$|\widehat{V}_t| = |\widehat{V}_t|\mathbf{1}_{[\sum_{i=1}^{T}\Delta_i^{(0)}(X_t)>c]}+|\widehat{V}_t|\mathbf{1}_{[\sum_{i=1}^{T}\Delta_i^{(0)}(X_t)\leq c]}$$
$$\leq \frac{|h(X_t)|}{\widehat{f}(X_t)}\mathbf{1}_{[\sum_{i=1}^{T}\Delta_i^{(0)}(X_t)>c]}+H\times\mathbf{1}_{[\sum_{i=1}^{T}\Delta_i^{(0)}(X_t)\leq c]}$$

Let us take $c=\frac{Tu}{2}$ with $u=\mathbb{E}\left[\Delta_i^{(0)}(X_t)\right]$. Consequently, we have $c_{1,0}F_{X_t}(h_0)\leq\mathbb{E}\left[\Delta_i^{(0)}(X_t)\right]\leq c_{2,0}F_{X_t}(h_0)$ from assumption **H2**(1). Since $F_{X_t}(h_0)$ is a probability, $0\leq F_{X_t}(h_0)\leq 1$ then $0\leq\mathbb{E}\left[\Delta_i^{(0)}(X_t)\right]\leq c_{2,0}$. If $\sum_{i=1}^{T}\Delta_i^{(0)}(X_t)>\frac{Tu}{2}$ then, $\widehat{f}(X_t)>\frac{Tu}{2TF_{X_t}(h_0)}>\frac{1}{2}>C$ and

therefore $\frac{|h(X_t)|}{\widehat{f}(X_t)} < C|h(X_t)|$, for T large enough. Consequently,

$$\|\widehat{V}_t\|_2 \leq C\|h(X_t)\|_2 + H \times \left(\mathbb{P}\left[\sum_{i=1}^{T} \Delta_i^{(0)}(X_t) \leq \frac{Tu}{2}\right]\right)^{1/2}$$

$$\|h(X_t)\|_2 = \frac{1}{T\mathbb{E}\left[\Delta_1^{(0)}(X_t)\right]} \left[\mathbb{E}\left(\sum_{i=1}^{T} \xi_i\right)^2\right]^{1/2},$$

where $\xi_i = \Delta_i^{(0)}(X_t)u_i$.

$$\mathbb{E}\left(\sum_{i=1}^{T} \xi_i\right)^2 = \sum_{i=1}^{T} \mathbb{E}[\xi_i^2] + \sum_{i=1}^{T} \sum_{j=1, j\neq i}^{T} \mathbb{E}[\xi_i \xi_j]$$

$$\begin{aligned}\sum_{i=1}^{T} \mathbb{E}[\xi_i^2] = T\mathbb{E}\left(\Delta_i^{(0)}(X_t)u_i\right)^2 &\leq T\mathbb{E}\left[(\Delta_i^{(0)}(X_t))^2\right]\mathbb{E}(u_i^2) \\ &\leq Tc_{2,0}^2 F_{X_t}(h_0)\sigma_u^2 \\ &= O\left(T\phi(h_0)\right)\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{T} \sum_{j=1, j\neq i}^{T} \mathbb{E}\left[\xi_i \xi_j\right] &= \sum_{i=1}^{T} \sum_{j=1, j\neq i}^{T} \mathbb{E}\left[\Delta_i^{(0)}(X_t)u_i \Delta_j^{(0)}(X_t)u_j\right] \\ &= \sum_{i=1}^{T} \sum_{j=1, j\neq i}^{T} \mathbb{E}\left[\Delta_i^{(0)}(X_t)\Delta_j^{(0)}(X_t)\right]\mathbb{E}[u_i u_j] \\ &= \sum_{i=1}^{T} \sum_{j=1, j\neq i}^{T} \mathbb{E}\left[\Delta_i^{(0)}(X_t)\Delta_j^{(0)}(X_t)\right] Cov(u_i, u_j)\end{aligned}$$

Let $i > j$ and $i = j + r$ with $r > 0$:

$$\begin{aligned}Cov(u_i, u_j) = Cov(u_{j+r}, u_j) = Cov(a_1 u_{j+r-1} + \epsilon_{j+r}, u_j) &= a_1 Cov(u_{j+r-1}, u_j) \\ &= a_1 Cov(a_1 u_{j+r-2} + \epsilon_{j+r-1}, u_j) \\ &= a_1^2 Cov(a_1\ u_{j+r-3} + \epsilon_{j+r-2}, u_j) \\ &= a_1^r Cov(u_j, u_j) = a_1^r \gamma_u(0) \\ &= a_1^r \sigma_u^2\end{aligned}$$

$$\begin{aligned}\sum\sum_{0<|i-j|<r} \mathbb{E}\left[\xi_i \xi_j\right] &= \sum\sum_{0<|i-j|<r} \mathbb{E}\left[\Delta_i^{(0)}(X_t)u_i \Delta_j^{(0)}(X_t)u_j\right] \\ &\leq \sum\sum_{0<|i-j|<r} \mathbb{P}[(X_i, X_j) \in \mathcal{B}(X_t, h_0) \times \mathcal{B}(X_t, h_0)]\mathbb{E}[u_i u_j] \\ &\leq \sum\sum_{0<|i-j|<r} F_{X_t, X_t}^{i,j}(h_0) Cov(u_i u_j) \\ &\leq C \sum\sum_{0<|i-j|<r} \psi_1(h_0) f_2(X_t) a_1^r \sigma_u^2 \\ &\leq CTr\phi(h_0)^2 \\ &= O(T\phi(h_0))\end{aligned}$$

with $r = \phi(h_0)^{-1}$ and using assumption **H3**(2).

$$\begin{aligned}
\sum_{|i-j|>r}\sum \mathbb{E}\left[\xi_i\xi_j\right] &= \sum_{|i-j|>r}\sum \mathbb{E}\left[\Delta_i^{(0)}(X_t)\Delta_j^{(0)}(X_t)\right] Cov(u_i,u_j) \\
&\leq CF^{i,j}_{X_t,X_t}(h_0)(|\mathbb{E}[u_i^\nu]|\,|\mathbb{E}[u_j^\nu]|)^{1/\nu}\sum_{|i-j|>r}\sum \alpha(|i-j|)^{1-2/\nu} \\
&\leq CF^{i,j}_{X_t,X_t}(h_0)T\sum_{s>r} s^{-\delta}s^{\delta}\alpha(s)^{1-2/\nu} \\
&\leq C\psi_1(h_0)f_2(X_t)Tr^{-\delta}\sum_{s>r} s^{\delta}\alpha(s)^{1-2/\nu} \\
&\leq C\phi(h_0)^2Tr^{-\delta} \\
&\leq C\phi(h_0)^2T\phi(h_0)^{\delta} \\
&= O(T\phi(h_0))
\end{aligned}$$

under assumption **H3**(2) and assumption **H4** with $\delta > 1 - \frac{2}{\nu}$ where $\nu > 2$. Consequently, $\mathbb{E}\left(\sum_{i=1}^T \xi_i\right)^2 = O\left(T\phi(h_0)\right)$. Then,

$$\begin{aligned}
\|h(X_t)\|_2 &= \frac{1}{T\mathbb{E}\left[\Delta_i^{(0)}(X_t)\right]} O\left((T\phi(h_0))^{1/2}\right) \\
&= O\left((T\phi(h_0))^{-1/2}\right)
\end{aligned}$$

In the proof of Theorem 5.1, we have found $\left(\mathbb{P}\left[\sum_{i=1}^T \Delta_i^{(0)}(X_t) \leq Tu/2\right]\right)^{1/2} = o\left((T\phi(h_0))^{-1/2}\right)$. Then, $\|\widehat{V}_t\|_2 \leq O\left((T\phi(h_0))^{-1/2}\right) + o\left((T\phi(h_0))^{-1/2}\right)$. In conclusion, $\|\widehat{V}_t\|_2^2 = O\left(\frac{1}{T\phi(h_0)}\right)$. We deduce that for all t, $|\widehat{V}_t| = O_p\left(\left(\frac{\log T}{T\phi(h_0)}\right)^{1/2}\right)$ then, $\sup_t|\widehat{V}_t| = O_p\left(\left(\frac{\log T}{T\phi(h_0)}\right)^{1/2}\right)$. Therefore

$$\frac{1}{T}\sum_{t=2}^T\left[(\sup_s|\widehat{V}_s|)^2 + (\sup_s|\widehat{B}_s|)^2 + 2\sup_s|\widehat{B}_s|\sup_s|\widehat{V}_s|\right] = O_p\left(\frac{\log T}{T\phi(h_0)} + h_0^{\beta}\right)$$

Part 2b Let us treat $\left|\frac{1}{T}\sum_{t=2}^T(-u_{t-1}\widehat{B}_t - u_{t-1}\widehat{V}_t - \widehat{B}_{t-1}u_t - \widehat{V}_{t-1}u_t)\right|$

$$\begin{aligned}
\diamondsuit \quad \frac{1}{T}\sum_{t=2}^T u_{t-1}\widehat{V}_t &= \frac{1}{T}\sum_{t=2}^T u_{t-1}\frac{\frac{1}{T}\sum_{i=1}^T\Delta_i^{(0)}(X_t)u_i}{\frac{1}{T}\sum_{i=1}^T\Delta_i^{(0)}(X_t)} \\
&\approx \frac{1}{T}\sum_{t=2}^T u_{t-1}\frac{\frac{1}{T}\sum_{i=1}^T\Delta_i^{(0)}(X_t)u_i}{\mathbb{E}\left(\Delta_i^{(0)}(X_t)\right)}
\end{aligned}$$

since $\frac{1}{TE(\Delta_i^{(0)}(X_t))}\sum_{i=1}^T\Delta_i^{(0)}(X_t)$ converges in probability to 1. Note that u_t has linear process representation $u_t = \sum_{j=0}^1 c_j\epsilon_{t-j}$. Then

$$\begin{aligned}
\frac{1}{T}\sum_{t=2}^T u_{t-1}\frac{1}{T}\sum_{i=1}^T\frac{\Delta_i^{(0)}(X_t)}{\mathbb{E}\left(\Delta_i^{(0)}(X_t)\right)}u_i &= \frac{1}{T}\sum_{t=2}^T\sum_{i=1}^T\frac{1}{T}\frac{\Delta_i^{(0)}(X_t)}{\mathbb{E}\left(\Delta_i^{(0)}(X_t)\right)}u_iu_{t-1} \\
&= \frac{1}{T}\sum_{t=2}^T\sum_{i=1}^T\frac{1}{T}\frac{\Delta_i^{(0)}(X_t)}{\mathbb{E}\left(\Delta_i^{(0)}(X_t)\right)}\left(\sum_{s=0}^1 c_s\epsilon_{i-s}\right)\left(\sum_{b=0}^1 c_b\epsilon_{t-1-b}\right)
\end{aligned}$$

In addition, notice that X and ϵ are independent; thus

$$\mathbb{E}\left(\frac{1}{T}\sum_{t=2}^{T} u_{t-1}\frac{1}{T}\sum_{i=1}^{T}\frac{\Delta_i^{(0)}(X_t)}{\mathbb{E}\left(\Delta_i^{(0)}(X_t)\right)}u_i\right)^2 = \frac{1}{T^2}\sum_{a=0}^{1}\sum_{b=0}^{1}\sum_{g=0}^{1}\sum_{s=0}^{1}\sum_{t=2}^{T}\sum_{p=2}^{T}\sum_{i=1}^{T}\sum_{j=1}^{T} c_a c_b c_g c_s \quad (5.8)$$
$$\times \mathbb{E}\left[\frac{1}{T^2}\frac{\Delta_i^{(0)}(X_t)}{\mathbb{E}\left(\Delta_i^{(0)}(X_t)\right)}\frac{\Delta_j^{(0)}(X_p)}{\mathbb{E}\left(\Delta_j^{(0)}(X_p)\right)}\right]$$
$$\times \mathbb{E}[\epsilon_{t-1-s}\epsilon_{p-1-g}\epsilon_{i-b}\epsilon_{j-a}]$$

Because ϵ's are i.i.d., the foregoing expectation is non zero when

1. $i-b=j-a$ and $t-1-s=p-1-g$;
2. $i-b=t-1-s$ and $j-a=p-1-g$;
3. $i-b=p-1-g$ and $j-a=t-1-s$;
4. $i-b=j-a=t-1-s=p-1-g$.

Case 1: $i-b=j-a$ **and** $t-s=p-g$

1.1 $a=b=g=s=0$ then the expression is non zero when $i=j$ and $t=p$ and

$$\begin{aligned}(5.8) &= \frac{c_0^4}{T^4}\sum_{t=2}^{T}\sum_{i=1}^{T}\mathbb{E}\left[\left(\frac{\Delta_i^{(0)}(X_t)}{\mathbb{E}\left(\Delta_i^{(0)}(X_t)\right)}\right)^2\right]\mathbb{E}[\epsilon_{t-1}^2\epsilon_i^2]\\ &= \frac{c_0^4}{T^4}\sum_{t=2}^{T}\sum_{i=1}^{T}\frac{c^2F_{X_t}(h_0)}{c^2F_{X_t}^2(h_0)}\gamma_\epsilon(0)\gamma_\epsilon(0) \qquad \text{by assumption } \mathbf{H3}(1)\\ &= C\frac{1}{T^4}\sum_{t=2}^{T}\sum_{i=1}^{T}\frac{1}{F_{X_t}(h_0)} \le \frac{C}{T^2\phi(h_0)} = O(T^{\epsilon_1-2}) = O(T^{-1})\end{aligned}$$

1.2 $a=b=g=s=1$; **1.3** $a=b=1$ and $g=s=0$; **1.4** $a=b=0$ and $g=s=1$ are similar to **1.1**

1.5 $a=b=g=0$ and $s=1$ then the expression is non zero when $i=j$ and $t=p+1$ and

$$\begin{aligned}(5.8) &= \frac{c_0^3c_1}{T^4}\sum_{t=2}^{T}\sum_{i=1}^{T}\mathbb{E}\left[\frac{\Delta_i^{(0)}(X_t)}{\mathbb{E}\left(\Delta_i^{(0)}(X_t)\right)}\frac{\Delta_i^{(0)}(X_{t-1})}{\mathbb{E}\left(\Delta_i^{(0)}(X_{t-1})\right)}\right]\mathbb{E}[\epsilon_{t-2}^2\epsilon_i^2]\\ &= \frac{c_0^3c_1}{T^4}\sum_{t=2}^{T}\sum_{i=1}^{T}\mathbb{E}\left[\frac{\mathbb{P}[(X_t,X_{t-1})\in\mathcal{B}(X_i,h_0)\times\mathcal{B}(X_i,h_0)]}{F_{X_t}(h_0)F_{X_{t-1}}(h_0)}\right]\mathbb{E}[\epsilon_{t-2}^2\epsilon_i^2]\\ &= \frac{c_0^3c_1}{T^4}\sum_{t=2}^{T}\sum_{i=1}^{T}\mathbb{E}\left[\frac{F_{X_i,X_i}^{t-1,t}(h_0)}{F_{X_t}(h_0)F_{X_{t-1}}(h_0)}\right]\mathbb{E}[\epsilon_{t-2}^2\epsilon_i^2]\\ &\le \frac{c_0^3c_1}{T^4}\sum_{t=2}^{T}\sum_{i=1}^{T}\mathbb{E}\left[\frac{\psi_1(h_0)f_2(X_i)}{F_{X_t}(h_0)F_{X_{t-1}}(h_0)}\right]\mathbb{E}[\epsilon_{t-2}^2\epsilon_i^2] \quad \text{by assumption } \mathbf{H3}(2)\\ &\le \frac{c_0^3c_1}{T^4}\sum_{t=2}^{T}\sum_{i=1}^{T}\mathbb{E}\left[\frac{f_2(X_i)\phi^2(h_0)}{F_{X_t}(h_0)F_{X_{t-1}}(h_0)}\right]\gamma_\epsilon(0)\gamma_\epsilon(0) \le \frac{C}{T^2} = o(T^{-1})\end{aligned}$$

1.6 $a=b=s=1$ and $g=0$; **1.7** $a=b=s=0$ and $g=1$; **1.8** $a=b=g=1$ and $s=0$; **1.9** $a=g=s=0$ and $b=1$; **1.10** $b=g=s=1$ and $a=0$; **1.11** $b=g=s=0$ and $a=1$; **1.12** $a=g=s=1$ and $b=0$ are similar to **1.5**

1.13 $a = g = 0$ and $b = s = 1$ then the expression is non zero when $i = j+1$ and $t = p+1$ and

$$
\begin{aligned}
(5.8) &= \frac{c_0^2 c_1^2}{T^4} \sum_{t=2}^{T} \sum_{i=1}^{T} \mathbb{E}\left[\frac{\Delta_i^{(0)}(X_t)}{\mathbb{E}\left(\Delta_i^{(0)}(X_t)\right)} \frac{\Delta_{i-1}^{(0)}(X_{t-1})}{\mathbb{E}\left(\Delta_{i-1}^{(0)}(X_{t-1})\right)} \right] \mathbb{E}[\epsilon_{t-2}^2 \epsilon_{i-1}^2] \\
&= \frac{c_0^2 c_1^2}{T^4} \sum_{t=2}^{T} \sum_{i=1}^{T} \mathbb{E}\left[\frac{F_{X_i, X_{i-1}}^{t,t-1}(h_0)}{F_{X_t}(h_0) F_{X_{t-1}}(h_0)} \right] \gamma_\epsilon(0)\gamma_\epsilon(0) \\
&\leq \frac{c_0^2 c_1^2}{T^4} \sum_{t=2}^{T} \sum_{i=1}^{T} \mathbb{E}\left[\frac{\psi_2(h_0) f_3(X_i, X_{i-1})}{F_{X_t}(h_0) F_{X_{t-1}}(h_0)} \right] \gamma_\epsilon(0)\gamma_\epsilon(0) \quad \text{by assumption } \mathbf{H3}(3) \\
&\leq \frac{c_0^2 c_1^2}{T^4} \sum_{t=2}^{T} \sum_{i=1}^{T} \mathbb{E}\left[\frac{f_3(X_i, X_{i-1}) \phi^2(h_0)}{F_{X_t}(h_0) F_{X_{t-1}}(h_0)} \right] \gamma_\epsilon(0)\gamma_\epsilon(0) \leq \frac{C}{T^2} = o(T^{-1})
\end{aligned}
$$

1.14 $a = s = 0$ and $b = g = 1$; **1.15** $a = g = 1$ and $b = s = 0$; **1.16** $a = s = 1$ and $b = g = 0$ are similar to **1.13**.

The other cases are treated in the same manner as above. Consequently, we have

$$
\begin{aligned}
&\mathbb{E}\left(\frac{1}{T} \sum_{t=2}^{T} u_{t-1} \frac{1}{T} \sum_{i=1}^{T} \frac{\Delta_i^{(0)}(X_t)}{\mathbb{E}\left(\Delta_i^{(0)}(X_t)\right)} u_i \right)^2 = O\left(T^{-1}\right) \\
&\left\| \frac{1}{T} \sum_{t=2}^{T} u_{t-1} \frac{1}{T} \sum_{i=1}^{T} \frac{\Delta_i^{(0)}(X_t)}{\mathbb{E}\left(\Delta_i^{(0)}(X_t)\right)} u_i \right\|_2 = O\left(T^{-1/2}\right)
\end{aligned}
$$

Then we have $\frac{1}{T} \sum_{t=2}^{T} u_{t-1} \widehat{V}_t = O_p\left(T^{-1/2}\right)$.

$$
\begin{aligned}
\diamondsuit \quad \left| \frac{1}{T} \sum_{t=2}^{T} u_{t-1} \widehat{B}_t \right| &= \left| \frac{1}{T} \sum_{t=2}^{T} u_{t-1} \frac{1}{T} \sum_{i=1}^{T} \frac{\Delta_i^{(0)}(X_t)}{\frac{1}{T}\sum_{i=1}^{T} \Delta_i^{(0)}(X_t)} \{r(X_t) - r(X_i)\} \right| \\
&\leq \frac{1}{T} \sum_{t=2}^{T} \sum_{i=1}^{T} \frac{1}{T} \left| \frac{\Delta_i^{(0)}(X_t)}{\mathbb{E}\left[\Delta_i^{(0)}(X_t)\right]} u_{t-1} \{r(X_t) - r(X_i)\} \right| \\
&\leq \frac{1}{T} \sum_{t=2}^{T} \sum_{i=1}^{T} \frac{1}{T} \left| \frac{\Delta_i^{(0)}(X_t)}{\mathbb{E}\left[\Delta_i^{(0)}(X_t)\right]} u_{t-1} \right| c\, d(X_t, X_i)^\beta \quad \text{using assumption } \mathbf{H1}(1) \\
&\leq \frac{1}{T} \sum_{t=2}^{T} \sum_{i=1}^{T} \frac{1}{T} \left| \frac{\Delta_i^{(0)}(X_t)}{\mathbb{E}\left[\Delta_i^{(0)}(X_t)\right]} u_{t-1} \right| c\, h_0^\beta
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}\left(\frac{1}{T} \sum_{t=2}^{T} \sum_{i=1}^{T} \frac{1}{T} \left| \frac{\Delta_i^{(0)}(X_t)}{\mathbb{E}\left[\Delta_i^{(0)}(X_t)\right]} u_{t-1} \right| c\, h_0^\beta \right)^2 \\
&\leq c^2\, h_0^{2\beta} \frac{1}{T^2} \mathbb{E}\left(\sum_{t=2}^{T} \sum_{i=1}^{T} \frac{1}{T} \left| \frac{\Delta_i^{(0)}(X_t)}{\mathbb{E}\left[\Delta_i^{(0)}(X_t)\right]} u_{t-1} \right| \right)^2 \\
&\leq \frac{c^2\, h_0^{2\beta}}{T^2} \sum_{a=0}^{1} \sum_{b=0}^{1} \sum_{t=2}^{T} \sum_{p=2}^{T} \sum_{i=1}^{T} \sum_{j=1}^{T} c_a c_b \mathbb{E}\left[\frac{1}{T^2} \frac{\Delta_i^{(0)}(X_t)}{\mathbb{E}\left[\Delta_i^{(0)}(X_t)\right]} \frac{\Delta_j^{(0)}(X_p)}{\mathbb{E}\left[\Delta_j^{(0)}(X_p)\right]} \right] \mathbb{E}[\epsilon_{t-1-a} \epsilon_{p-1-b}]
\end{aligned}
\tag{5.9}
$$

The previous expression is non zero if $t-1-a=p-1-b$ therefore if $t-a=p-b$

1. If $a=b=0$ then the expression is non zero when $t=p$
2. If $a=1$ and $b=0$ then the expression is non zero when $t-1=p$
3. If $a=0$ and $b=1$ then the expression is non zero when $t=p-1$ ($p=t+1$)
4. If $a=b=1$ then $t-1=p-1$ then the expression is non zero when $t=p$

Case 1: $a=b=0$ and $t=p$

$$
\begin{aligned}
(5.9) &\leq \frac{c^2h_0^{2\beta}}{T^2}\sum_{t=2}^{T}\sum_{i=1}^{T}\sum_{j=1}^{T}c_0^2\mathbb{E}\left[\frac{1}{T^2}\frac{\Delta_i^{(0)}(X_t)}{\mathbb{E}\left[\Delta_i^{(0)}(X_t)\right]}\frac{\Delta_j^{(0)}(X_t)}{\mathbb{E}\left[\Delta_j^{(0)}(X_t)\right]}\right]\mathbb{E}[\epsilon_{t-1}\epsilon_{t-1}] \\
&\leq \frac{c^2h_0^{2\beta}}{T^2}\frac{c_0^2}{T^2}\sum_{t=2}^{T}\sum_{i=1}^{T}\sum_{j=1}^{T}\mathbb{E}\left[\frac{F^{i,j}_{X_t,X_t}}{F_{X_t}(h_0)F_{X_t}(h_0)}\right]\gamma_\epsilon(0) \\
&\leq \frac{c^2h_0^{2\beta}}{T^2}\frac{c_0^2}{T^2}\sum_{t=2}^{T}\sum_{i=1}^{T}\sum_{j=1}^{T}\mathbb{E}\left[\frac{\psi_1(h_0)f_2(X_t)}{F_{X_t}(h_0)F_{X_t}(h_0)}\right]\gamma_\epsilon(0) \quad \text{by assumption } \mathbf{H3}(2) \\
&\leq \frac{Ch_0^{2\beta}}{T} = o(h_0^{2\beta})
\end{aligned}
$$

Similarly, we have **Case 2:** $a=1$ and $b=0$ then $p=t-1$

$$
\begin{aligned}
(5.9) &\leq \frac{c^2h_0^{2\beta}}{T^2}c_0c_1\sum_{t=2}^{T}\sum_{i=1}^{T}\sum_{j=1}^{T}\mathbb{E}\left[\frac{1}{T^2}\frac{\Delta_i^{(0)}(X_t)}{\mathbb{E}\left[\Delta_i^{(0)}(X_t)\right]}\frac{\Delta_j^{(0)}(X_{t-1})}{\mathbb{E}\left[\Delta_j^{(0)}(X_{t-1})\right]}\right]\mathbb{E}[\epsilon_{t-2}\epsilon_{t-2}] \\
&\leq \frac{c^2h_0^{2\beta}}{T^2}c_0c_1\frac{1}{T^2}\sum_{t=2}^{T}\sum_{i=1}^{T}\sum_{j=1}^{T}\mathbb{E}\left[\frac{F^{i,j}_{X_t,X_{t-1}}}{F_{X_t}(h_0)F_{X_{t-1}}(h_0)}\right]\gamma_\epsilon(0) \\
&\leq \frac{c^2h_0^{2\beta}}{T^2}c_0c_1\frac{1}{T^2}\sum_{t=2}^{T}\sum_{i=1}^{T}\sum_{j=1}^{T}\mathbb{E}\left[\frac{\psi_2(h_0)f_3(X_t,X_{t-1})}{F_{X_t}(h_0)F_{X_{t-1}}(h_0)}\right]\gamma_\epsilon(0) \text{ by assumption } \mathbf{H3}(3) \\
&= o(h_0^{2\beta})
\end{aligned}
$$

Case 3: $a=0$ and $b=1$ then $p=t+1$ it is like to Case 2.
Case 4: $a=b=1$ then $t-1=p-1$ it is like to Case 1.

In conclusion, $(5.9)=O(h_0^{2\beta})$. Then, $\left(\mathbb{E}\left(\frac{1}{T}\sum_{t=2}^{T}u_{t-1}\widehat{B}_t\right)^2\right)^{1/2}=o(h_0^{\beta})$

$$
\left|\frac{1}{T}\sum_{t=2}^{T}(-u_{t-1}\widehat{B}_t-u_{t-1}\widehat{V}_t-\widehat{B}_{t-1}u_t-\widehat{V}_{t-1}u_t)\right| = O_p\left(h_0^{\beta}+T^{-1/2}\right)
$$

We deduce from Part a and Part b that $|\widehat{a}_{12}-\overline{a}_{12}|=O_p\left(\frac{\log T}{T\phi(h_0)}+h_0^{\beta}+T^{-1/2}\right)$.

- $$
\begin{aligned}
|\overline{a}_{11}-\widehat{a}_{11}| &= \left|\frac{1}{T}\sum_{t=2}^{T}(u_{t-1}^2-\hat{u}_{t-1}^2)\right| \\
&= \left|\frac{1}{T}\sum_{t=2}^{T}(u_{t-1}^2-(u_{t-1}-\widehat{B}_{t-1}-\widehat{V}_{t-1})^2)\right| \\
&= \left|\frac{1}{T}\sum_{t=2}^{T}(2u_{t-1}\widehat{B}_{t-1}+2u_{t-1}\widehat{V}_{t-1}-2\widehat{B}_{t-1}\widehat{V}_{t-1}-\widehat{B}_{t-1}^2-\widehat{V}_{t-1}^2)\right| \\
&= O_p\left(h_0^{\beta}\right)+O_p\left(T^{-1/2}\right)+O_p\left(\frac{\log T}{T\phi(h_0)}\right)
\end{aligned}
$$

since we show that $|\widehat{B}_t| = O_p\left(h_0^{\beta}\right)$, $|\widehat{V}_t| = O_p\left(\frac{\log T}{T\phi(h_0)}\right)$, $\frac{1}{T}\sum_{t=2}^{T} u_{t-1}\widehat{V}_t = O_p\left(T^{-1/2}\right)$ and $\frac{1}{T}\sum_{t=2}^{T} u_{t-1}\widehat{B}_t = O_p\left(h_0^{\beta}\right)$. Then, $\widehat{a}_{11} \longrightarrow \overline{a}_{11}$ in probability as $T \to \infty$ and therefore we have

- $(\overline{a}_{11}) \to \mathbb{E}(\overline{a}_{11}) = \dfrac{1}{T}\displaystyle\sum_{t=2}^{T}\mathbb{E}(u_{t-1}^2) = \dfrac{1}{T}\sum_{t=2}^{T}\gamma_u(0) = \gamma_u(0)$

Then, $|\widehat{a}_1 - \overline{a}_1| = O_p\left(h_0^{\beta} + T^{-1/2} + \frac{\log T}{T\phi(h_0)}\right)$.

$$\begin{aligned}
\|\widetilde{r}_{21B}(x)\|_2 &= \frac{1}{T\mathbb{E}\left[\Delta_t^{(1)}(x)\right]}\left\|\sum_{t=1}^{T}(\hat{a}_1 - \overline{a}_1)u_{t-1}\Delta_t^{(1)}(x)\right\|_2 \\
&\leq \frac{C}{TF_x(h_1)}\sum_{t=1}^{T}\left\|(\hat{a}_1 - \overline{a}_1)u_{t-1}\Delta_t^{(1)}(x)\right\|_2 \\
&\leq \frac{C}{TF_x(h_1)}\sum_{t=1}^{T}|\hat{a}_1 - \overline{a}_1|\left\|u_{t-1}\Delta_t^{(1)}(x)\right\|_2
\end{aligned}$$

Notice that $\{u_t\}$ is independent of $\{X_t\}$, then $\left\|u_{t-1}\Delta_t^{(1)}(x)\right\|_2 \leq C(\gamma_u(0)\phi(h_1))^{1/2}$. We have

$$\begin{aligned}
\widetilde{r}_{21B}(x) &= O_p\left(h_0^{\beta}\phi(h_1)^{-1/2}\right) + O_p\left((T\phi(h_1))^{-1/2}\right) + O_p\left(\frac{\log T}{T\phi(h_0)\phi(h_1)^{1/2}}\right) \\
&= O_p\left(\sqrt{\frac{1}{T\phi(h_1)}}\right), \text{by assumption } \mathbf{H5}.
\end{aligned}$$

In conclusion, $Q_{T1} = O_p\left(\sqrt{\frac{1}{T\phi(h_1)}}\right)$. ■

Proof of Lemma 5.11

Recall that

$$\widehat{B}_t = -\frac{1}{T}\sum_{i=1}^{T} W_i(X_t)\{r(X_t) - r(X_i)\} \qquad \text{and} \qquad \widehat{V}_t = \frac{1}{T}\sum_{i=1}^{T} W_i(X_t)u_i$$

We know that, $\widehat{r}(X_t) - r(X_t) = \widehat{B}_t + \widehat{V}_t$ and $Q_{T_2} = \frac{\widetilde{r}_{22}(x)}{\widetilde{r}_1(x)}$. Then,

$$\widetilde{r}_{22}(x) = \frac{1}{T\mathbb{E}\left[\Delta_t^{(1)}(x)\right]}\sum_{t=1}^{T} a_1\widehat{B}_{t-1}\Delta_t^{(1)}(x) + \frac{1}{T\mathbb{E}\left[\Delta_t^{(1)}(x)\right]}\sum_{t=1}^{T} a_1\widehat{V}_{t-1}\Delta_t^{(1)}(x)$$

In Lemma 5.10, we have calculated $\widehat{B}_t$ and $\widehat{V}_t$ and we have found that $|\widehat{V}_t| = O_p\left(\frac{\log T}{T\phi(h_0)}\right)$ and $|\widehat{B}_t| = O_p(h_0^{\beta})$ and we know that $0 \leq |a_1| \leq 1$. Then,

$$\widetilde{r}_{22}(x) = O_p\left(h_0^{\beta} + \frac{\log T}{T\phi(h_0)}\right) = O_p(h_0^{\beta}) + o_p\left(\left(\frac{1}{T\phi(h_1)}\right)^{1/2}\right).$$

This ends the proof. ■

Proof of Lemma 5.12

Recall that

$$\begin{aligned}\widetilde{r}_{23}(x) &= \frac{1}{T\mathbb{E}\left[\Delta_t^{(1)}(x)\right]}\sum_{t=1}^{T}(\widehat{a}_1 - a_1)\widehat{B}_{t-1}\Delta_t^{(1)}(x) + \frac{1}{T\mathbb{E}\left[\Delta_t^{(1)}(x)\right]}\sum_{t=1}^{T}(\widehat{a}_1 - a_1)\widehat{V}_{t-1}\Delta_t^{(1)}(x)\\ &= \widetilde{r}_{231}(x) + \widetilde{r}_{232}(x)\end{aligned}$$

We proved in Lemma 5.10 that $|\overline{a}_1 - a_1| = O_p\left(T^{-1/2}\right)$, $|\widehat{a}_1 - \overline{a}_1| = O_p\left(h_0^{\beta}\right) + O_p\left(T^{-1/2}\right) + O_p\left(\frac{\log T}{T\phi(h_0)}\right)$. Then $(\widehat{a}_1 - a_1) = (\widehat{a}_1 - \overline{a}_1) + (\overline{a}_1 - a_1) = O_p\left(h_0^{\beta}\right) + O_p\left(T^{-1/2}\right) + O_p\left(\frac{\log T}{T\phi(h_0)}\right)$. Moreover, we have $|\widehat{B}_{t-1}| = O_p(h_0^{\beta})$ and $\frac{1}{T\mathbb{E}\left[\Delta_t^{(1)}(x)\right]}\sum_{t=1}^{T}\Delta_t^{(1)}(x) \to 1$. Therefore,

$$\begin{aligned}\widetilde{r}_{231}(x) &= \frac{1}{T\mathbb{E}\left[\Delta_t^{(1)}(x)\right]}\sum_{t=1}^{T}(\widehat{a}_1 - a_1)\widehat{B}_{t-1}\Delta_t^{(1)}(x)\\ &= O_p\left(h_0^{\beta}\right) + O_p\left(T^{-1/2}\right) + O_p\left(\frac{\log T}{T\phi(h_0)}\right)\end{aligned}$$

Similarly, with $|\widehat{V}_{t-1}| = O_p\left(\frac{\log T}{T\phi(h_0)}\right)$, we have

$$\begin{aligned}\widetilde{r}_{232}(x) &= \frac{1}{T\mathbb{E}\left[\Delta_t^{(1)}(x)\right]}\sum_{t=1}^{T}(\widehat{a}_1 - a_1)\widehat{V}_{t-1}\Delta_t^{(1)}(x)\\ &= O_p\left(h_0^{\beta}\right) + O_p\left(T^{-1/2}\right) + O_p\left(\frac{\log T}{T\phi(h_0)}\right)\end{aligned}$$

Therefore,

$$\begin{aligned}\widetilde{r}_{23}(x) &= O_p\left(h_0^{\beta}\right) + O_p\left(T^{-1/2}\right) + O_p\left(\frac{\log T}{T\phi(h_0)}\right)\\ &= O_p\left(h_0^{\beta}\right) + o_p\left(\left(\frac{1}{T\phi(h_1)}\right)^{1/2}\right).\end{aligned}$$

This ends the proof. ■

Proof of Theorem 5.7

The proof is similar to that of Theorem 5.6 since $\widetilde{r}(x) = \overline{r}(x) - Q_{T_1} + Q_{T_2} + Q_{T_3}$ with $Q_{T_l} = \frac{\widetilde{r}_{2l}(x)}{\widetilde{r}_1(x)}$, for $l = 1, 2, 3$. It then follows directly from Proposition 1, Lemmas 5.10, 5.11 and 5.12, which are key results for proof of this theorem. ■

Part III

Applications

Chapter 6

Application of the kernel density and mode estimations to clustering

Contents

Résumé en français

Dans ce chapitre, nous appliquons l'estimateur de la fonction de densité spatiale étudié au chapitre 2, et du mode associé, à la classification non supervisée, d'une part, sur des données simulées et, d'autre part, sur des données environnementales relatives aux moussons d'Asie. En effet, à la section 2.3.2, nous avons présenté une méthode de classification descendante hiérarchique basée sur l'estimation du mode spatial. Ce dernier est déduit de l'estimation de la fonction de densité spatiale. Ces applications montrent l'importance de tenir compte de la dépendance spatiale en particulier dans le cadre de la classification considérée.

6.1 Introduction

In this chapter, we focus on the importance of taking into account the spatial dependence toward some simulation studies and a real data application in the case of clustering. We first simulate some spatial random fields located in three different regions, with piecewise stationarity. Then, we estimate the density of this field at some sites using the observations in these three regions. After that, we perform the clustering procedure, which should be able to detect the spatial homogeneity on each region and heterogeneity between regions. We end this part by applying our procedure to a real data set where spatial homogeneity and heterogeneity should be detected. Along these two applications, the spatial density is computed using the Epanechnikov kernel for K_1 and the indicator function of $[0, 1]$ for K_2. We deal here with the case where the spatial dimension is $N = 2$. In the following, we compute the density estimate $f_{\mathbf{n}}$ at any point of $O_{\mathbf{n}}$ using the other observations. Note that all notations considered in this chapter are the same as in chapter 2.

Recall that the spatial density estimation of the margins $X_{\mathbf{i}}$, of a discretely indexed spatial process, $(X_{\mathbf{i}}, \mathbf{i} \in \mathbb{Z}^N)$ where $\mathbb{Z}^N$ is the integer lattice points, is given by

$$f_{\mathbf{n}}(x_{\mathbf{j}}) = \frac{1}{\widehat{\mathbf{n}} b_{\mathbf{n}}^d \rho_{\mathbf{n}}^N} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i} - \mathbf{j}\|),$$

for each fixed (not random) observation $x_{\mathbf{j}} \in \mathbb{R}^d$ located at $\mathbf{j} \in \mathcal{I}_{\mathbf{n}}$, where $\mathcal{I}_{\mathbf{n}} := \{\mathbf{i} \in (\mathbb{N}^*)^N, 1 \leq i_k \leq n_k, k = 1, \ldots, N\}$ is the set of sites of observations. Moreover, we define $K_{2,\rho_{\mathbf{n}}}(\|\mathbf{i} - \mathbf{j}\|) = C_{\mathbf{nj}} K_2\left(\frac{\|t_{\mathbf{i}} - t_{\mathbf{j}}\|}{\rho_{\mathbf{n}}}\right)$ where $t_{\mathbf{i}} = \frac{\mathbf{i}}{\mathbf{n}} =: \left(\frac{i_1}{n}, \ldots, \frac{i_N}{n}\right)$, $C_{\mathbf{nj}} > 0$ is a normalized constant eventually equal to one, K_1 and K_2 are kernels respectively defined on $\mathbb{R}^d$ and $\mathbb{R}$. In addition, an estimation of the mode is obtained from the density estimate as follows:

$$\widehat{\omega} = \arg \sup_{\substack{x_{\mathbf{i}} \in R \\ \mathbf{i} \in V_R}} f_{\mathbf{n}}(x_{\mathbf{i}}).$$

with R be a set in $\mathbb{R}^d$ supposed to be a compact and $V_R \subseteq \mathbb{R}^N$ be the finite set of sites $\mathbf{j}$ contained in $\mathcal{I}_{\mathbf{n}}$ such that the corresponding $x_{\mathbf{j}}$ are in R.

We now specify some elements that will be used in the following.

- The *bandwiths selection*. Basically, our method suggests to first select the bandwidths. Here, since we deal with a classification procedure (presented in Section 2.3.2), we aim to detect largest heterogeneity as possible. This is why we prefer to use the optimal bandwidths obtained by maximizing the empirical entropy of $f_{\mathbf{n}}$ defined by $-\int f_{\mathbf{n}}(x) \log\left(f_{\mathbf{n}}(x)\right) dx$ (as proposed in Ferraty and Vieu (2006)).
- For the *error measurement*, we evaluated it using either:
  - ☐ the mean integrated squared error (MISE):

$$\mathbb{E}\left(\int \left(f_{\mathbf{n}}(x) - f(x)\right)^2 dx\right)$$

    which will be approximated as an average of the integrated squared error over some subsamples.
  - ☐ or the Kullback-Leibler (KL) divergence measure:

$$\mathbb{E}\left(\log\left(f_{\mathbf{n}}(X)\right)\right) - \mathbb{E}\left(\log\left(f_{\mathbf{n}}^{(-\mathbf{i})}(X)\right)\right)$$

    which will be estimated by $KL = \frac{1}{\widehat{\mathbf{n}}} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} \left(\log\left(f_{\mathbf{n}}(X_{\mathbf{i}})\right) - \log\left(f_{\mathbf{n}}^{(-\mathbf{i})}(X_{\mathbf{i}})\right)\right)$ where $f_{\mathbf{n}}^{(-\mathbf{i})}(X_{\mathbf{i}})$ is the leave-one-out estimator of $f(X_{\mathbf{i}})$.

6.2 Simulations

The simulated data are located on the area $\mathcal{I}_{(26,26)} = \{(i,j), 1 \leq i, j \leq 26\}$ and let $(X_{(i,j)})$ be the field of interest. In the following, we denote by $GRF(m, \sigma^2, s)$ any stationary Gaussian Random Field with mean m and the covariance function defined by $C(h) = \sigma^2 \exp\left(-\left(\frac{\|h\|}{s}\right)^2\right)$, $h \in \mathbb{R}^2$ and $s > 0$.

In order to illustrate the fact that our method works for multidimensional random variables, we consider the case where $X_{(i,j)}$'s belong to $\mathbb{R}^d$ with $d = 5$ (naturally the procedure is valid if $d = 1$). Then, the data set $(X_{(i,j)}, (i,j) \in \mathcal{I}_{(26,26)})$ is partitioned

in three clusters of observations such that each component of $X_{(i,j)}$ denoted as $X_{(i,j)}^{(p)}$, $p = 1, \ldots, 5$, is:

$$X_{(i,j)}^{(p)} = \begin{cases} F_{(i,j)}(t_p - 0.5)^3 + (\varepsilon_{(i,j)})/4 & \text{for } (i,\ j) \in R_1 \text{ (Group 1)} \\ Z_{(i,j)}(t_p) + a_1 & \text{for } (i,\ j) \in R_2 \text{ (Group 2)} \\ F_{(i,j)} \cos(2\pi t_p)^5 + a_2 + \varepsilon_{(i,j)} & \text{for } (i,\ j) \in R_3 \text{ (Group 3)} \end{cases}$$

where $t_1 = 0$, $t_2 = 2.5$, $t_3 = 5$, $t_4 = 7.5$, $t_5 = 10$, $\varepsilon = GRF(0, 2, 5)$ as well as $(Z_{(i,j)}(u))$ is a Brownian motion, $a_1 = 10$ and $a_2 = 30$. In this case, called **Case 1**, the field $F_{(i,j)}$ is built such that the five component fields of $\left(X_{(i,j)}^{(p)}\right)$, $(i,j) \in \mathcal{I}_{(26,26)}$, and $p = 1, \ldots, 5$ have the shapes presented in Figure 6.1. The areas R_1, R_2, and R_3 are disjoint sets of sites represented respectively in red, blue and green.

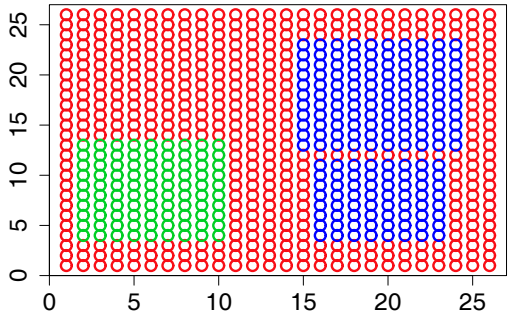


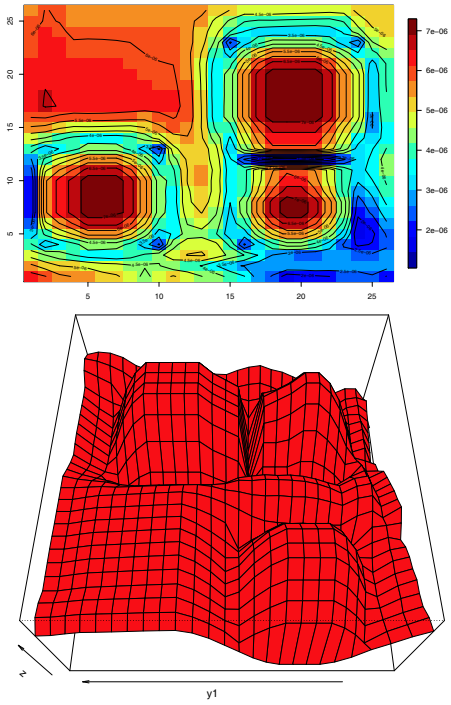
Figure 6.1 – The three regions considered in Case 1
Group 1 in red, Group 2 in blue and Group 3 in green

We first perform the spatial density estimation. The bandwidths selection rule computed over [0.08, 1] and [10, 30] (here $S(b)$ is a regular sequence of 30 values in [0.08, 1] and $S(\rho)$ is also a regular sequence of 30 values in [10, 30] for $b_{\mathbf{n}}$) has led to $\rho_{\mathbf{n},\,opt} \simeq 0.15$, $b_{\mathbf{n},opt} \simeq 19$ and $KL \simeq 0.002$. That is, we have computed $f_{\mathbf{n},opt}(x_{\mathbf{i}})$, for each $\mathbf{i} \in O_{\mathbf{n}}$. These values identified with $\bar{f}_{\mathbf{n}}(\mathbf{i})$ are displayed in Figure 6.2a, representing the density in the area of observations. In fact, when $N = 2$, a 3D plot of $f_{\mathbf{n}}$ can be obtained even if it is defined on $\mathbb{R}^d$, $d > 3$. It suffices to identify $f_{\mathbf{n}}$ with the function $\overline{f}_{\mathbf{n}}$ defined by: $\bar{f}_{\mathbf{n}}(\mathbf{i}) = f_{\mathbf{n}}(x_{\mathbf{i}})$, $x_{\mathbf{i}} \in \mathbb{R}^d$, $\mathbf{i} \in \mathbb{Z}^N$. Figure 6.2a shows that the density estimation detects four spatial distributions with three which are similarly distributed. Note that both graphics in Figure 6.2a are two representations of the same function. Figure 6.2b shows that the spatial clustering procedure 2.3.2 better detects the three groups than the density estimation. One can observe that, except for a few points, the clustering procedure efficiently detects the three sets; the well-classified rate is $92,8\%$.

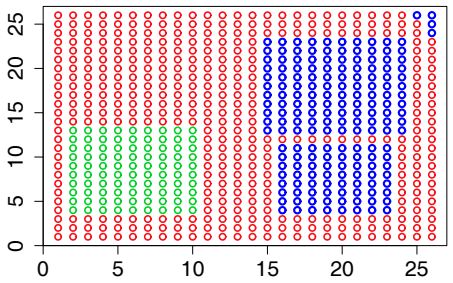
Having tested our method on this relatively simple case, we moved on to other situations. Then, we studied the behavior of our procedure over three other scenarios, called Cases 2, 3 and 4 which are detailed below. We first study the behavior of the clustering algorithm. In each case, the performance of the procedure is measured in well-classified rate meaning. For that purpose, we have done 100 simulations of each situation and computed the distribution of the well classified. The results are summarized in Table 6.1 and interpreted below. We notice that the three cases differ by the gap between the values taken by the subfields. This contrast is controlled by the values of a_1 and a_2. In fact, the lower a_1 and a_2 are, the closer the subfields.

Case 2 : $a_1 = 10$ and $a_2 = 30$.

In this situation, presented in Figure 6.4, the subfields are the same as in Case 1 but



(a) Two representations of the estimated spatial density



(b) The spatial clusters obtained using spatial clustering procedure

Figure 6.2 – The results of the procedure based on simulated data in Case 1

the spatial regions change (see Figure 6.3). Our aim here is to test whether or not the procedure is able to detect non-rectangular spatial regions, based on the same spatial random fields as in Case 1.
As can be observed in the first column of Table 6.1, the well-classified rate, over the 100 simulations, is distributed between 58.98 and 89.64%.

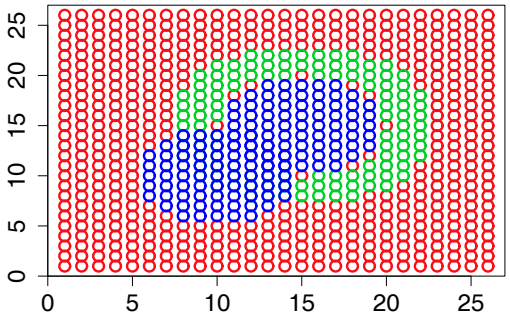
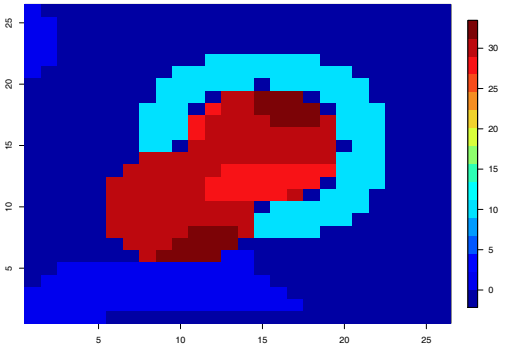


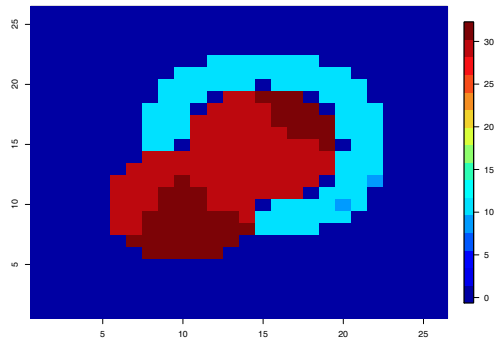
Figure 6.3 – The three regions considered in Cases 2, 3 and 4
Group 1 in red, Group 2 in blue and Group 3 in green

Case 3 : $a_1 = 0$ and $a_2 = 0$.

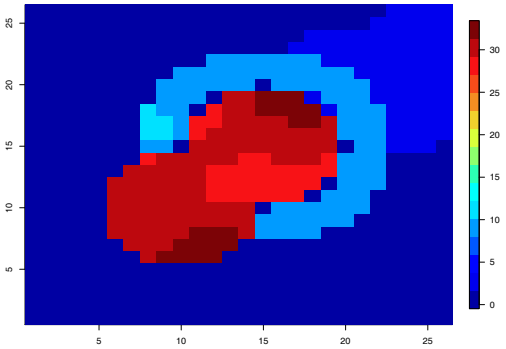
It can be seen from Figure 6.5 that, in this case, the differences among the three regions are less obvious. This case is motivated by the fact that we want to know how the procedure behaves in such a situation where the frontiers between the regions is



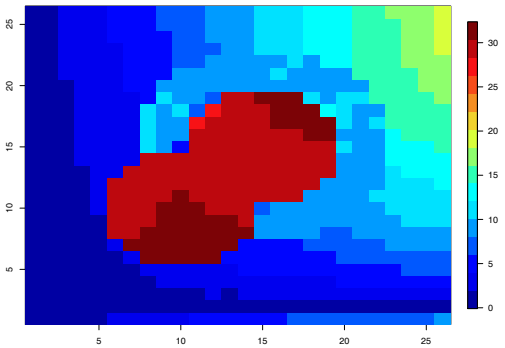
(a) Plot of the observations of the field $\left(X^{(1)}_{(i,j)}\right)$



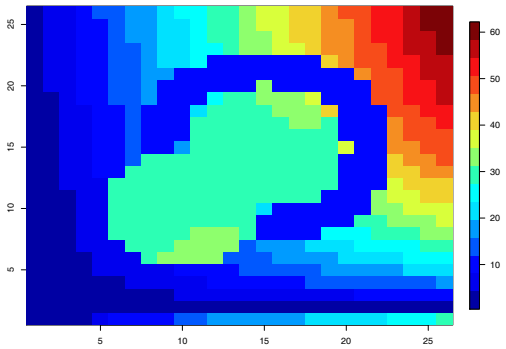
(b) Plot of the observations of the field $\left(X^{(2)}_{(i,j)}\right)$



(c) Plot of the observations of the field $\left(X^{(3)}_{(i,j)}\right)$



(d) Plot of the observations of the field $\left(X^{(4)}_{(i,j)}\right)$



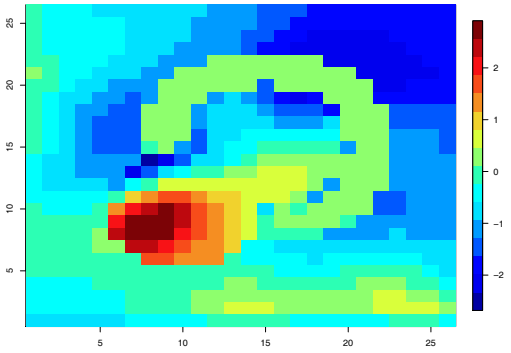
(e) Plot of the observations of the field $\left(X^{(5)}_{(i,j)}\right)$

Figure 6.4 – A simulation of the random fields in Case 2.

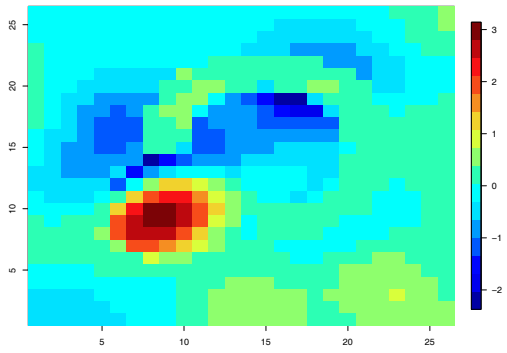
not obvious. Due to the lack of real frontiers between the subfield in many parts of the image, one can expect that our procedure will not detect the three regions. This is what the second column of Table 6.1 shows where well-classified rate that varies between 24.11 and 67.75%, over the 100 simulations is the worst compared to other cases.

Case 4 : $a_1 = 30$ and $a_2 = 60$.

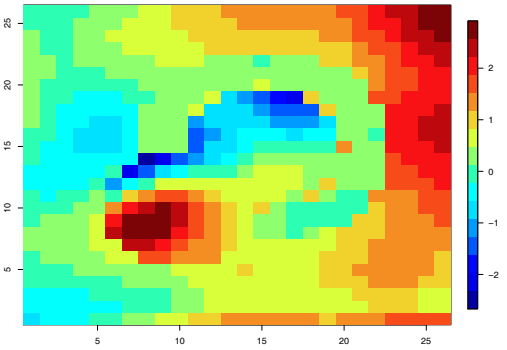
In this case presented in Figure 6.6, it is easier to distinguish the difference between the three regions. We can then expect that our clustering procedure should detect



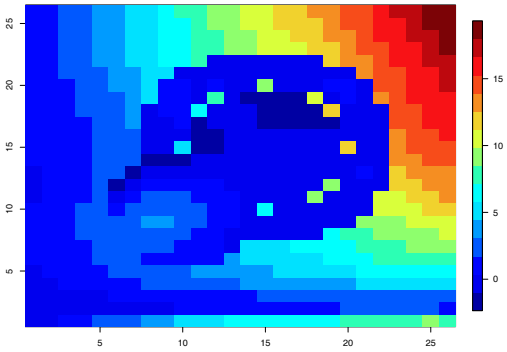
(a) Plot of the observations of the field $\left(X^{(1)}_{(i,j)}\right)$



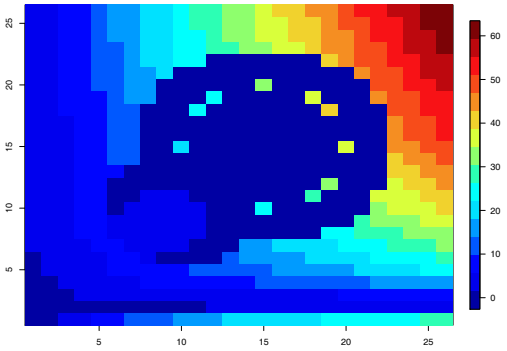
(b) Plot of the observations of the field $\left(X^{(2)}_{(i,j)}\right)$



(c) Plot of the observations of the field $\left(X^{(3)}_{(i,j)}\right)$



(d) Plot of the observations of the field $\left(X^{(4)}_{(i,j)}\right)$

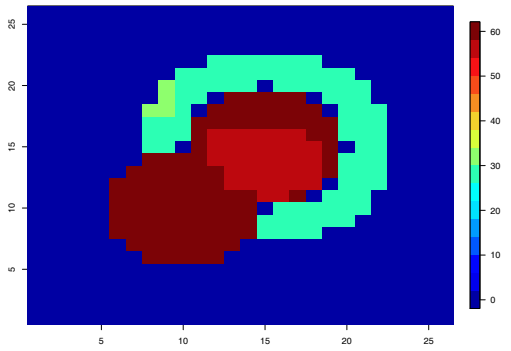


(e) Plot of the observations of the field $\left(X^{(5)}_{(i,j)}\right)$

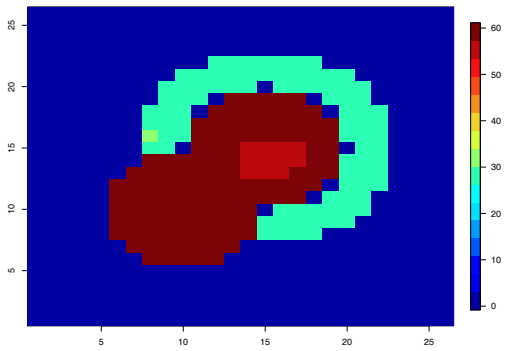
Figure 6.5 – A simulation of the random fields in Case 3.

these regions. In fact, the well-classified rate over the 100 simulations is the best compared to the two previous cases since it takes values between 61.39 and 89.79%. It is presented in the third column of Table 6.1.

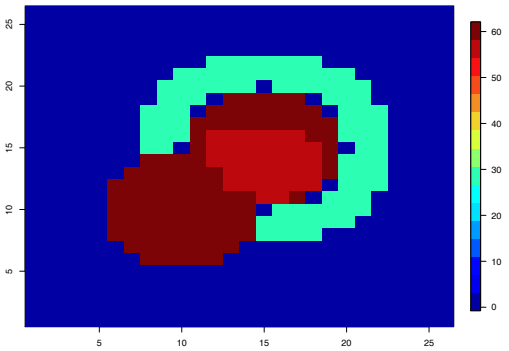
We notice that the last two lines of Table 6.1 provide, for each case, an example of the optimal bandwidths, selected among a regular sequence of 30 values in [0.2, 1] for $\rho_{\mathbf{n}}$ and a regular sequence of 30 values in [10, 30] for $b_{\mathbf{n}}$, still maximizing the estimation of the entropy.



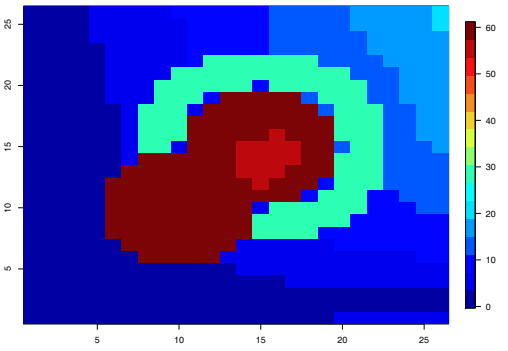
(a) Plot of the observations of the field $\left(X^{(1)}_{(i,j)}\right)$



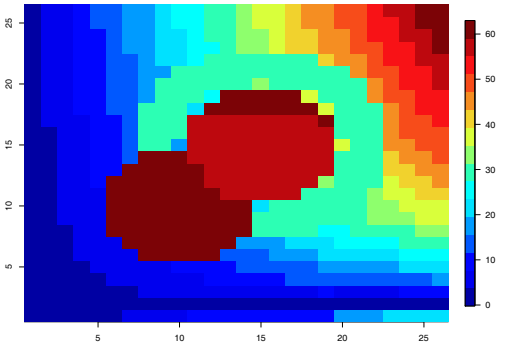
(b) Plot of the observations of the field $\left(X^{(2)}_{(i,j)}\right)$



(c) Plot of the observations of the field $\left(X^{(3)}_{(i,j)}\right)$



(d) Plot of the observations of the field $\left(X^{(4)}_{(i,j)}\right)$



(e) Plot of the observations of the field $\left(X^{(5)}_{(i,j)}\right)$

Figure 6.6 – A simulation of the random fields in Case 4.

Evaluation of the performance of the kernel density estimation

Besides the classification performance, we were also interested in calibrating the performance of the density estimator. For this purpose, we measure the quality of estimation using the approximation of the MISE: $MISE^S = \frac{1}{S}\sum_{j=1}^{S}\int\left(f_{\mathbf{n}}(x) - f_{\mathbf{n}}^{(-i)}(x)\right)^2 dx$ computed from $S = 100$ subsamples. Each subsample is obtained by sampling without replacement B observations among the $\hat{\mathbf{n}} = 26 \times 26$ former observations.

Then, using the same sets of bandwidths as before, we selected $(b_{\mathbf{n},opt},\ \rho_{\mathbf{n},opt})$ which minimizes the $MISE^S$. We notice that $MISE^S$ were computed with respect to subsamples

	Case 2 $a_1 = 10$ - $a_2 = 30$	Case 3 $a_1 = 0$ - $a_2 = 0$	Case 4 $a_1 = 30$ - $a_2 = 60$
Minimum	0.5898	0.2411	0.6139
1st Quartile	0.7139	0.4357	0.8802
Median	0.8831	0.4822	0.8846
3rd Quartile	0.8920	0.5237	0.8935
Maximum	0.8964	0.6775	0.8979
$b_{\mathbf{n},opt}$	15.86	17	16.55
$\rho_{\mathbf{n},opt}$	0.67	0.73	0.72

Table 6.1 – Summary of the distributions of the well-classified rate, based on 100 simulations

with the same length B. We also present the corresponding values of $KL^S = \frac{1}{S}\sum_{i=1}^{S} KL_i^B$ which is the mean of the KL over the S subsamples with length B.

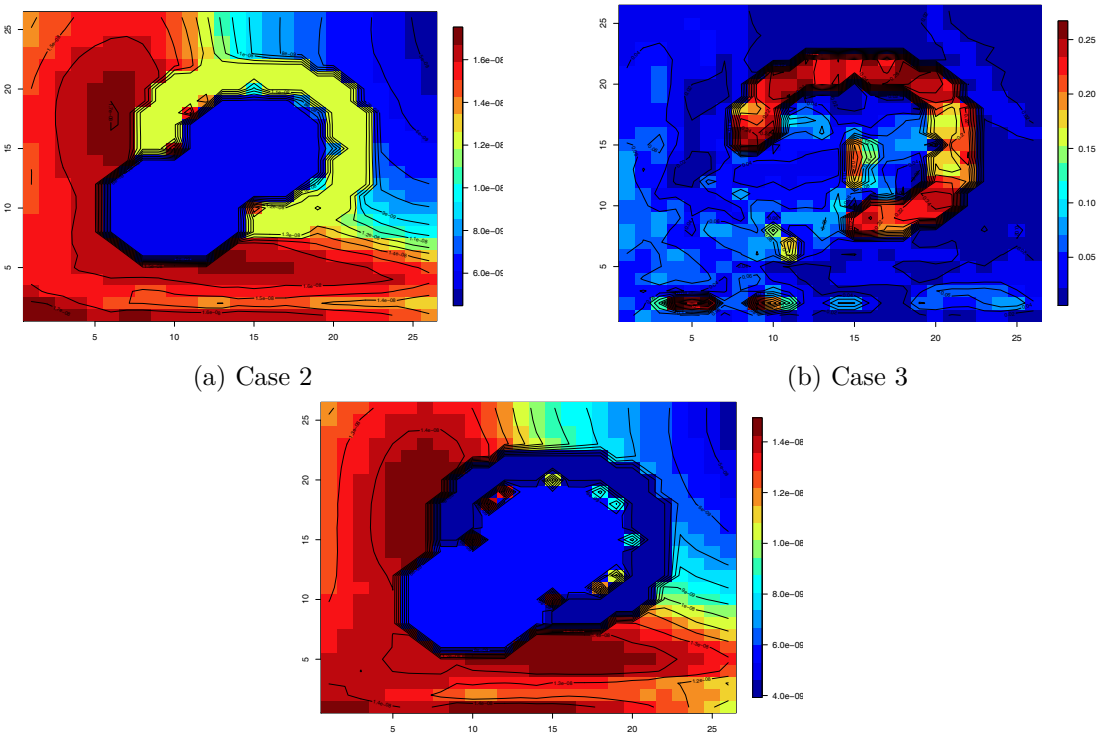
The results are given in Table 6.2. Since the subsamples are not the same from one case to the other, we just compare the behavior of the estimator when B varies. One can observe that $MISE^S$ (or the KL^S) is the worst when estimating $B = 50$ observations to estimate the density. However, the kernel density estimate becomes more efficient when B increases, as the errors ($MISE^S$ and KL^S) decrease to reach their smallest value for $B = 500$. The gap of the error from $B = 50$ to $B = 100$ is obvious. Figure 6.7 shows an example of the estimated density in each case. As expected, the lack of frontiers in Case 3 can be observed here, which explains the bad results of the clustering procedure in Case 3.

B		Case 2 $a_1 = 10$ - $a_2 = 30$	Case 3 $a_1 = 0$ - $a_2 = 0$	Case 4 $a_1 = 30$ - $a_2 = 60$
50	$\min MISE^S$	9×10^{-9}	1.32×10^{-9}	1.75×10^{-6}
	$\min KL^S$	0.02	0.02	0.02
	$b_{\mathbf{n},opt}$	16.9	17	16.55
	$\rho_{\mathbf{n},opt}$	0.75	0.73	0.72
100	$\min MISE^S$	1.79×10^{-9}	8.47×10^{-10}	5.64×10^{-10}
	$\min KL^S$	8.8×10^{-5}	3.7×10^{-5}	2×10^{-4}
	$b_{\mathbf{n},opt}$	13.79	15.86	16.21
	$\rho_{\mathbf{n},opt}$	0.58	0.66	0.69
200	$\min MISE^S$	2.13×10^{-10}	1.86×10^{-10}	2.97×10^{-10}
	$\min KL^S$	5.9×10^{-5}	8×10^{-6}	1.5×10^{-4}
	$b_{\mathbf{n},opt}$	16.55	16.21	15.51
	$\rho_{\mathbf{n},opt}$	0.72	0.7	0.64
500	$\min MISE^S$	2.67×10^{-11}	5.9×10^{-11}	1.56×10^{-11}
	$\min KL^S$	1×10^{-5}	7.8×10^{-6}	1.6×10^{-5}
	$b_{\mathbf{n},opt}$	16.21	15.52	15.86
	$\rho_{\mathbf{n},opt}$	0.7	0.64	0.67

Table 6.2 – Estimation of the empirical $MISE$ and KL based on B sub-samples

Remark 6.1.

- At the bandwidth selection step, the calculation time can be long, depending on the length of the bandwidth-sets.
- Many other situations could be considered such as testing our method in Case 1 or other scenarios with different sample sizes. We have done this in Case 1 and others and obtain similar conclusions as in the other cases.



(a) Case 2

(b) Case 3

(c) Case 4

Figure 6.7 – Examples of the density estimation

- In Table 6.2, in Case 4, the corresponding $MISE^S$ is equal to 1.56×10^{-11} which is lower than in the two other cases. For KL^S, the lowest value is obtained in Case 3. We notice that according to the simulations, the order of the best KL^S or $MISE^S$ can change from one case to another. Basically contrary to the clustering where we were comparing the rate of convergence, the kernel density estimate does not give the same results. In fact, since we have artificially constructed the spatial regions, the procedure is unable to detect the underlying distribution if its frontiers do not coincide with those of the artificial spatial region. That is, one could get a bad well-classified rate but good error measurement based on the same observations.

We now study the behavior of our procedures in a setting of real data application.

6.3 Application to the Monsoon Asia Drought Atlas (MADA)

Description of the MADA dataset

Monsoon failures, megadroughts and extreme flooding events are of critical importance to human populations and ecosystems and have been affecting the agrarian population of Asia over the past millennia. The Asian monsoon system affects more than half of humanity worldwide. Therefore, it is important to study the spatio-temporal variability of the Asian monsoon system.

Here, we consider the MADA spatial reconstruction data over the past millennium, presented in Cook et al. (2010). This data set shows, for example, the occurrence and severity of previously unknown monsoon megadroughts and their close linkages to large-scale patterns of tropical Indo-Pacific sea surface temperatures. We refer to Cook et al.

(2010) for more details on this data set. These authors reconstruct the seasonalized Palmer Drought Severity Index (PDSI) for the summer (June-July-August) monsoon season, using a well-known gridded measure of relative drought and wetness for the globe's land areas. The observations are 534 time points reconstructed on a 2.5×2.5 grid, and the total time period covered equals $1300 - 2005$. The area of study is presented in Figure 6.8 where one can also see an example of variation of the PDSI observed in 2005. In fact, it represents the PDSI observations for the summer monsoon season over the MADA domain: India, East Asia and south into northern Australia. Red (respectively blue) areas are dry (respectively wet) in 2005.

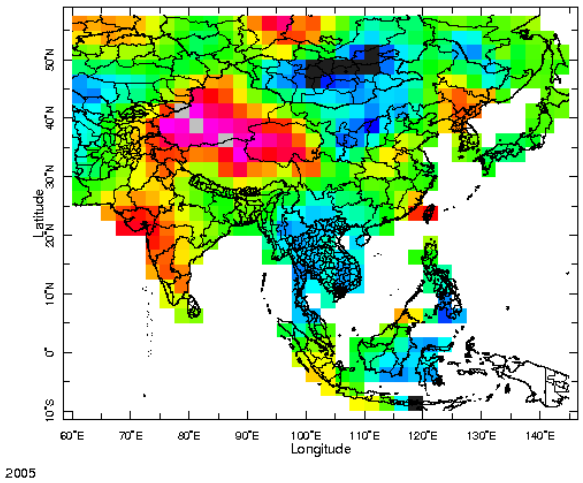


Figure 6.8 – The Palmer Drought Severity Index (PDSI) map in MADA domain for the year 2005

We want to use our method to quantify the spatio-temporal variation in the distribution of PDSI. In view of the large period (705 years), we cannot present all the results here. So we have chosen several dates that will be used to illustrate the use of our method both in the case where the $X_{(i,j)}$ is univariate and in the case where it is multivariate. We first formatted our procedure to evaluate the variation in spatial distribution for several years that are 1500, 1998, and 2005. Then, we studied the variation in spatial distribution using all the observations of a 10-year period $1996 - 2005$. These observations then have values in $\mathbb{R}^{10}$. Naturally, this approach can be used for smaller or larger periods between other observation dates.

The observations of the PDSI in the period of interest are presented in Figure 6.9, and some information about the spatial distribution of the PDSI, fitted empirical semi-variograms (diagonal) and cross semi-variograms (off-diagonal), is presented in Figure 6.10. The latter allows us to have an idea of the spatial dependence structure (in correlation meaning) for the years 1997 and $2002-2005$. To ease the presentation, the semi-variograms corresponding to the other years are removed. We turn now to the application of our procedure to the PDSI distribution.

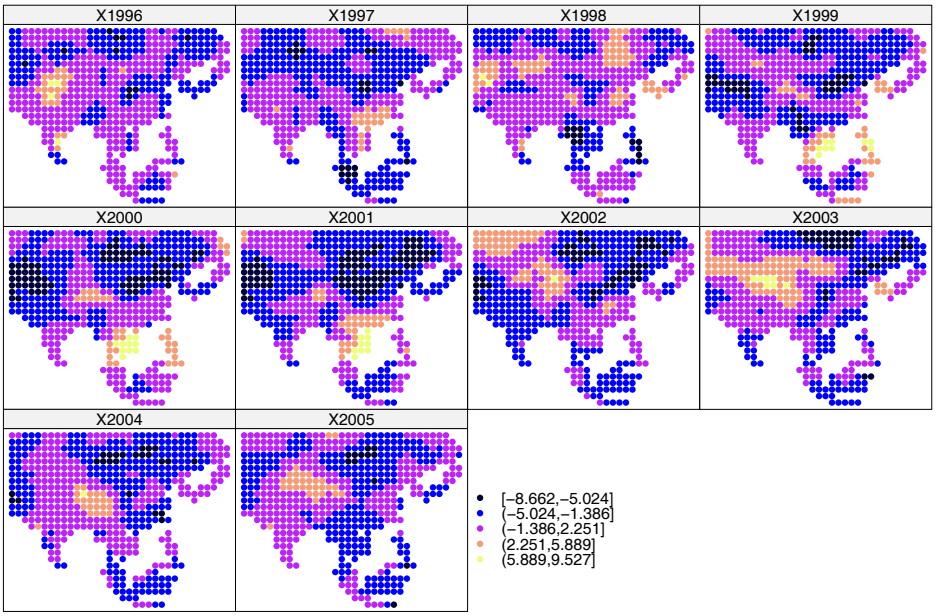


Figure 6.9 – The observations of the Palmer Drought Severity Index (PDSI) in MADA domain for the period 1996 – 2005

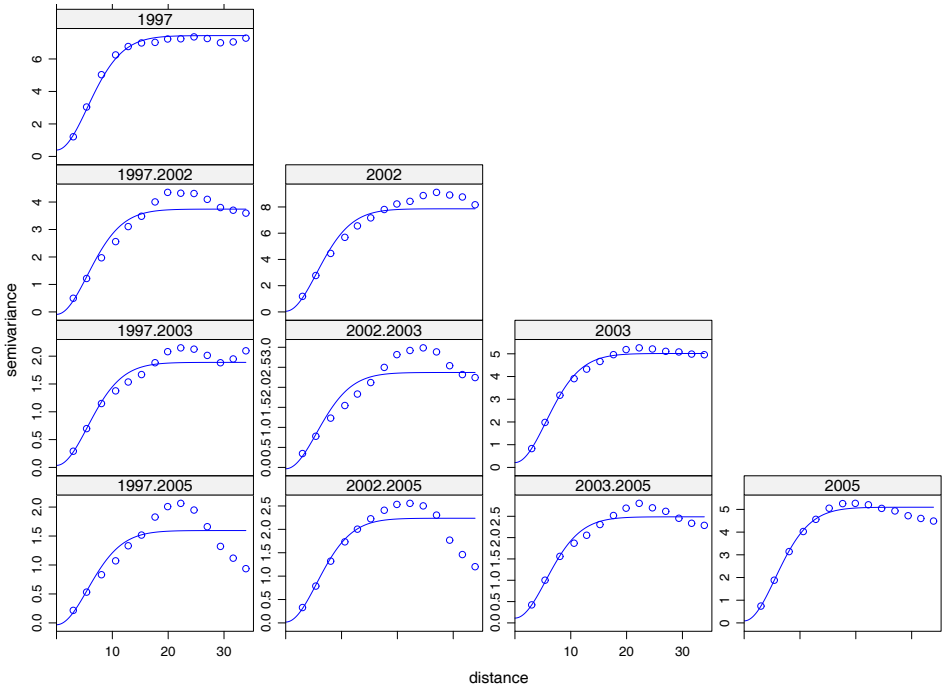
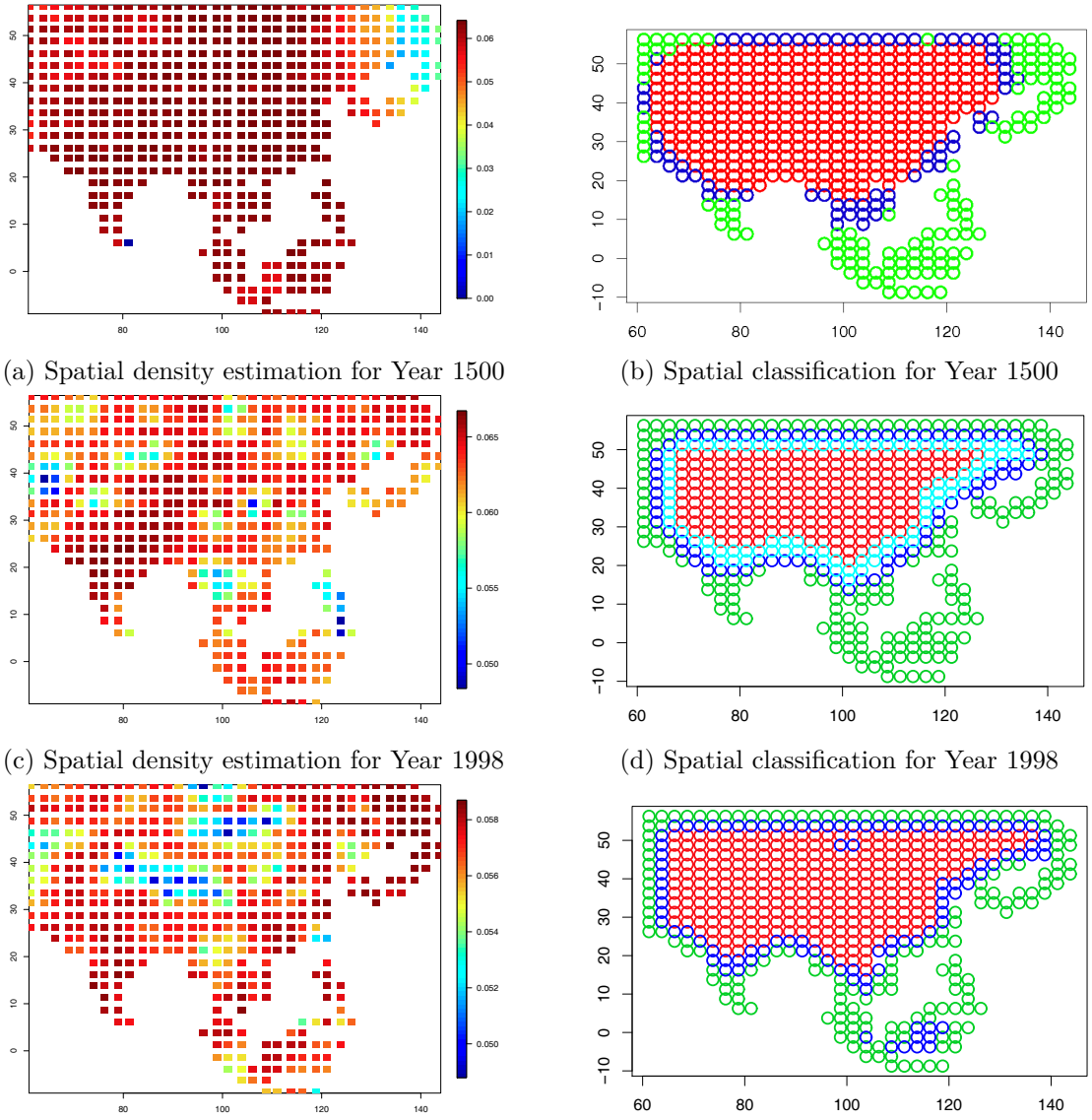


Figure 6.10 – Example of direct variograms (diagonal) and cross variograms (off-diagonal) for the period of interest

The spatial density estimation and classification of PDSI data in the MADA area

The study of the spatial classification of the PDSI for the years 1500, 1998 and 2005

For each year, the dataset is such that $X_{(i,j)} \in \mathbb{R}$ with $60 \leq i \leq 145$ (longitude) and $10 \leq j \leq 60$ (latitude). We first performed the estimation of the spatial density of the PDSI



(a) Spatial density estimation for Year 1500

(b) Spatial classification for Year 1500

(c) Spatial density estimation for Year 1998

(d) Spatial classification for Year 1998

(e) Spatial density estimation for Year 2005

(f) Spatial classification for Year 2005

Figure 6.11 – Examples of Spatial distribution detected by our procedure based on observations of the PDSI in MADA domain for years 1500, 1998, and 2005
On the left: the map of the estimated densities. On the right: the spatial clusters detected by our spatial clustering method

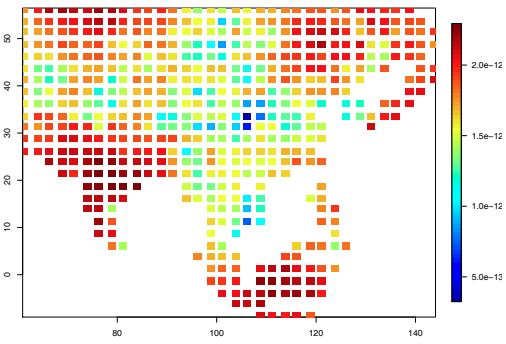
in the MADA domain. The bandwidths for years 1500, 1998 and 2005 are respectively: $\rho_{\mathbf{n},opt}(1500) \simeq 0.1$ and $b_{\mathbf{n},opt}(1500) \simeq 11.5$, $\rho_{\mathbf{n},opt}(1998) \simeq 0.15$ and $b_{\mathbf{n},opt}(1998) \simeq 11$, $\rho_{\mathbf{n},opt}(2005) \simeq 0.1$ and $b_{\mathbf{n},opt}(2005) \simeq 12.5$ obtained using a cross-validation rule. The result is displayed in Figures 6.11a, 6.11c and 6.11e which raise three different spatial regions in each case. Moreover, if their spatial densities seem to have different shapes, the results of the spatial classification procedure presented in Figures 6.11b, 6.11d and 6.11f show that there is almost no change in spatial structures on the inland region when there are some changes in the structure of the frontier region from one year to another. Particularly, in 1500, lower spatial densities are in the northeast contrary to the costal region. The spatial clustering procedure (Figures 6.11b, 6.11d and 6.11f) with this specificity of the northeast

area also shows that the southeast of the land shares this specificity. For the year 1998, it is difficult to detect some differences between the inland and costal areas based on the density estimation when the spatial clustering method detects some differences among the inland region, an intermediate region and a coastal area. Let us notice that the density estimation of PDSI for 2005 raises similar regions as shown in the map of Figure 6.8 when the clustering procedure detects three regions: an inland region, intermediate region and coastal area.

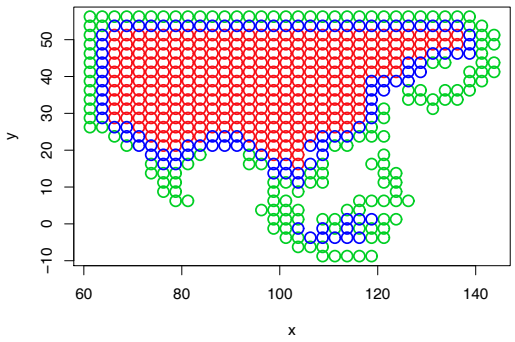
The study of the spatial classification of the PDSI for the 10-year period: 1996-2005

First of all, we formatted the data of this period as observations of $X_{(i,j)}$s with $60 \leq i \leq 145$ (longitude), $10 \leq j \leq 60$ (latitude), $X_{(i,j)} \in \mathbb{R}^{10}$ and $X^{(1)}_{(i,j)}$ corresponding to observations of year 1996, $X^{(2)}_{(i,j)}$ corresponds to observations of year 1997,..., $X^{(10)}_{(i,j)}$ corresponding to observations of year 2005. In this setting, we first performed the estimation of the spatial density of the PDSI in the MADA domain with $\rho_{\mathbf{n},opt} \simeq 0.075$ and $b_{\mathbf{n},opt} \simeq 19$ obtained using a cross-validation rule. The result is displayed in Figure 6.12a which shows that the smallest values of the spatial density are observed in the inland areas contrary to costal areas (India, East Asia and south into northern Australia).

In order to make the spatial homogeneity or heterogeneity precise, we used our procedure to cluster the data set. The result is given in Figure 6.12b The latter confirms the existence of the two classes raised by Figure 6.12a and in addition, reveals some intermediate regions between inland areas and coastal areas that could hardly be observed in Figure 6.12a.



(a) The map of the estimated density



(b) The spatial clusters detected by our spatial clustering method

Figure 6.12 – Spatial distribution detected by our procedure based on observations of the PDSI in MADA domain for the period 1996 – 2005

To conclude, in this chapter, some simulations and a real data application are carried out to illustrate the behavior of the density estimate and the clustering procedure proposed in chapter 2.

Chapter 7

Streamflow hydrograph classification using functional data analysis

Contents

Résumé en français

Précédemment, nous avons appliqué l'estimateur de la fonction de densité spatiale à la classification non supervisée. Dans ce chapitre, nous travaillons également sur le problème de classification non supervisée mais lorsque les données sont fonctionnelles et ne sont plus spatialisées. Les données traitées concernent les débits journaliers de certaines rivières du Québec. Chaque année est représentée par un hydrogramme (graphique de la variation temporelle du débit) considéré dans ce travail comme des données fonctionnelles (courbes). L'objectif est de construire des classes d'hydrogrammes. Nous commençons par introduire le contexte hydrologique et les motivations à classifier les hydrogrammes. Puis, nous rappelons les principes de plusieurs méthodes rencontrées dans la littérature fonctionnelle à savoir la méthode de classification descendante hiérarchique basée sur l'estimation du mode (voir les chapitres 2 et 6), la méthode des k-moyennes avec les divergences de Bregman, mais aussi celle basée sur les projections. Nous appliquons ensuite ces méthodes sur les hydrogrammes de plusieurs stations hydrologiques. Une étude comparative des résultats obtenus avec ces différentes techniques est donnée ainsi qu'une interprétation environnementale des résultats. Nous appliquons également une méthode de statistique multivariée

qui est habituellement utilisée dans la littérature hydrologique. Nous expliquons les avantages et inconvénients à considérer une méthode pour données fonctionnelles plutôt que multivariées.

Ce travail est issu d'un projet de collaboration Franco-Québécoise entre le laboratoire EQUIPPE de Lille et l'Institut National de la Recherche Scientifique de Québec. Ce projet s'intitule "*Étude des variables hydrologiques dans le cadre de l'analyse de données fonctionnelles*" et est financé par la Commission Permanente de Coopération Franco-Québécoise (CPCFQ). Il fait l'objet d'un article en révision et a été réalisé avec Mohamed Ali Ben Alaya (INRS, Québec), Fateh Chebana (INRS, Québec), Sophie Dabo-Niang (Université Charles de Gaulle) et Taha B. M. J. Ouarda (Masdar Institute, Abu Dhabi).

7.1 Introduction

The hydrograph as a graphical representation of the temporal variation of flow, is the main source of information to study flow behavior. The information provided by the hydrograph is essential to determine the severity and frequency of extreme hydrological events, especially floods and droughts. The stream hydrograph is an integration of spatial and temporal variations in water input, storage and transfer processes within a catchment. Thus, hydrographs may present different hydrological regimes for a given watershed (e.g. Hannah et al. (2000)). Therefore, hydrographs of a given watershed may not be similar from year to year. Classifying hydrographs into homogeneous classes is of interest to identify and understand the different regimes, to characterize groups, to separate events, and to detect possible changes. In addition, hydrograph classification is essential to characterize the impacts of climate disturbances on hydrological regimes e.g. Kingston et al. (2011). Hydrograph classification is then very important, particularly where changes in the frequency and/or in the intensity of various forms of extreme weather events could occur. The classification of hydrographs can allow characterizing different hydrological regimes which leads to a better understanding and specific treatment of the behavior of extreme flows and the associated water resource activities (e.g. Harris et al. (2000)).

From a water resources management point of view, hydroelectric utilities are interested in classifying hydrographs based on their shape, and eventually linking the shape to a risk measure (return period for instance). Previous efforts to derive a rational classification procedure were limited mainly by technique availability. Yue et al. (2002), for instance, considered the two-parameter beta probability density function to represent the shape of hydrographs, and used two shape variables (shape mean and shape variance) to classify flood hydrographs. This approach, although simplistic, was useful for the classification of hydrographs for practical purposes in the Province of Quebec, Canada. The hydrological community needs hydrograph classification methods that can provide a full representation of the hydrograph and a full use of all the information contained in it.

In terms of methods, hierarchical classification (HC) is the most commonly used technique for hydrograph classification. Hannah et al. (2000) proposed a multidimensional technique to classify diurnal discharge hydrographs from glacier basins separately according to their shape and magnitude. Their procedure involves two separate classifications of the hydrographs that have been combined. The aim of the first classification is to derive a set of distinct diurnal hydrograph shape classes using the HC approach based on principal component analysis (PCA). The second classification is based on four magnitude indices: the mean, minimum, maximum and variance of monthly observations. This method was adapted by Harris et al. (2000) to riparian systems on four British rivers, where flow regimes are defined by monthly mean flow series. Bower et al. (2004) used

this same method to develop a regime classification to identify spatial and temporal patterns in intra-annual hydro-climatological response as well as an index to assess river flow regime climatic sensitivity. Assani and Tardif (2005) do not use HC approach but proposed 11 hydrological variables based upon monthly discharge data considered with a PCA to identify three significant components used to characterize hydrological regimes in Quebec. Recently, Belmar et al. (2011) proposed a hydrological classification using $\beta-$flexible clustering technique based on weighted PCA scores. The latter are obtained using 73 hydrological indices describing natural flow regimes in Segura River Basin, in Spain. These indices include, for instance, measures of drought duration, as well as flow magnitude, central tendency and dispersion.

In the hydrological literature, including the above mentioned studies, a hydrograph is generally characterized by a limited number of characteristics. However, since the hydrograph represents the variation of flow over a period of time (Yue et al. (2002)), a flood cannot be characterized only by a finite, even large, number of characteristics, but instead by its entire hydrograph as a curve. We propose an illustration which is, for simplicity, based on the main flood features. Figure 7.1(A), shows, in the left panel, two hydrograph types characterized by the same volume and different peaks and durations. In the right panel, the two hydrograph types present the same peak, same volume and same duration. The only difference between them is that the second occurs with a lag time from the first. Multidimensional classification taking into account only the peak, volume and duration can detect the differences between the two hydrograph types of the first example, but is unable to do so for the last example. This last remark is valid whatever the finite number of hydrograph features considered. This is because the continuous character of the hydrograph, as a function, cannot be reduced to any limited number of its features where the hydrograph cannot be fully represented. Figure 7.1(B) illustrates another situation in which two hydrographs can correspond to the same peak value, duration and volume and hence would not be differentiable through a classical multidimensional classification. However, these two hydrographs correspond to two completely different behaviors. Hydrograph I corresponds to a steep rising limb, which would lead to a large volume of water entering the reservoir in a short period of time. This does not give enough time to the operator to evacuate the excess water (due to the capacity of the spillway) and could lead to dam toppling and serious security consequences. On the other hand, Hydrograph II presents a slow rising limb, giving ample opportunity to spill excess water and reduce the risk level. A classification of the hydrographs based on their shape is hence important. This example represents a simple illustration of the importance of the shape of the hydrograph. It would have been possible to add an additional variable representing the time to peak. However, other considerations may require other variables, making the process difficult and highly dimensional, especially in the case of multi-modal hydrographs for example. A general classification procedure is needed and can be provided by the functional framework.

The examples above illustrate that the multidimensional approach depends on the indices used to characterize the phenomenon, and that, not taking into account some indices (e.g. Julian date) can influence the multidimensional classification results. On the other hand, when a large number of features, such as 73 indices (Belmar et al. (2011)), is used, a large quantity of information can be extracted and the hydrograph could be almost represented. However, other drawbacks occur, such as the increase of dimensionality, redundancy, and subjectivity. When the number of variables to include increases, the number of choices and possibilities of subsets of variables increases as well. Some variable selection techniques (see, e.g. Fraiman et al. (2008), Andrews and McNicholas (2013)) can be used but are often computationally intensive and are based on the iterative use of hypothesis

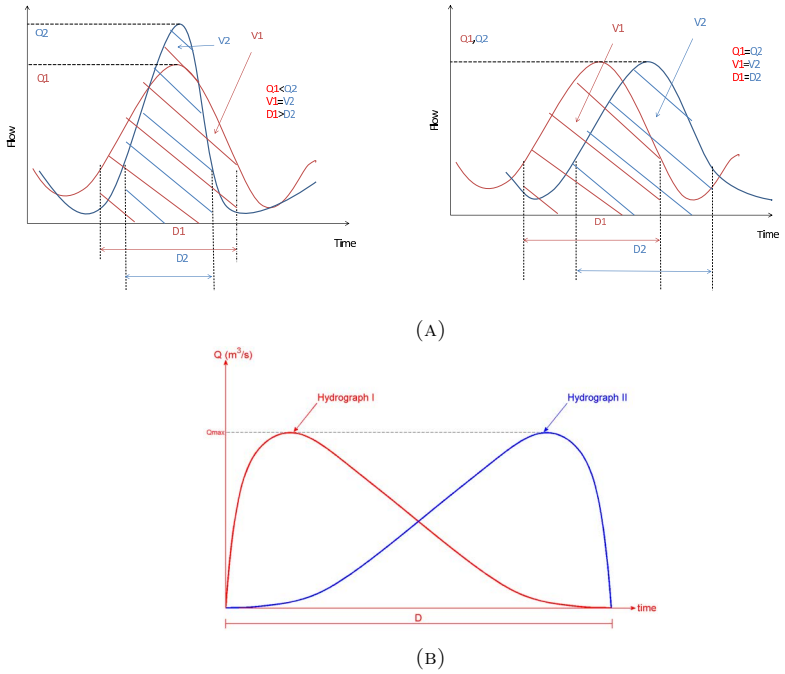


Figure 7.1 – Examples of flood hydrographs
(A) Flood hydrographs characterized by different flood types - (B) Flood hydrographs characterized by different behaviors. V, Q and D denote the volume, the peak and the duration, respectively

testing which induces errors at each step. In addition, some variables are not directly available, such as the volume, and require extraction from the raw data, which can cause an increase in uncertainty due to the lack of accuracy in their computation. A number of the considered variables or indices are also usually taken at monthly or annual time scales which reduces capturing the temporal variability of the hydrological phenomenon. The above considerations have negative impacts on hydrograph classification especially in terms of information loss and they consist in a substantial simplification of the overall hydrological phenomenon.

Recently, Chebana et al. (2012) introduced the functional data analysis (FDA) statistical framework to the hydrological context. It is important to mention that in hydrology, the term 'functional' is mainly used to refer, e.g. to the function of a catchment as its input-output conversion. Here, the term FDA is employed to refer to a statistical framework based on functional data. Specifically, Chebana et al. (2012) focused on exploratory analysis as well as outlier detection of hydrographs. They showed that the FDA framework is adapted to the hydrological context with a number of advantages. Indeed, the functional framework is more general, flexible and representative of the real hydrological phenomena than multidimensional analysis. In fact, the former treats the whole hydrograph as a functional observation (function or curve) taking into account the maximum of the available information and constitutes a natural extension of multidimensional approaches. Hence, a classification approach based on the whole hydrograph of an annual discharge time series as a single observed function could lead to more representative classes. It is relevant to mention that Pappenberger and Beven (2004) proposed a hydrograph classification approach using the multidimensional HC method and based on graphical visualizations of hydrographs. Ganora et al. (2009) considered the flow duration as a curve for regionalization purposes. However, even though these studies indicate the need to consider the

whole hydrograph in the classification process, they do not have a functional statistical foundation.

In the functional framework, a variety of hydrographs could be covered and all their features would be included without necessarily increasing the uncertainty and subjectivity. Active research is targeting the development of adapted statistical methods to analyze functional data. A number of classical approaches are extended to the functional context (e.g. Ramsay and Silverman (2005), Dabo-Niang et al. (2006), Cadre and Paris (2010), Fischer (2010)). This is also the case for classification methods. The functional versions are often extensions of the classical classification ones, in particular the HC and the k-means methods. The aim of the present chapter is to introduce the FDA framework for classifying streamflow hydrographs, by considering discharge time series as continuous curves.

The theoretical background of functional classification methods is presented in Section 7.2, in its general form. In Section 7.3, these methods are adapted to the hydrological context and applied to daily streamflows of the Romaine River station, from the province of Quebec, Canada. For more general results, functional methods are also applied to 13 other stations in the province of Quebec and the results are briefly presented. A comparison with a multidimensional classification method is also given in Section 7.3. A discussion of all results is carried out in Section 7.4 and conclusions are reported in Section 7.5.

7.2 Functional data classification methods

The purpose of this section is to present some recently developed procedures for classifying functional data. Let $x_i = (x_i(t_1), \ldots, x_i(t_T))$, $i = 1, \ldots, n$, be a set of n discrete observations, where each t_j, $j = 1, \ldots, T$ is the j^{th} record time point from a given continuous time subset $\mathcal{T}$ which includes the set $\{1, \ldots, T\}$. For instance, an observation x_i could be daily flow temperatures series within a given i^{th} year with $T = 365$. Before going further, note that the statistical object of FDA is a function (curve). However, the curves are not completely observed, instead, only discrete measurements of the curves are available. Then, a first step is to prepare the data to be used in a FDA context. For a fixed observation i, each set of measurements $(x_i(t_1), \ldots, x_i(t_T))$ is converted to be a functional data and denoted $\{X_i(t),\, t \in \mathcal{T} \subset \mathbb{R}^+\}$, by using a smoothing technique. In the case where data series are of good quality and with long enough records, one can simply interpolate the measurements to obtain the curves. Otherwise, smoothing can be required. This is typically the case for diffusive processes like in the present study of streamflows. However, even in the first case, smoothing could be necessary depending on the objective of the study (e.g. Ramsay and Silverman (2005)). The reader is referred to Chebana et al. (2012) for more details related to this issue in the hydrological context.

Generally, in any classification framework, data inside each class should be as similar as possible but different from those in other classes. Note that classification can be supervised or unsupervised (e.g. Hartigan (1975)). In the first one, the number of classes is known in advance or chosen according to the study constraints. Otherwise, unsupervised approaches are considered. This is the case in the present chapter where a number of classification approaches are available in the functional literature. In particular, k-means techniques are adjusted to functional data, hierarchical algorithm and some of its variants are proposed as well (e.g Abraham et al. (2003), Dabo-Niang et al. (2007), Auder and Fischer (2012)).

An appropriate classification should lead to homogeneous classes and heterogeneity between classes, thus avoiding unnecessary classes. Consequently, the obtained number of classes k is important (e.g. Milligan and Cooper (1985)). Some classification methods do not automatically determine k. Techniques are developed in the literature to overcome

this difficulty. One of the techniques consists in selecting k that optimizes a given class homogeneity index (e.g. Krzanowski and Lai (1988)). In the application section below, for the k-means classification, an initial and arbitrary choice of k is taken at the beginning of the procedure which is not necessary the final choice. Note that an extensive literature exists for the initialization of the k-means algorithm (e.g. Khan and Ahmad (2004)). In addition, the size of the obtained classes could be a concern, according to the classification aims. For instance, if the purpose is inference, such as modeling or estimation, then large size classes are necessary for reliable results. On the contrary, an exploratory or descriptive analysis does not require any size constraints and small classes could be of interest.

The set of the curves $X_1, \ldots, X_n$ is denoted S. The following presented approaches should lead to a splitting of S into some k distinct representative and interpretable classes $S_1, \ldots, S_k$. Firstly, we present the method studied by Dabo-Niang et al. (2007) which is a descendant HC procedure based on distances between the modal and mean curves of a set of curves. Secondly, the k-means classification is presented, consisting in partitioning the observations in classes by minimizing some distortion measures defined below (Fischer (2010)).

7.2.1 Descendant hierarchical classification based on centrality curves

Recent advances in nonparametric FDA allow to define centrality features for a sample of curves (see e.g. Ferraty and Vieu (2006)). Dabo-Niang et al. (2007) indicated that both the mean and the median curves are interesting when dealing with homogeneous data, while the modal curve would be more useful for detecting possible different structures in the data. Consequently, Dabo-Niang et al. (2007) used a descendant HC method based on comparing the modal curve either with the mean or the median. In this chapter, this method is applied in hydrology where location curves can be used to characterize a given basin and to proceed to comparison or grouping of a set of basins.

Location measures (mean, mode and median) summarize the data and aim to provide a representative element of the sample. In the FDA context, for a set of curves S, we define the mean curve as $X_{mean,S} = \dfrac{1}{\text{Card}(S)} \displaystyle\sum_{X_i \in \mathcal{S}} X_i$, where $\text{Card}(S)$ denotes the number of elements in a set S. The median curve is estimated by $X_{median,S} = \underset{X \in S}{\text{argmin}} \displaystyle\sum_{X_i \in \mathcal{S}} m(X, X_i)$, where X_i and X are curves in the set S and $m(\cdot, \cdot)$ is a proximity measure. The modal curve is defined such that the sample of curves is the most dense. From a theoretical point of view, the mode, when it exists, is an observation whose probability is locally maximum. So, the modal curve of a sample S can be estimated as:

$$X_{modal,S} = \underset{X \in S}{\text{argmax}} \sum_{X_i \in \mathcal{S}} K\left(\frac{m(X, X_i)}{h}\right)$$

where $K(\cdot)$ is a kernel function, $h = h_n$ is a sequence of positive numbers called bandwidth, considered as a smoothing parameter and $\arg\max_{x \in S} f(x)$ stands for the element in the set S that maximizes a function f. The elements m, K and h are essential in nonparametric estimation. In the functional context, a semi-metric $m(\cdot, \cdot)$ is often used as a proximity measure. In particular, a semi-metric based on derivatives, as used in the application section, is defined by

$$m_q^{deriv}(X_i, X_j)^2 = \int \left(X_i^{(q)}(t) - X_j^{(q)}(t)\right)^2 dt$$

where $X^{(q)}$ denotes the qth derivative of X (see Ferraty and Vieu (2006) for more details). A kernel K is a weighting function used in nonparametric estimation techniques. There exists a large variety of kernels in the FDA context, the most classical ones are the positive and symmetrical kernels such as the box kernel, the triangle kernel, the quadratic kernel and the gaussian kernel (see Ferraty and Vieu (2006)).

The proposed methodology performs iteratively by splitting S into increasingly homogeneous classes. To measure the heterogeneity of a given sample S of curves, Dabo-Niang et al. (2007) compared modal and mean curves by computing the subsampling heterogeneity index (SHI). The median curve can also be used instead of the mean, e.g. when one wants to assign to the same group all curves that have the same shape but which are affected by some clearly horizontal shift (see Dabo-Niang et al. (2007)). The SHI is computed by using a large number L of randomly generated subsamples $S^{(l)} \subset S$ of the same size

$$SHI_{mean}(S) = \frac{1}{L}\sum_{l=1}^{L} \frac{m(X_{modal,S^{(l)}}, X_{mean,S^{(l)}})}{m(X_{mean,S^{(l)}}, 0) + m(X_{modal,S^{(l)}}, 0)} \tag{7.1}$$

where $m(X, 0)$ denotes the proximity measure between a function X and the constant null function 0. A large value of $m(X_{modal,S^{(l)}}, X_{mean,S^{(l)}})$ indicates that $X_{modal,S^{(l)}}$ and $X_{mean,S^{(l)}}$ have different behaviors according to $m(\cdot,\cdot)$. The larger $SHI_{mean}(S)$ is, the more heterogeneous the sample S will be. However, since the goal is to decide if the set S should be split into G classes $S_1, \ldots, S_G$ another index is required. The splitting will be accepted if the heterogeneity in each class is smaller than before splitting. To this end, the Partitioning Heterogeneity Index (PHI) is considered. It is defined as a weighted average of the SHI over classes

$$PHI_{mean}(S; S_1, \ldots, S_G) = \frac{1}{\mathrm{Card}(S)} \sum_{g=1}^{G} \mathrm{Card}(S_g) SHI_{mean}(S_g) \tag{7.2}$$

The larger PHI is, the more heterogeneous each class $S_1, \ldots, S_G$ is. Both $SHI_{mean}(S)$ and $PHI_{mean}(S; S_1, \ldots, S_G)$ are employed to define a score SC given by

$$SC = SC_{mean}(S; S_1, \ldots, S_G) = \frac{SHI_{mean}(S) - PHI_{mean}(S; S_1, \ldots, S_G)}{SHI_{mean}(S)} \tag{7.3}$$

A positive score SC indicates a gain of homogeneity inside classes. The splitting is accepted if SC is greater than a fixed threshold τ. For instance, $\tau = 5\%$ indicates that a splitting is accepted if it brings more than 5% of homogeneity within classes. If the score is negative, then S does not require this splitting. A value of τ that is too small indicates that the considered splitting is not required and the gain in terms of homogeneity is not significant. The value of τ is chosen according to the type of the data and the purpose of the classification. It is analogous to the choice of the first kind error in hypotheses testing. All the details concerning this methodology are given in Ferraty and Vieu (2006).

Aside from the above splitting criteria, it is required to define classes of S. Ferraty and Vieu (2006) proposed a procedure to establish the subgroups $S_1, \ldots, S_G$ as well as their number G. The procedure is related to the choice of the bandwidth parameter h by using the small ball probabilities. They play a key role in the theoretical properties of the mode estimate, (see Dabo-Niang et al. (2007)). A small ball probability is defined as $\mathbb{P}[X_i \in B(X, h)]$ for $X, X_i \in S$ which is the probability that a curve $X_i \in S$ belongs to the ball $B(X, h)$ with center X and radius h. For a given bandwidth h, one has at hand n probability points $\mathbb{P}[X_i \in B(X, h)], i = 1, \ldots, n$ for which the corresponding density $d_{S,h}$

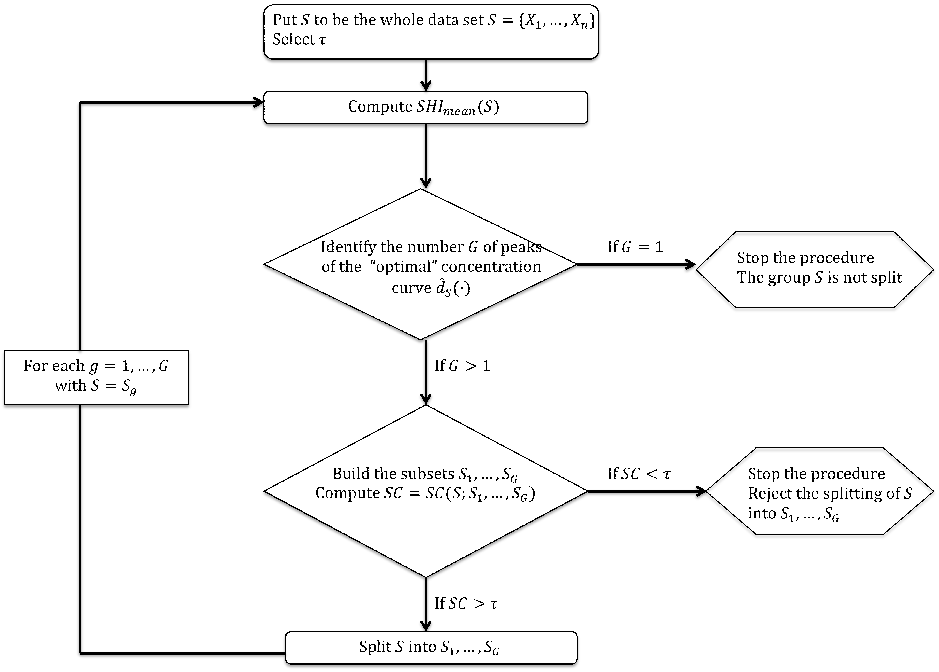


Figure 7.2 – Algorithm of the descendant HC

can be estimated by $\hat{d}_{S,h}$. The estimated density can be computed using, for instance, the package *np* of the R language (see Hayfield and Racine (2008)). The number of groups G will be determined by the number of peaks of $\widehat{d}_{S,\widehat{h}_S}$. In practice, the bandwidth is selected using the entropy such that $\widehat{h}_S = \arg\min_h \int \widehat{d}_{S,h}(t) \log \widehat{d}_{S,h}(t)\, dt$.
The main algorithm of this classification approach is illustrated in Figure 7.2. The reader is referred to Ferraty and Vieu (2006) and Dabo-Niang et al. (2007) for more details.

7.2.2 k-means functional classification

The k-means classification is a widely used method. In the classical analysis, it consists in partitioning observations $y_i \in \mathbb{R}^p$ into k classes, by minimizing a certain quantity, named distortion in this context (similar to a distance). Given k, the preliminary number of classes, the general k-means algorithm is the following (e.g. Hartigan (1975)):

1. Choose k initial centers $c_1, \dots, c_k$;
2. Identify for each observation i the nearest center c_l. Each center c_l defines a class l. The proximity between y_i and c_l is measured by a distortion measure $d(y_i, c_l)$;
3. Define new centers: for each l, c_l is the mean of the observations of the new class l;
4. If the composition of each class is the same as in Step 2 then stop the algorithm and save the obtained classes ;
5. Otherwise, go to Step 2.

This algorithm requires an adaptation to the functional setting. The main difficulty consists in dealing with the infinite dimensionality of the data curves. Note that a functional random variable takes values in an infinite dimensional space (such as functional space, e.g. space of continuous functions on $[0, T]$) (see Ferraty and Vieu (2006)). Firstly, the classical k-means algorithm can be used with Bregman divergences as the distortion measure (see Fischer (2010)). In the following, this method is presented. Secondly, projection

approaches proposed by Abraham et al. (2003) and James and Sugar (2003) is another adaptation. It consists in projecting the curves on a basis, and get classes by considering the k-means algorithm applied on the coefficients of the projection. The k-means method developed in Auder and Fischer (2012) is also presented in this section, where the infinite dimension is reduced by considering only the first p coefficients of the projection on a basis and then performing classification in $\mathbb{R}^p$.

For the descendant HC based on centrality curves, the heterogeneity inside classes is measured through the above heterogeneity indices (SHI, PHI, SC). Although, in the literature, these indices are not generalized to other methods, in the application section, they are computed for the classifications obtained with the k-means methods as well for comparison purposes.

Bregman divergence

In the general k-means algorithm, a distortion measure $d(\cdot,\cdot)$, between an observation and the center of a class, is needed. The main purpose of the k-means classification is to allocate each observation y_i to a class l by minimizing the mean of the distances between each observation and its nearest class center c_l that is

$$\min_{l=1,\dots,k} \frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{k} z_i^l d(y_i, c_l) \tag{7.4}$$

where $z_i^l = 1$ if y_i belongs to class l, otherwise $z_i^l = 0$.

In $\mathbb{R}^p$, different well-known distances are used as distortion measures such as the Euclidean distance. However, in the context of curves classification, it is necessary to consider a notion of distance adapted to high dimensions. To this end, Banerjee et al. (2005) showed that the k-means algorithm should be generalized by replacing the classical distance by the Bregman divergence.

The Bregman divergence was introduced by Bregman (1967) in the multidimensional context. Most of the widely used distortion measures, as the Euclidean distance, are particular cases of Bregman divergences. The Bregman divergences are generalized to the functional context and are defined by

$$d_\phi(x,y) \quad = \quad \phi(x) - \phi(y) - D_y\phi(x-y) \geq 0 \tag{7.5}$$

where ϕ, x and y are functions, $\phi : \mathcal{C} \to \mathbb{R}$ with $\mathcal{C}$ is a convex set, $D_y\phi$ denotes the differential of the function ϕ at y. According to the choice of the function ϕ, the functional Bregman divergence may be, for instance, the squared L^2 distance, the quadratic bias, the classical and generalized Kullback-Leibler divergences (e.g. Fischer (2010)). The divergence choice depends on the data and on the type of partition required. In particular, the quadratic bias, used in the application section, is defined by

$$d_\phi(x,y) \quad = \quad \left[\int (x-y)d\mu\right]^2 \qquad \text{using} \qquad \phi(x) = \left[\int_{\mathcal{I}} x(t)d\mu(t)\right]^2 \tag{7.6}$$

with $\mathcal{I}$ is an interval in $\mathbb{R}$ and μ is a finite positive measure.
Similar idea is employed in Chebana and Ouarda (2011) to measure errors between multivariate quantile curves applied to hydrological variables.

Projection-based curve clustering

Auder and Fischer (2012) proposed an approach for curve classification by reducing the infinite dimension based on projecting the curves $X_1, \ldots, X_n$ onto a finite lower-dimensional space. Then, k-means classification is performed on the first p coefficients of the corresponding basis projection.

Since the k obtained classes may depend on the basis choice, several projection bases are used in practice such as Fourier, Haar and functional principal component. Another basis is proposed by Auder and Fischer (2012) as the Best-Entropy basis.

Given the centers $\mathbf{c} = (c_1, \ldots, , c_k)$ of the k classes, the goal of this method of classification is to find the basis minimizing the following distortion:

$$W_{p,n}(\mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{l=1,\ldots,k} \|\Pi_p(X_i) - \Pi_p(c_l)\|^2 \tag{7.7}$$

where Π_p is the orthogonal projection on $\mathbb{R}^p$. To this end, according to different values of the projection dimension p, the k-means algorithm is implemented on the projected coefficients resulting from the above basis. The selected classification is the one where the projection dimension p and the basis minimize distortion (7.7).

In the application section below, the above presented functional approaches are applied to real world data and a comparison with a classical multidimensional approach is carried out. For a quantitative comparison of the different classification approaches, the Silhouette index is considered (see Kaufman and Rousseeuw (1990), Chapter 2). It is presented briefly here.

For a given classification, and for each object x_i of a dataset $x_1, \ldots, x_n$, the corresponding Silhouette index $s(i)$ is defined as

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}$$

where, for a given x_i belonging to a class A, (with Card(A) ≥ 2), and a distance $d(\cdot, \cdot)$,

$$a(i) = \frac{1}{\text{Card}(A) - 1} \sum_{\substack{x_j \in A \\ j \neq i}} d(x_i, x_j), \qquad \text{and} \qquad b(i) = \min_{A \neq C} \frac{1}{\text{Card}(C)} \sum_{x_k \in C} d(x_i, x_k)$$

for all classes C different from A, from the same classification.

Note that, Kaufman and Rousseeuw (1990) proposed to use the Euclidean distance for the distance $d(\cdot, \cdot)$. For each object x_i, $s(i)$ is between -1 and 1. An observation x_i is considered well classified when $s(i)$ is large. Consequently, for each classification, the mean $\bar{s}$ of the $s(i)$, $i = 1, \ldots, n$ is evaluated and the best classification (with compact and well separated classes) corresponds to the largest $\bar{s}$. In the application section, the Silhouette index is computed on the raw data.

7.3 Applications

The methods presented in Section 7.2 are adapted to hydrological discharge time series in Section 7.3.1. Then, they are applied to the case study data from the province of Quebec described in Section 7.3.2. The obtained functional results are presented and compared with those obtained from a multidimensional ascendant hierarchical classification method in Section 7.3.3. The aim of the case study is to illustrate the functional framework with a number of sub-methods and provide a comparison between frameworks (functional and classical) rather than select a specific method within a given framework.

7.3.1 Adaptation to discharge time series

We consider flow series recorded for a given station. These data can be recorded at different time scales such as hourly, daily or monthly. In the following, we focus on daily flow data assumed to be available for n hydrological events. These hydrological events can be, for instance, floods and are denoted by $x_i(t) = \{x_i(t_j),\, j = 1, \ldots, T\},\ i = 1, \ldots, n$, where T represents the number of days in the period of year covering the hydrological event, such as $T = 365$ for the whole year or $T = 184$ for Spring floods. The flow value $x_i(t_j)$ is measured at day j for the i^{th} event. Usually, these discrete observations have the same size T. These observations $x_i(t_j)$ are converted into smooth functions $X_i(t)$ on a continuous period $\mathcal{T} = [1, T]$. The obtained function $X_i(t)$ constitutes a hydrograph of one event.

In this chapter, for a given hydrological year, we only consider the spring high flow event. Then, the three different functional classification methods presented in Section 7.2 (descendant HC based on centrality curves, k-means with Bregman divergence and k-means with projection-based curve), are applied to the smooth functions $X_i(t)$. We consider the mode and the mean for measuring the heterogeneity index, (7.1) and (7.2). Note that, in this chapter, we consider unsupervised classification, since the number of classes k is usually unknown in advance. Thus, for the k-means approaches, the initial values of k are chosen arbitrarily: different values have been tested to make the choice.

7.3.2 Data description

The data series is represented by daily flow ($\text{m}^3\ \text{s}^{-1}$) from the Romaine River station with reference number $02VC001$. The area of the drainage basin is 13 000 km^2. The focus in this study is on spring high flow events occurring between March 1st and August 31st that is $j = 1, \ldots, 184$, for $T = 184$. Indeed, in the province of Quebec, the largest streamflow events are mainly caused by snow melt during the Spring season and can continue until the end of Summer. Data is available from 1961 to 2000. According to the present data set, we have $n = 40$ years of observations $x_i(t_j)$, $t_j = 1, ..., 184$, $i = 1, \ldots, 40$. The i^{th} observation $\{x_i(t_j),\, j = 1, \ldots, 184\}$ denotes the daily flow measurements for the i^{th} year which is converted to a smooth curve $\{X_i(t),\, t \in \mathcal{T} = [1, 184]\}$. This is performed through Fourier series expansion as indicated in Chebana et al. (2012). Figure 7.3 presents all the obtained smooth curves for the studied period. It shows that these curves have several different shapes. Therefore, it is appropriate to classify these curves.

As indicated in Section 7.2, S represents the whole set of flow curves $X_1, \ldots, X_{40}$. The presented methods in Section 7.2 are applied. The results obtained from the different functional approaches are also compared with a classical classification methodology, that is the multidimensional HC. Since we focus on unsupervised classification, the number k of classes is unknown. Moreover, because of the exploratory illustrative character of the present study, no constraint on the size of the classes is imposed.

Then, 13 other stations are considered. They represent pristine basins and were selected as part of the reference hydrometric basin network (RHBN) to help provide an understanding of the physical processes within and account for the impacts of climate variation across the province of Quebec (Ouarda et al. (1999)). These stations are listed in Table 7.1.

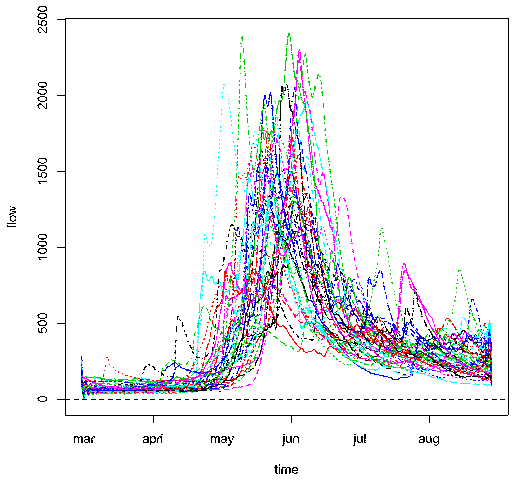


Figure 7.3 – The curves corresponding to the studied period, March 1st to August 31st, for the Romaine River station

Number	Federal Number	Station name	Area (km^2)
1	01BH005	Darmouth	645
2	02QA002	Rimouski	1610
3	02PJ007	Beaurivage	704
4	02OE007	Eaton	642
5	02LH004	Picanoc	1290
6	02NE011	Croche	1570
7	02PB006	Sainte-Anne	642
18	02RG005	Metabetchouane	2280
19	02RF005	Chamouchouane	15300
10	02RD002	Mistassibi	9320
11	02UC002	Moisie	19000
12	04NA001	Harricana	3680
13	03MB002	À la Baleine	29800

Table 7.1 – List of stations

7.3.3 Results

The functional approaches are applied to the data presented previously and results are given in Section 7.3.3. Results for the Romaine River station, in Quebec, are presented in detail. Furthermore, the main results for 13 other stations, across the province of Quebec, are briefly presented. For comparison purposes, a multidimensional approach is also applied to the data and the results are presented in Section 7.3.3.

Functional classification results

Firstly, the functional descendant hierarchical classification method based on modal curve (CM), as explained in the second section, is performed on the Romaine River station. The function m, which measures the proximity between curves, is a crucial element in some of the above functional methods. The choice of m can be related to practical hydrological classification aim. Ideally, hydrographs with very similar shapes but affected by a horizontal shift should be assigned to different groups. As a first consequence, a good proximity function between such curves can vary under translation (because an invariant translation distance would produce small values whereas one expects large values). Note

that the distance chosen here is the semi-metric, defined above, based on 2 derivatives. Other tested semi-metrics do not lead to better interpretable classification. As a second consequence, taking the mean curve as a centrality curve makes sense in the computation of heterogeneity indices. Then CM is based on the comparison between modal and mean curves. To confirm our choice, the case where the method is based on the deviation between modal and median curves has been tested: the obtained classes were not of interest in the present hydrological context. The corresponding algorithm is implemented on the set of curves S with a threshold τ fixed at 5%. Indeed, we consider that a gain of homogeneity lower than 5% is not significant. The SHI given in (7.1) is computed for the whole data set S where $SHI_{mean}(S) = 0.77$ which indicates heterogeneity level that is high enough. At the first iteration, the data set is split into two classes, denoted CM_1 and CM_2 with sizes 23 and 17, respectively. Compositions of these classes are presented in Figure 7.4 in terms of occurrence years. Figure 7.4 presents all the obtained classes by each one of the considered methods in terms of time occurrence. For instance, for the CM method, the year 1961 belongs to the class CM_1 whereas the year 1964 belongs to CM_2. We note a discontinuity of years within each of these two classes. The partitioning heterogeneity index PHI and the score SC, defined respectively in (7.2) and (7.3), are evaluated for CM_1 and CM_2. The values are given in Table 7.2. $PHI_{mean}(S; CM_1, CM_2) = 0.66$ indicates that this splitting is temporary accepted since it allows to increase homogeneity inside classes by $SC = 14\%$ which exceeds $\tau = 5\%$. In the second iteration, CM_1 is split into two subgroups, but this splitting is not accepted since the SC is around 1% which is lower than the threshold 5%. The class CM_2 is not split because the density estimation of the concentration curves has only one peak. At this stage, the algorithm is stopped and leads to a final classification in CM_1 and CM_2 of S.

$SHI_{mean}(S) = 0.77$			
		$PHI_{mean}(\mathcal{S}; S_1, S_2, (S_3))$	SC
Descendant HC	CM	0.66	0.14
k-means with	$KMB^{(a)}$	0.77	0.00
Bregman divergence	$KMB^{(b)}$	0.69	0.11
k-means with	$KMP^{(a)}$	0.72	0.07
projection-based curve	$KMP^{(b)}$	0.62	**0.20**

Table 7.2 – Heterogeneity Indices for the Romaine River station
Bold character indicates highest score SC

For description and interpretation of these classes, Figure 7.5a represents all the curves of the set S of each class with different colors according to class, black curves for CM_1 and grey curves for CM_2. In addition, for a concise view, the centrality curves, namely the mean, median and modal, of each class are given in Figures 7.5b-7.5d. We observe that the mean and modal flows of the class CM_2 are higher than those of class CM_1. We note that mean curves do not reflect properly the structure of the studied curves, since they are not observations contrary to median or modal curves. The median curve (Figure 7.5c) of class CM_1 is more elevated than the median curve of CM_2. On the other hand, event durations in CM_2 are longer than in CM_1.

Secondly, the same data are classified using the functional k-means method with Bregman divergence (KMB). The number of $k = 2$ classes is initially considered with initial centers as the modes of classes CM_1 and CM_2 obtained from the CM method. These modes correspond to the curves of the years 1991 and 1970 respectively (illustrated in Figure 7.5d). According to these considerations, the algorithm leads to the two classes denoted

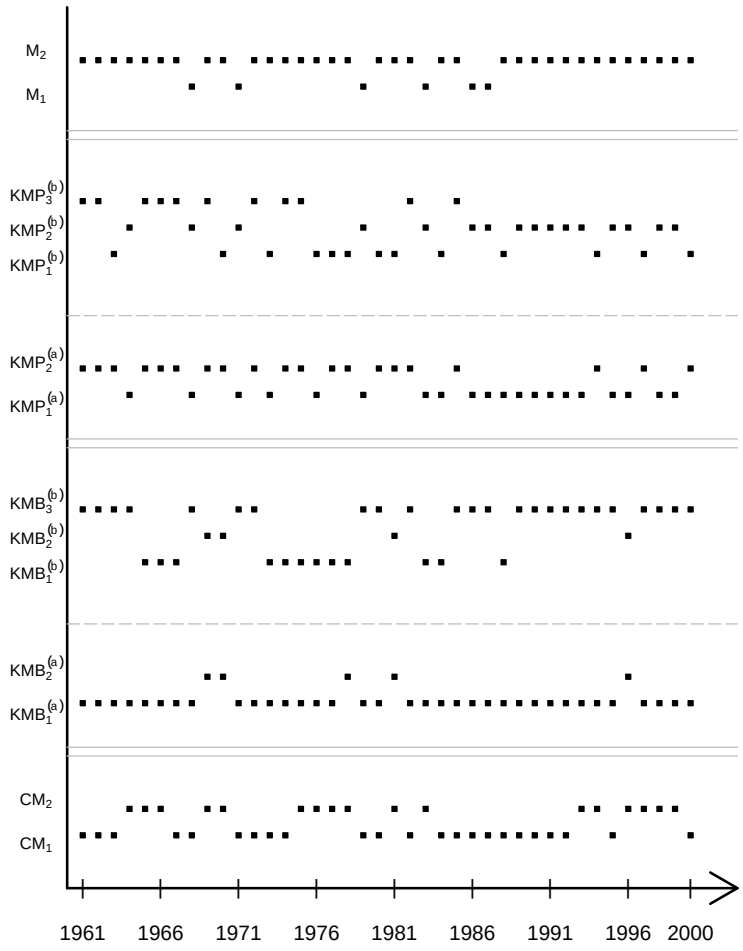


Figure 7.4 – Composition of classes obtained for the Romaine River station, according to the different methods

CM (Functional hierarchical classification based on modal curve), KMB (Functional k-means with Bregman divergence), KMP (Functional k-means with projection-based curve), M (Multidimensional hierarchical classification)

by $KMB_1^{(a)}$ and $KMB_2^{(a)}$, with sizes 35 and 5 respectively. Their composition is shown in Figure 7.4. The sizes of these classes are not of the same order of magnitude since no constraint was imposed in this sense. Therefore, the class $KMB_2^{(a)}$ of 5 curves could be seen as a class of unusual curves. To measure the heterogeneity in these classes, PHI and SC are evaluated and given in Table 7.2. Note that $PHI_{mean}(S; KMB_1^{(a)}, KMB_2^{(a)}) = 0.77$ equals to $SHI_{mean}(S)$ and is higher than $PHI_{mean}(S; CM_1^{(a)}, CM_2^{(a)}) = 0.66$. Consequently, the classification in $KMB_1^{(a)}$ and $KMB_2^{(a)}$ does not increase the homogeneity within classes which is confirmed by the null value of $SC = 0$.

For the above reasons, i.e. "unbalanced" sizes and $SC = 0$, the number of classes is increased to $k = 3$. The obtained three classes, denoted $KMB_1^{(b)}$, $KMB_2^{(b)}$ and $KMB_3^{(b)}$, are respectively of sizes 12, 4 and 24 and the corresponding compositions are presented in Figure 7.4. Table 7.2 indicates that the value of $PHI_{mean}(S; KMB_1^{(b)}, KMB_2^{(b)}, KMB_3^{(b)})$ is 0.69 which is smaller than 0.77 as the value obtained from $KMB_1^{(a)}$ and $KMB_2^{(a)}$

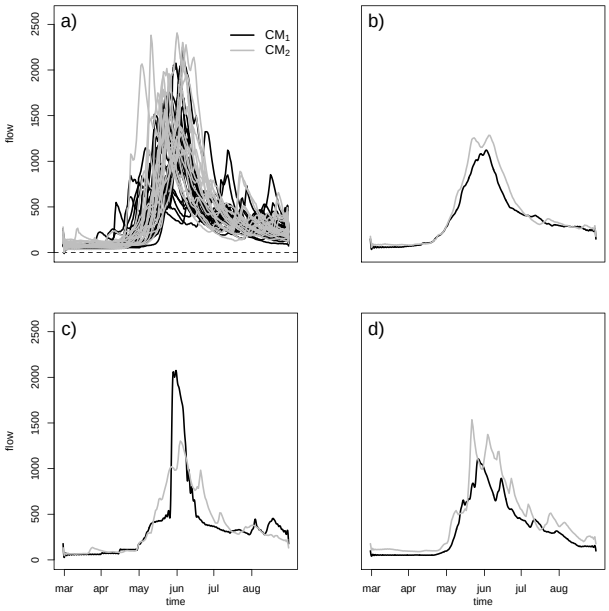


Figure 7.5 – Curves and centrality curves of each of the two clusters, using the descendant HC based on centrality curves, for the Romaine River station
a) all curves b) mean curves c) median curves d) modal curves

and the SC increases to be 0.11. Therefore, the classification into $KMB_1^{(b)}$, $KMB_2^{(b)}$ and $KMB_3^{(b)}$ reduces the heterogeneity in classes compared to the classification into $KMB_1^{(a)}$ and $KMB_2^{(a)}$. Figure 7.6f illustrates the corresponding three distinct mean curves. In terms of mean, class $KMB_3^{(b)}$ represents the curves with low peak and short duration whereas class $KMB_2^{(b)}$ represents the curves with higher peak and longer duration. However, class $KMB_1^{(b)}$ corresponds to curves with intermediate features between those of $KMB_2^{(b)}$ and $KMB_3^{(b)}$. Figures 7.6g and 7.6h show that median and modal curves corresponding to class $KMB_3^{(b)}$ are smaller than for $KMB_1^{(b)}$ and $KMB_2^{(b)}$.

Note that we have tested the four Bregman divergences above mentioned in the second section. The quadratic bias (7.6) is chosen because the corresponding results lead to clearly distinguished and meaningful classes.

Thirdly, the functional k-means method with projection-based curve (KMP) is implemented on the data from the Romaine River station. The initial number of classes is taken to be $k = 2$. The k-means algorithm has been run on the projected coefficients for the four bases indicated in the second section, namely Best-Entropy, Haar, Fourier and functional PCA (principal component analysis) basis, with a number of projection dimension of p values. For each basis and different values of the dimension projection p, the empirical distortion given in (7.7), is computed for the Romaine River station and presented in Table 7.3. According to Table 7.3, the distortion is minimal for a projection onto $p = 18$ functional PCA basis then, for the classification, we consider projections onto $p = 18$ functional PCA basis. The size of both obtained classes, denoted $KMP_1^{(a)}$ and $KMP_2^{(a)}$, is 20 and their composition is shown in Figure 7.4. Class $KMP_2^{(a)}$ contains most years before 1981 whereas $KMP_1^{(a)}$ mainly contains years after 1981. The value of the $PHI_{mean}(S; KMP_1^{(a)}, KMP_2^{(a)}) = 0.72$ is not the smallest compared to the previous tested methods and the score $SC = 0.07$ is not the highest (Table 7.2). Figure 7.7b indi-

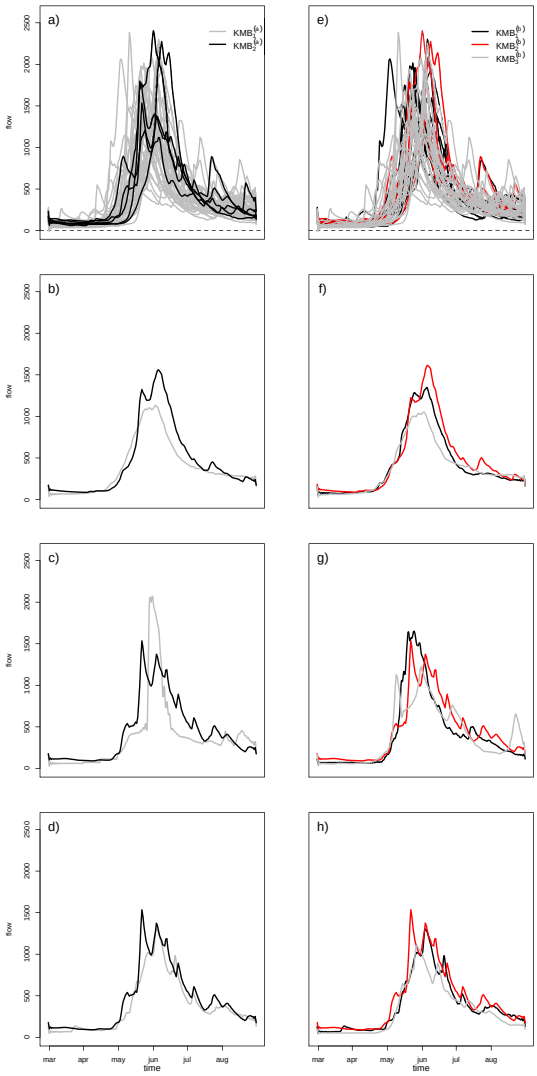


Figure 7.6 – Curves and centrality curves of each of the two and three classes, using the classification by k-means with Bregman divergences, for the Romaine River station
a, e) all curves b, f) mean curves c, g) median curves d, h) modal curves

cates that class $KMP_2^{(a)}$ is clearly characterized by a mean curve higher than in $KMP_1^{(a)}$. This figure shows also a time lag between these mean curves. The spring high flow event seems to occur one month earlier in class $KMP_1^{(a)}$. The feature related to the time lag is also valid for the modal and median curves. Consequently, the spring high flow event generally becomes less important and occurs earlier from the year 1981.

A classification in $k = 3$ classes, $KMP_1^{(b)}$, $KMP_2^{(b)}$ and $KMP_3^{(b)}$ is also applied by this method. The composition of these classes of respective sizes 13, 16 and 11 is shown in Figure 7.4. The heterogeneity index $PHI_{mean}(S; KMP_1^{(b)}, KMP_2^{(b)}, KMP_3^{(b)}) = 0.62$ is the lowest value among those obtained with other classifications. Consequently, its score is the largest, $SC = 0.20$ (Table 7.2). Therefore, in terms of homogeneity gain, this classification is the best. Figure 7.4 shows that $KMP_3^{(b)}$ mainly represents the years prior to 1975, while $KMP_2^{(b)}$ mainly represents the years after 1986 and $KMP_1^{(b)}$ the years between 1975 and 1986. According to Figure 7.7f-h, class $KMP_2^{(b)}$ seems to correspond to earlier and smaller

Basis \ p	2	6	10	14	18
Fourier	308 744	48 266	37 571	36 957	36 875
Functional PCA	36 577	36 427	36 408	36 391	**36 385**
Haar	300 893	46 900	43 944	39 880	39 153
Best-Entropy	75 573	75 033	74 942	74 900	74 893

Table 7.3 – Distortions for the Romaine River station
Bold character indicates smallest value

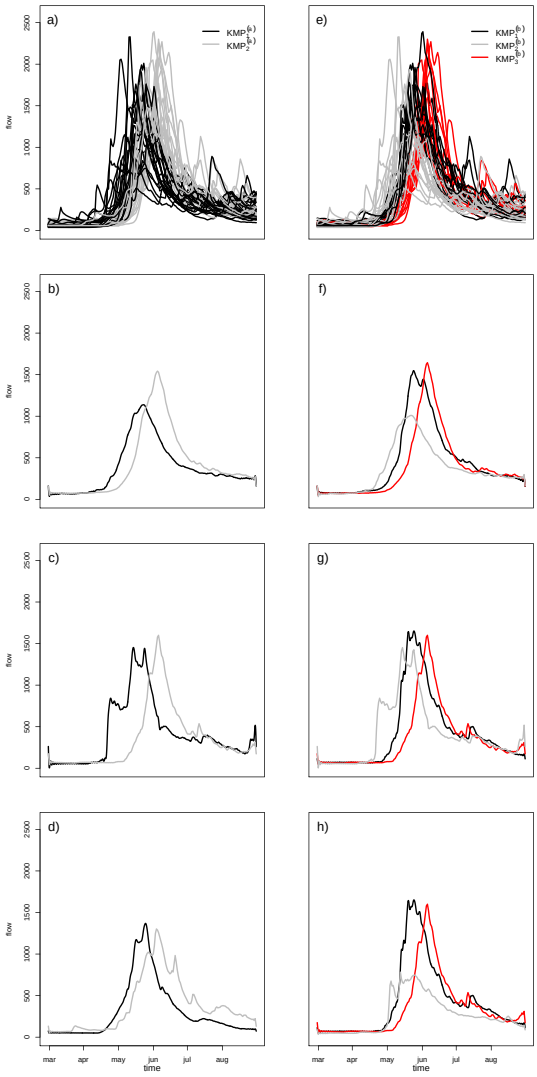


Figure 7.7 – Centrality curves for the two and three classes obtained using the projection-based curve clustering method, for the Romaine River station
a, e) all curves b, f) mean curves c, g) median curves d, h) modal curves

hydrographs and $KMP_3^{(b)}$ corresponds to higher and latter hydrographs. Class $KMP_1^{(b)}$ is the intermediary class between $KMP_2^{(b)}$ and $KMP_3^{(b)}$. Spring streamflow events seem to be evolving towards smaller streamflows which occur earlier.

Finally, the different functional classification methods are applied to 13 other RHBN stations. It is noticed that results are generally similar to those of the Romaine River sta-

tion. Due to size limitations, we only present in Figure 7.9 the mean curves corresponding to two or three classes, according to stations, using the KMP method. According to the obtained results of the various functional classification methods, we can conclude that:

1. KMP method can clearly distinguish between the spring hydrograph types. Figure 7.9 shows that classes obtained using this method are visually distinct.
2. KMB approach is more appropriate to detect unusual years. The classification into two classes using this method gives classes with very different sizes.
3. CM method provides classes that are less distinctive than those from the KMP method.
4. Based on the KMP method, there are two main spring hydrograph types for each station. The first type gathers hydrographs with large volume, large peak and late start date, and the second includes hydrographs with a low volume, low peak and early start date.

Comparison between multidimensional and functional results

The comparison with an usual multidimensional classification method is based on 25 hydrological variables which describe three of the five characteristics of hydrologic types suggested by Richter et al. (1996), namely magnitude, duration and rate of change. The 25 indices are as follows: monthly discharges (6 variables), monthly maximum and minimum discharges (12 variables), and monthly discharge ratios (5 variables), duration of event (1 variable) and the Julian date of the maximum flow (1 variable). The results of the multidimensional classification are given in detail for the Romaine River station and are resumed for the RHBN stations.

First, for the Romaine River station, a PCA is performed to isolate the first five principal components which explain 81% of the variance in the data described by the 25 variables. Then, an ascendant HC was applied on scores of these five components to identify the different classes. The dendrogram, represented on Figure 7.8a, indicates that a classification in two classes is appropriate. Figure 7.4 illustrates the composition of the two obtained classes M_1 and M_2. Class M_1 contains 6 years, and the second class M_2 contains 34 years. According to Figures 7.8b-7.8d, class M_1 gathers hydrographs with smallest peak and volume and early start date. On the other hand, class M_2 is characterized by highest peak and volume and late start date.

The Silhouette index (introduced in Section 7.2) is computed, on the raw data $x_i = (x_i(t_1), \ldots, x_i(t_T))$, in order to compare the different classification results obtained using functional and multidimensional approaches. Table 7.4 presents the Silhouette index corresponding to the different methods. Accordingly, the best classifications for the Romaine River station would be the 2 and 3 classes obtained by the KMP method ($KMP^{(a)}$ and $KMP^{(b)}$). The corresponding $\bar{s}$ are the highest (0.2748 and 0.2620 respectively). These functional classifications outperform classes obtained by the multidimensional approach (M) for which $\bar{s}$ is 0.2293. The worst classification results for the Romaine River station are those produced by the CM approach with $\bar{s} = 0.0824$ and the KMB approach ($KMB^{(a)}$ and $KMB^{(b)}$) with $\bar{s} = 0.0841$ and $\bar{s} = 0.1076$.

The multidimensional (M) approach is also performed on the RHBN stations. The Silhouette indices corresponding to the obtained classes by the functional and M methods are given in Table 7.4. According to this criterion, the M method often gives better results than CM and KMB methods. However, the KMP method leads to better results for all stations, except for Metabetchouane where the M method is the best and the $\bar{s}$ cannot be evaluated for the KMP method. It is also noticed that, in some cases, the Silhouette index

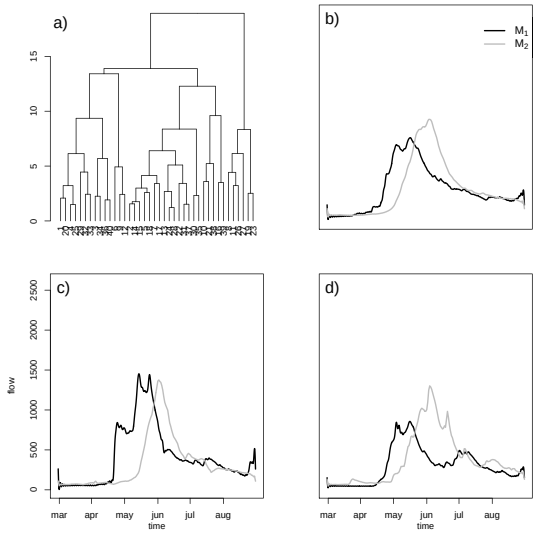


Figure 7.8 – Dendrogram and centrality curves corresponding to the multidimensional ascendant HC, for the Romaine River station
a) dendrogram b) mean curves c) median curves d) modal curves

cannot be calculated for the classes obtained using the KMB method. This is due to the existence of classes containing a single element. In terms of the Silhouette criterion, this method (KMB) gives the worst results compared to the two other functional approaches (KMP and CM), due to the proximity measure used (quadratic bias) which takes mainly into account the volume. Therefore, the conclusions obtained for the RHBN stations are generally similar to those obtained for the Romaine River station.

7.4 Discussion

In the case of the Romaine River station, in terms of homogeneity gain, a classification in three classes by the KMP method is the best among those presented, since its $PHI = 0.62$ is the lowest (see Table 7.2). Moreover, Figure 7.7e shows three clearly distinct groups of curves which is confirmed by the corresponding mean curves (Figure 7.7f). Then, the CM method leads to a classification in two classes CM_1 and CM_2, which is in second rank in terms of PHI (see Table 7.2). However, Figures 7.7a-7.7d show that the obtained two classes $KMP_1^{(a)}$ and $KMP_2^{(a)}$ are clearly distinctive compared to classes CM_1 and CM_2. Consequently, the index PHI should be always accompanied by a graphical checking. For the 13 other stations, classifications in two or three classes obtained by the KMP method (Figure 7.9) are also visually distinct.

For the Romaine River station, the comparison of the functional classification methods with the classical approach (M) indicates that the first class M_1 is included in class $KMP_2^{(b)}$. Furthermore, the classification obtained with the M method takes only into account the dates and does not consider the peak height. In fact, class M_1 regroups years with early events and irregular curves while class M_2 includes the 34 others curves without any regard for a particular feature; nor the date (e.g. years 1985 and 1994), nor the height (years 1972 and 1997), nor the speed (years 1978 and 1982), see Figure 7.10. However, the classification by the KMP method takes into account the starting and ending dates, the peak height, the hydrograph duration and shape. Indeed, the KMP

	Functional					Multidimensional
	CM	$KMB^{(a)}$	$KMP^{(a)}$	$KMB^{(b)}$	$KMP^{(b)}$	M
Romaine	0.0824	0.1076	**0.2748**	0.0841	0.2620	0.2293
Darmouth	0.1204	0.0682	**0.2191**	0.0737	-	0.1602
Rimouski	0.1511	0.0149	**0.2261**	0.0155	0.2259	0.1226
Beaurivage	0.0733	-0.0074	**0.1923**	-0.0038	0.1306	0.1238
Eaton	0.0932	-0.0272	0.1409	-0.0762	**0.1472**	0.0945
Picanoc	0.1858	-	0.3073	-	**0.3190**	0.1486
Croche	0.2309	-	0.2323	-	**0.2719**	0.2148
Sainte-Anne	0.1806	0.0638	**0.2043**	-	0.1870	0.1414
Metabetchouane	0.2365	-	-	-	-	**0.2678**
Chamouchouane	0.0657	0.0303	**0.2333**	0.0417	-	0.1945
Mistassibi	0.1021	-	**0.2328**	-	0.2246	0.1437
Moisie	0.1405	0.1305	0.3518	-	**0.3525**	0.2653
Harricana	0.1363	0.0267	**0.2645**	0.0048	0.2427	0.2248
À la Baleine	0.1487	0.1197	**0.3861**	0.0863	0.3613	0.3473

Table 7.4 – Classification results based on the mean $\bar{s}$ of the Silhouette indices $s(i)$, for the Romaine River and RHBN stations
Bold character indicates highest value for each station.
A dash (-) indicates that the Silhouette index cannot be computed (e.g. class with one observation)
CM: Functional hierarchical classification based on modal curve,
KMB$^{(a)}$ (and **KMB**$^{(b)}$): Functional k-means with Bregman divergence in 2 (and 3) classes,
KMP$^{(a)}$ (and **KMP**$^{(b)}$): Functional k-means with projection-based curve in 2 (and 3) classes,
M: Multidimensional hierarchical classification

method separates the previous cited years in two different classes (see Figure 7.10), namely $KMP_1^{(b)}$ and $KMP_3^{(b)}$. More precisely, class $KMP_1^{(b)}$ is characterized by the high speed of the hydrograph rise, the early start date and the lower peak whereas class $KMP_3^{(b)}$ is characterized by lower speed of the hydrograph, the later start date and the higher peak. The cutting in three classes with the KMP method is possible thanks to a homogeneity gain but a cutting in three with the M method seems unreasonable (Figure 7.8a).

For the Romaine River and the RHBN stations, according to the Silhouette criterion (Table 7.4), the M method performs better than the two functional methods KMB and CM. This indicates, that numerical compression of discharge time series by using appropriate indices or variables can preserve most information contained in the data. Indeed, hydrological time series usually show strong internal dependencies and high autocorrelations allowing a better numerical compression (see Weijs et al. (2013)). Therefore, by adequately choosing the indices, a large part of the information can be preserved. On the other hand, given a list of available indices or variables, the choice of which to include in the multidimensional classification is important but subjective, and the number of possible subsets of variables to include could be great. Thus, the choice of indices directly influences the classification results. However, a functional classification method has the advantage of being automatic without making choices of indices and thus, it can improve results. This is the case of the KMP approach which shows better performance compared to the M approach, in terms of the Silhouette criterion. Although, there are still subjective choices to be made with functional methods such as the divergence, the bandwidth parameter (to estimate the mode), they are secondary. Indeed, these choices are less fundamental than variable selection in the multidimensional context. In the latter, the result is directly and significantly related to the way the series and the variables to include are extracted. Other choices are common to both settings, for instance, thresholds or centers of the k-means.

Figure 7.11 summarizes results obtained for the Romaine River station. From this

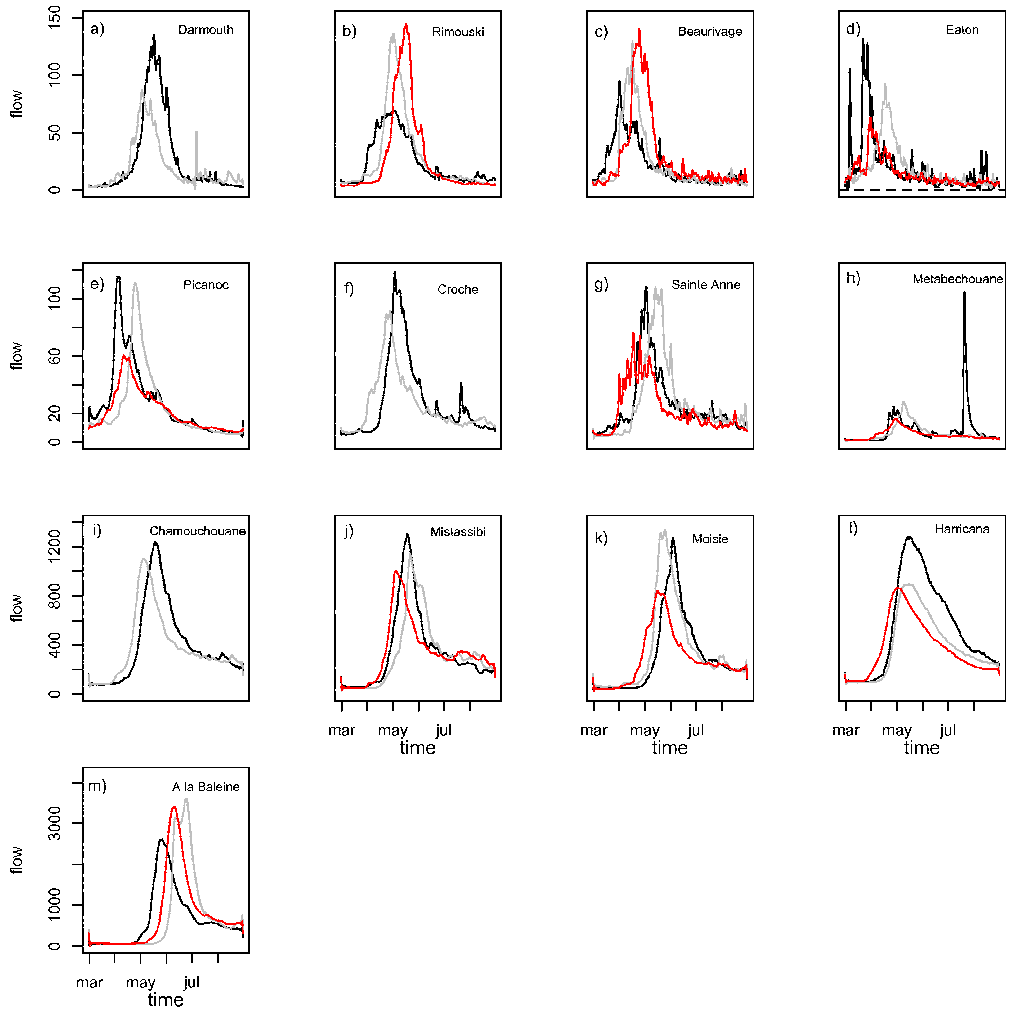


Figure 7.9 – Mean curves obtained using the k-means method with projection-based curve into two or three classes for different stations

figure, we can notice that classes $KMB_2^{(b)}$, $KMP_2^{(a)}$ and $KMP_3^{(b)}$ contain hydrographs that start late and have high volume and peak. On the other hand, classes $KMB_3^{(b)}$, $KMP_1^{(a)}$ and $KMP_2^{(b)}$ contain hydrographs that start early and have low peak and low volume. In the same way, the result from the 13 RHBN stations indicates that there are two main hydrograph types in Quebec (see Figure 7.9). A first hydrograph type is characterized by a large volume, large peak and late start date, and a second type is characterized by a low volume, low peak and early start date. In Quebec, high spring runoff is caused by the melting of large quantities of snow. Indeed, during the summer, liquid precipitation contributes directly to surface runoff since precipitation accelerates the snowmelt process. Consequently, when snow melt occurs late in the Summer, it combines with heavy rainfall to lead to extreme events.

From Figure 7.4 and Figure 7.11, the appearance frequency of the curves in classes $KMP_1^{(a)}$ and $KMP_2^{(a)}$, for the Romaine River station, seems to have changed during the last years (approximately since year 1981). Indeed, the frequency of late hydrographs has decreased whereas hydrographs that start early and that are characterized by a low peak and a low volume became more frequent. This behavior could be explained by climate

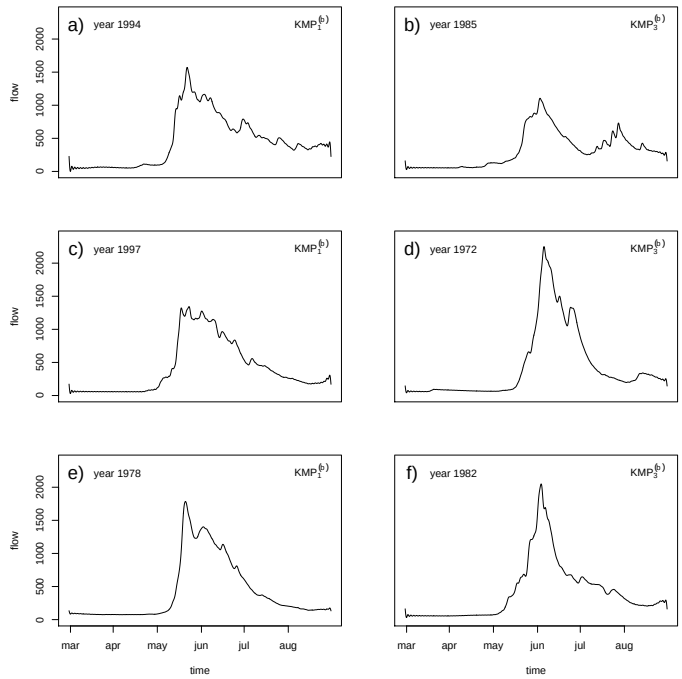


Figure 7.10 – Some curves of class M_2 according to classes $KMP_1^{(b)}$ or $KMP_3^{(b)}$

variability. However, the relatively short length of the series cannot affirm this. Indeed, climate has a significant influence on rainfall, river streamflow, and snowmelt. The variability and trends in the climate are influenced by oceanic and atmospheric oscillations on a large scale, known as teleconnections (e.g. Hurrell and Van Loon (1997), Rogers (1997)). Among the most known atmospheric oscillations, one can mention the El Niño-Southern Oscillation ($ENSO$). $ENSO$ was shown to have a strong impact on hydro-climatic variables in a number of regions throughout the globe (e.g. Cullen et al. (2002), Nazemosadat et al. (2006), Modarres and Ouarda (2013) and Ouachani et al. (2013)). These oscillations determine the large scale atmospheric circulation and can affect the watershed hydrological regime for a given year. Figure 7.12 shows that the period preceding the year 1981 was dominated by low phases of $ENSO$, while the period posterior to 1981 was dominated by high phases of $ENSO$. This seems to be related to the two classes identified by methods $KMP^{(a)}$ (see Figure 7.4). Indeed, events classified before 1981 seem to be mostly characterized by large peaks, large volumes and late starts, while events classified after 1981 are characterized by low peaks, low volumes and early starts. Thus, the developed application in this work can be extended by the characterization of different obtained classes using the climatic oscillations indices and other climatic factors.

7.5 Summary and concluding remarks

The purpose of the present chapter is the classification of streamflow hydrographs using the FDA framework. Three functional classification methods are considered, namely, descendant hierarchical classification based on modal curve (CM), k-means method with Bregman divergences (KMB) and k-means method with projection-based curve (KMP). These functional classification methods are presented and adapted to streamflows. Although this work covers the classification of streamflow hydrographs, the presented method-

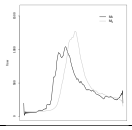
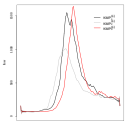
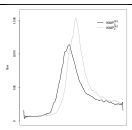
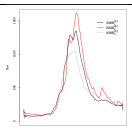
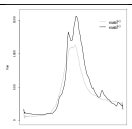
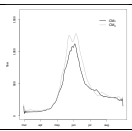
Method	Classes	Size			Main features of classes	Mean curves
		Total	1961-1980	1981-2000		
Multi-dimensional	M_1	6	3	3	- Low peak and early start	
	M_2	34	17	17	- Medium peak and late start	
Functional framework	$KMP_1^{(b)}$	13	7	6	- Large peak, large volume and large duration	
	$KMP_2^{(b)}$	16	4	12	- Low peak, low volume, and early start	
	$KMP_3^{(b)}$	11	9	2	- Large peak and late start	
	$KMP_1^{(a)}$	20	6	14	- Low peak, low volume, and early start	
	$KMP_2^{(a)}$	20	14	6	- Large peak, large volume and late start	
	$KMB_1^{(b)}$	12	9	3	- Medium peak, large volume and large duration	
	$KMB_2^{(b)}$	4	2	2	- Large volune	
	$KMB_3^{(b)}$	24	9	15	- Low peak and low volume	
	$KMB_1^{(a)}$	35	17	18	- Medium peak, medium volume	
	$KMB_2^{(a)}$	5	3	2	- Large peak, large volume and large duration	
	CM_1	23	11	12	- No big difference between the two classes	
	CM_2	17	9	8		

Figure 7.11 – Main features of the obtained classes for the Romaine River station

ology is general and can therefore be applied to other hydrological events, for example, to the classes of droughts curves, storms, and rainfall.

Applications are carried out for hydrological stations from the province of Quebec (Canada), including the Romaine River and 13 other RHBN stations. Functional approaches are compared with those obtained using a multidimensional method based on 25 extracted hydrological variables which describe magnitude, duration, and rate of change of a hydrograph. For the Romaine River station, an appropriate functional classification is obtained with the k-means method with projections, in two and three classes. In fact, in terms of the Silhouette criterion, classification using this method gives better results than classification obtained using the multidimensional method. An advantage of functional approaches is that they allow to cover the whole spring streamflow event and not only partially through some of its features. For all the RHBN considered stations, the different classification results allow to identify two main spring streamflow types. The first is characterized by large volume, large peak and late start date, and the second is characterized by low volume, low peak and early start date.

The presented method is applied to streamflow hydrographs in each station separately. Consequently, the different obtained spring streamflow types characterize only the temporal variability of the spring streamflow events at a given station. Since the stream hydrograph is an integration of spatial and temporal variations in water input, the presented method could be adapted and extended in order to account for spatial variability in a regional classification context.

A natural extension to the work presented herein consists in linking the shape classi-

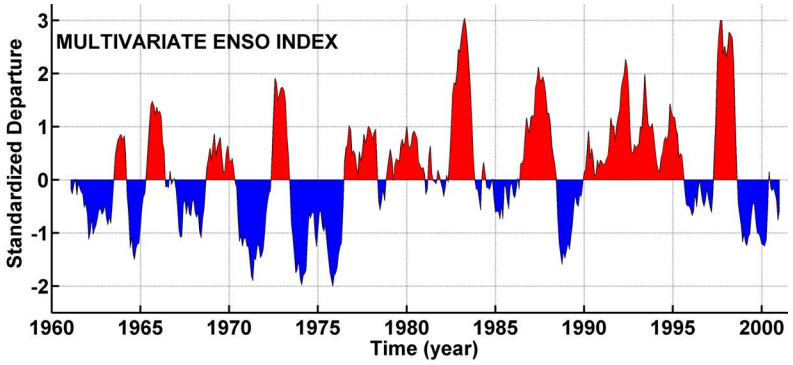


Figure 7.12 – El Niño Southern Oscillation time series over the studied period

fication obtained by FDA to a risk measure (return period for instance). The idea is to derive hydrograph shapes that correspond to different return periods, identifying hence (in a quantitative manner) the types of hydrographs that correspond to highly recurrent events on one side, and extremely rare events on the other. This hydrograph classification is very useful for design and management purposes. This is in fact an extension to the concept of flood quantile, except that the focus is no longer on quantiles corresponding to a single variable (peak for instance) or based on the joint distribution of two or more variables (peak, volume and duration for instance), but on quantiles representing the whole hydrograph. Future efforts can also focus on adapting the distance measure to the classification objective (risk analysis, total inflow quantification, etc.) and to the data record length and quality. This can only be achieved through the application of the functional classification method to a number of case studies.

To conclude, functional methods allow to cover all the information contained in a discharge time series. In fact, classical approaches have the problem of the dimensionality growth when the hydrograph is characterized by a large number of variables, but also algorithm, error and subjectivity in the evaluation. This is important when one is interested in classification of the whole hydrograph and not only a part of its features.

Chapter 8

Change point detection of flood events using a functional data framework

Contents

Résumé en français

Au cours de la classification non supervisée des hydrogrammes, considérée dans le chapitre précédent, nous avons obtenu des classes dont les courbes de chacune d'elles correspondent à une période de temps rapprochée. L'interprétation des classes obtenues nous a ainsi amené à considérer l'éventualité d'un changement de régime des crues au cours du temps. Dans ce chapitre, afin d'étudier plus en détails cette hypothèse, nous avons appliqué des méthodes de détection de rupture sur ces hydrogrammes afin de tester l'éventualité d'un tel changement. Nous considérons les méthodes pour données fonctionnelles qui testent l'existence d'un (ou plusieurs) changement(s) dans la moyenne fonctionnelle des données. Ces techniques sont rappelées avant d'être appliquées à deux stations hydrologiques du Québec. Avant de conclure, les résultats sont comparés avec ceux obtenus grâce à une approche multivariée Bayesienne, très utilisée dans le cadre hydrologique.

Ce chapitre entre également dans le cadre du projet de collaboration, cité précédemment au chapitre 7, entre le laboratoire EQUIPPE et l'INRS. Les résultats ont été obtenus avec les contributions de Mohamed Ali Ben Alaya (INRS, Québec), Fateh Chebana (INRS, Québec), Sophie Dabo-Niang (Université Charles de Gaulle) et Taha B. M. J. Ouarda (Masdar Institute, Abu Dhabi).

8.1 Introduction

Detection of changes in hydrological data is of interest to better understand hydrological regimes, to detect changes and separate events. Change in a series can occur in

numerous ways, for instance, gradually or abruptly, and can affect the mean, median, variance, autocorrelation, or almost any other aspect of the data. In the future, regions that are relatively sheltered from wind storms, heat waves, droughts and floods, may no longer be in a warmer climate (see Goudie (2006)). Detection of changes in long time series of hydrological data is an important and difficult issue of increasing interest. Change point detection in hydrology are essential to characterize the impacts of climate disturbances on hydrological regimes (see Kingston et al. (2011)). It is then very important, particularly where we observe changes in the frequency and/or in the intensity of various forms of extreme weather events. Detection of eventual changes in collected data of hydrologic time series is thus obviously an important step before performing any descriptive or predictive analysis.

There is a large literature on change point testing in scalar or vector time series. For example, Kundzewicz and Robson (2004) gave a general guidance on the methodology for change detection in hydrological records. Wong et al. (2006) proposed a relational method for discrete data. Change point analysis is addressed both in classical and Bayesian statistics. Classical statistical methods usually consist of performing several kinds of tests to confirm or reject the hypothesis of change. Most of them address slope or intercept change in linear regression models, see for instance the works of Solow (1987), Easterling and Peterson (1995) and Vincent (1998). Bayesian statistics are performed to obtain a statistical distribution for the change point and eventually a distribution for the other model parameters. Seidou and Ouarda (2007) proposed a Bayesian method of multiple change point detection in multiple linear regression. This method is numerically efficient and does not involve time consuming Markov Chain Monte Carlo (MCMC) simulation as opposed to other Bayesian change point detection methods. The procedure was initially designed to detect a change in the relationship between a set of explanatory variables and dependent variables. Using time variable as an explanatory variable, this approach can detect the change point in a given time series.

A flood event is an integration of spatial and temporal variations in water input, storage and transfer processes within a catchment (see Hannah et al. (2000)). Particularly, discharge time series is the main source of information for studying flood events. Thus, information that it contains is essential to determine the severity and frequency of a flood event. This information may present a change in flood characteristics for a given watershed and change point detection methods are used to detect the presence of such change. However, the major drawback of classical change point detection approaches is that they consider only a single feature of the flood event (such as maximum flows) and do not exploit all the information contained in the discharge time series of the flood event. In fact, these classical methods consist in a substantial simplification of the overall hydrological event. To circumvent this problem, a pragmatic solution is to consider the discharge time series as a curve, using the functional data analysis (FDA) framework.

Indeed, currently a dynamic research touches the development of adapted statistical tools that allow analyzing functional data. Many tools existing in the univariate and multivariate literature are adapted to the functional context, see e.g. Dabo-Niang et al. (2010), Fischer (2010) and Cuevas (2014). The first application of FDA to the hydrological context refers to Chebana et al. (2012) and focused on exploratory analysis as well as outlier detection of hydrographs. Chebana et al. (2012) showed that FDA is more general, flexible and representative of the real hydrological phenomena. Recently, Ternynck et al. (2014) proposed various alternatives to adapt FDA to streamflow hydrographs classification, and showed that classes obtained using functional approaches are more representative than those obtained using a traditional multivariate hierarchical classification method. Indeed,

functional classification takes into account the whole characteristic of a flood event including, for example, its shape, the peak or the starting date.

Furthermore, some authors also investigated the change point detection method in the context of functional data, see e.g. Chapters 6, 14 and 16 in Horváth and Kokoszka (2012). More precisely, we are interested in the work of Berkes et al. (2009) in which an approach is proposed to test the assumption of a common functional mean in presence of independent functional data. Their test is invalid for functional time series since it does not take into account the temporal dependence. Hörmann and Kokoszka (2010) as well as Aston and Kirch (2011) recognized the limitation of the test of Berkes et al (2009) and proposed a modification of this test by introducing a consistent long run variance estimator. The drawback in using this estimate is the necessity to choose a bandwidth parameter. To avoid selecting this parameter, Shao and Zhang (2010) proposed a self-normalization based test in the univariate time series setup, where an inconsistent normalization matrix is introduced to accommodate the dependence. Then, Zhang et al. (2011) adapted this work to the case of functional dependent data.

The aim of the present work is to introduce and adapt the FDA framework to change point detection of flood event. The chapter is structured as follows. After a presentation of the data set and the study area in Section 8.2, the proposed functional change point detection method is presented in Section 8.3. The proposed method is then applied, in Section 8.4, to the case of flood events in two stations from the province of Quebec, Canada. Results are compared with those obtained using the Bayesian method of Seidou and Ouarda (2007). Finally, a discussion and a conclusion are given in Section 8.5.

8.2 Data description and data smoothing

We consider the data series of a daily flow (m^3 s^{-1}) for two hydrological stations in Quebec, namely the Romaine River and the Moisie River stations. Figure 8.1, indicates the geographical locations of these two stations in the province of Quebec, Canada.

Figure 8.1 – Geographical locations of Romaine and Moisie rivers stations in the province of Quebec, Canada

For the Romaine river station, the area of the drainage basin is 13000 km^2 and available data cover the period from 1961 to 2000. Thus, according to the available data set, we have $n = 40$ years of observations for the Romaine river station. For the Moisie river station, the area of the drainage basin is 19000 km^2 and the data series is available from 1968 to 1991. Hence, we have $n = 24$ years of observations. For a given station, this chapter consider daily flows recorded from March 1 to August 31, corresponding to flood events occurring in spring and summer in Quebec. In fact, in Quebec, floods are mainly caused by snow melting during this period.

Let $x_i = (x_i(t_1), \ldots, x_i(t_T))$, $i = 1, \ldots, n$, be the set of n discrete observations of daily flows, where each $t_j \in \mathcal{T} \subset \mathbb{R}^+$, $j = 1, \ldots, T$, is the j^{th} record time point from time subset $\mathcal{T}$ from March 1 to August 31 which include the set $\{1, \ldots, T\}$. For a fixed observation i, corresponding to a year, each set of measurements $(x_i(t_1), \ldots, x_i(t_T))$ is converted to a functional data denoted $\{X_i(t), t \in \mathcal{T} \subset \mathbb{R}^+\}$, by using a smoothing technique. In this work, this is done through the technique based on Fourier series expansion. In fact, the periodicity of the data can justify the use of Fourier basis. The obtained curves, for the two considered stations, are displayed in Figure 8.2.

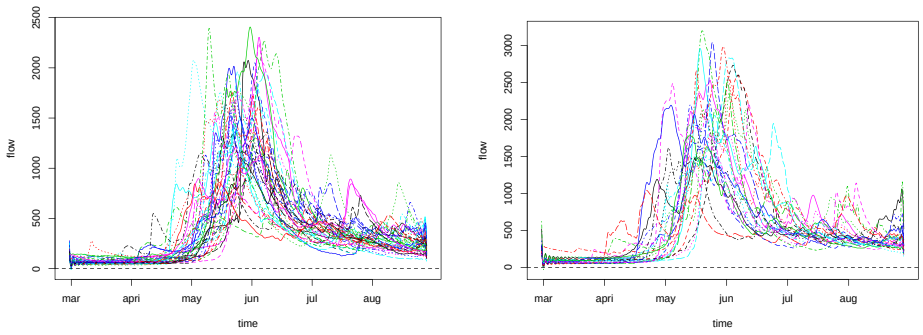


Figure 8.2 – Obtained curves for the Romaine (left panel) and Moisie (right panel) rivers stations, considering the period March-August

In the two following sections, some functional change point detection methods are explained and applied on these two stations.

8.3 Functional change point detection method

The main idea of this work is to test whether the mean of the functional observations $X_1, \ldots, X_n$ remains constant over time. Note that in this situation, the mean is a function, and the change can be not only in the average level of this function, but also in its shape. We assume that $X_i(t) = \mu_i(t) + \epsilon_i(t)$, $i = 1, \ldots, n$, where $\mu_i(t)$ denotes the functional mean and $\epsilon_i(t)$ is a zero-mean functional sequence. We wish to test the null hypothesis $H_0 : \mu_1(t) = \mu_2(t) = \cdots = \mu_n(t)$ against the alternative H_1 in which the data can be divided into several consecutive segments. The existence of change points means that the mean is constant within each segment, but changes from segment to segment. The change can occur at any point i and we want to test whether it occurs or not. The simplest case is the situation of one unknown change point k in the mean, that is two consecutive segments. In this case, the alternative hypothesis is written $H_1 : \mu_1(t) = \mu_2(t) = \cdots = \mu_k(t) \neq \mu_{k+1}(t) = \cdots = \mu_n(t)$.

Berkes et al. (2009) proposed an approach to test the assumption of a common functional mean for independent data. This approach is based on the quantity $P_k(t)$ which measures a deviation between the mean of the functional observations $X_1, \ldots, X_k$ and that of $X_{k+1}, \ldots, X_n$. This quantity is defined by

$$P_k(t) = \frac{k(n-k)}{n}\{\widehat{\mu}_k(t) - \widetilde{\mu}_k(t)\}$$

where $\widehat{\mu}_k(t) = \frac{1}{k}\sum_{1 \leq i \leq k} X_i(t)$ and $\widetilde{\mu}_k(t) = \frac{1}{n-k}\sum_{k \leq i \leq n} X_i(t)$. We note that the variability at the end points is attenuated by a parabolic weight function. If the mean changes, the

difference $P_k(t)$ is large for some values of k and t. To deal with the infinite dimension of the observations (curves), we consider the projections of the functions $P_k(\cdot)$ on the principal components of the data. In fact, principal component analysis represents functional data as $X_i(t) = \mu(t) + \sum_{1 \leq l \leq \infty} \eta_{i,l} \nu_l(t)$ where $\mu(t)$ is the functional mean, $\eta_{i,l}$ are the scores and $\nu_l(t)$ are the eigenfunctions of the covariance operator (see, e.g. Hall and Hosseini-Nasab (2006), Horváth and Kokoszka (2012)). These projections can be expressed in terms of functional scores which can be easily computed by using the R package *fda*. We consider the estimated scores $\widehat{\eta}_{i,l}$ corresponding to the L largest eigenvalues given by

$$\widehat{\eta}_{i,l} = \int \{X_i(t) - \overline{X}_n(t)\} \widehat{\nu}_l(t) dt, \qquad i = 1, \ldots, n, \quad l = 1, \ldots, L$$

with $\overline{X}_n(t)$ is the sample mean function and $\widehat{\nu}_l(t)$, $l = 1, \ldots, L$, are the estimated eigenfunctions of the covariance operator. It is supposed that $k = \lfloor n\alpha \rfloor$ where $\alpha \in (0, 1)$ and $\lfloor \cdot \rfloor$ denotes the integer part. Note that $P_k(t)$ does not change if the $X_i(t)$ are replaced by $X_i(t) - \overline{X}_n(t)$. Hence, we can rewrite $P_k(t)$ as

$$P_k(t) = \sum_{1 \leq i \leq k} \Big(X_i(t) - \overline{X}_n(t) \Big) - \frac{k}{n} \sum_{1 \leq i \leq n} \Big(X_i(t) - \overline{X}_n(t) \Big)$$

Consequently, the projections are defined by $\int P_k(t) \widehat{\nu}_l(t) dt = \sum_{1 \leq i \leq n\alpha} \widehat{\eta}_{i,l} - \frac{\lfloor n\alpha \rfloor}{n} \sum_{1 \leq i \leq n} \widehat{\eta}_{i,l}$ and are used for testing the constancy of the mean function. For this purpose, the following statistic is considered

$$S_{n,L} = \frac{1}{n^2} \sum_{l=1}^{L} \widehat{\lambda}_l^{-1} \sum_{k=1}^{n} \left(\sum_{1 \leq i \leq k} \widehat{\eta}_{i,l} - \frac{k}{n} \sum_{1 \leq i \leq n} \widehat{\eta}_{i,l} \right)^2$$

where $\widehat{\lambda}_1 > \ldots > \widehat{\lambda}_L$ denote the L-estimated eigenvalues. The test rejects the hypothesis H_0 of constant functional mean if $S_{n,L}$ is greater than the corresponding critical value, tabulated in Berkes et al. (2009) and in Horváth and Kokoszka (2012).

The previous test does not take the temporal dependence into account. Some authors recognize this limitation and propose some improvements. Hörmann and Kokoszka (2010) as well as Aston and Kirch (2011) proposed a modification of this test by introducing a consistent long run variance estimator but a bandwidth parameter must be selected. To avoid selecting this parameter, Shao and Zhang (2010) proposed a self-normalization based test in the univariate time series setup, where an inconsistent normalization matrix is introduced to accommodate the dependence. Zhang et al. (2011) adapted this work to the case of functional dependent data and gave the corresponding critical values of the modified test. The normalization matrix which takes into account the alternative, considered in Zhang et al. (2011), is defined as

$$V_{n,\hat{\eta}}(k, L) = \frac{1}{n^2} \left[\sum_{t=1}^{k} \left\{ S_{n,\hat{\eta}}(1, t) - \frac{t}{k} S_{n,\hat{\eta}}(1, k) \right\} \left\{ S_{n,\hat{\eta}}(1, t) - \frac{t}{k} S_{n,\hat{\eta}}(1, k) \right\}' \right.$$
$$\left. + \sum_{t=k+1}^{n} \left\{ S_{n,\hat{\eta}}(t, n) - \frac{n-t+1}{n-k} S_{n,\hat{\eta}}(k+1, n) \right\} \left\{ S_{n,\hat{\eta}}(t, n) - \frac{n-t+1}{n-k} S_{n,\hat{\eta}}(k+1, n) \right\}' \right]$$

with $S_{n,\hat{\eta}}(t_1, t_2) = \sum_{i=t_1}^{t_2} \hat{\eta}_i$ for $1 \leq t_1 \leq t_2 \leq n$ and with the estimated score vector $\hat{\eta}_i = (\hat{\eta}_{i,1}, \hat{\eta}_{i,2}, \ldots, \hat{\eta}_{i,L})'$, $i = 1, 2, \ldots, n$. The process $T_{n,\hat{\eta}}(k, L)$ is also considered and is defined as

$$T_{n,\hat{\eta}}(k, L) = \frac{1}{\sqrt{n}} \left\{ S_{n,\hat{\eta}}(1, k) - \frac{k}{n} S_{n,\hat{\eta}}(1, n) \right\}, \qquad k = 1, 2, \ldots, n$$

and the statistic of the test is defined as

$$G_{n,\hat{\eta}}(L) = \sup_{k=1,2,\ldots,n-1} \left\{ T_{n,\hat{\eta}}(k,L)' V_{n,\hat{\eta}}^{-1}(k,L) T(k,L) \right\}.$$

The test rejects hypothesis H_0 if $G_{n,\hat{\eta}}(L)$ is greater than the corresponding critical value. The critical values are tabulated in the work of Shao and Zhang (2010).

In the following section, the approaches studied in Berkes et al. (2009) and in Zhang et al. (2011) are applied to the discharge times series of Romaine and Moisie rivers stations.

8.4 Results and discussion

The functional change point detection methods explained in Section 8.3 are applied to the data set presented in Section 8.2. Results are compared with those obtained using the change point detection method of Seidou and Ouarda (2007) applied separately to the peak, the duration and the volume of the flood event occurring in spring and summer.

The Romaine river station

To apply these functional change point detection methods, it is necessary to perform a functional principal component analysis. In fact, we deal with the scores corresponding to the first principal components. For the Romaine river station, we choose the first four principal components because it represents 83% of the explained variance. In the statistical hypothesis testing, we fix the error of the first kind to 5%. By applying the functional method of Berkes et al. (2009) to the Romaine river station, we obtain a change point at the year 1984. It means that we can split the set of curves into the two following segments $1961 - 1984$ and $1985 - 2000$, of size 24 and 16 respectively. Figure 8.3 shows the mean curves of each two segments. From this figure, we note that the two obtained segments have different peaks; the peak of the first segment is significantly higher than the second. One can see that not only the peak changes, but also the duration of flood event, the volume, and the date of peak as well. Indeed, in both segments, usually, the floods usually start at the same time, but it remains longer in the first.

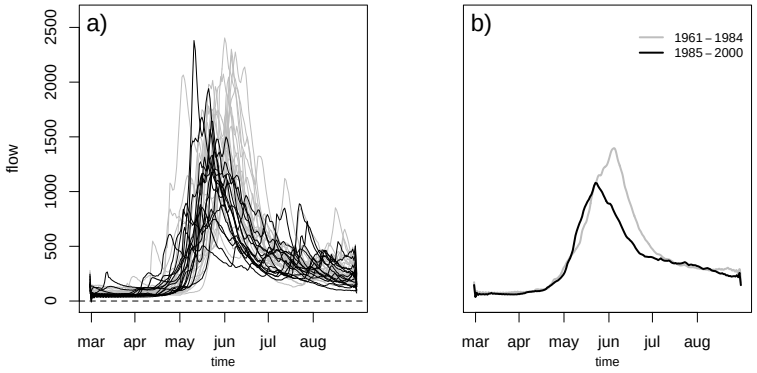


Figure 8.3 – Segments obtained at the first iteration using method of Berkes et al. (2009) for the Romaine river station
In the left: all the curves by segments. In the right: the mean curves of each two segments.

In a second step, we reiterate the procedure on the obtained two segments and a change point is detected on the first segment at year 1968. Consequently, on this station, we obtain

three segments corresponding to the periods 1961 – 1968, 1969 – 1984 and 1985 – 2000 of respective size 8, 16 and 16. Figure 8.4 displays the obtained segments. According to the mean curves, we can notice that flood events of the second segment begin before those of the first, however floods in both segments end at the same time. It means that the duration of flow during the second segment is larger than during the first. Moreover, we note also that these two segments have almost the same peak. Even if the two segments have similar peak, this method seems to detect difference in the duration of the flood event. Concerning the third segment, its floods begin at the same time than those of the second segment but end before. Thus, the second segment can be considered as an intermediate period which enables the transition of flood characteristics from the characteristics of the first segment to characteristics of the third segment. To conclude, for the Romaine river station, the functional change point method of Berkes et al. (2009) has detected two change points, the first at year 1968 and the second at year 1984. Flood events of this station are divided into three periods: the first with very large floods which begin later, a second intermediate period, and a third period characterized by less important floods which start early.

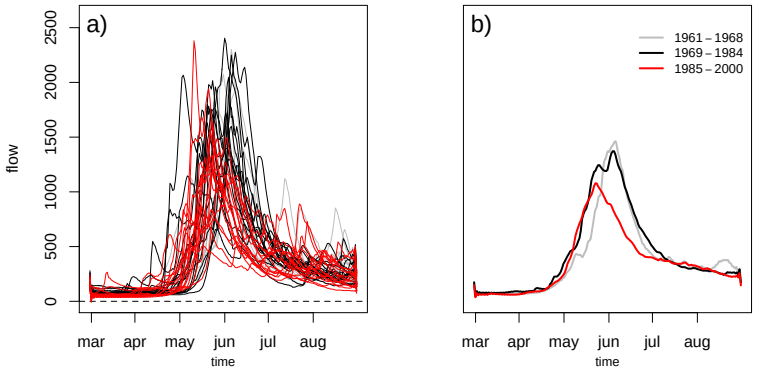


Figure 8.4 – Segments obtained at the second iteration using method of Berkes et al. (2009) for the Romaine river station
In the left: all the curves by segments. In the right: the mean curves of each two segments.

We now apply the change point detection method of Zhang et al. (2011) which accommodates the temporal dependence. A change point is detected at year 1978 which produces the two following segments, 1961 – 1978 and 1979 – 2000, of respective size 18 and 22, displayed on Figure 8.5. We note that this change point is between the two change points detected by the method of Berkes et al. (2009). This arises because the method of Zhang et al. (2011) takes into account the temporal dependence of the series during the change point detection process, which is not the case with the method of Berkes et al. (2009).

To compare functional change point results with a traditional method, we apply the Bayesian approach of Seidou and Ouarda (2007) to the peak, the volume and the duration of flood events separately. The method of Seidou and Ouarda (2007) based on the duration detects a change point at year 1987. However, the same method based on the volume and the peak detects a change point at year 1985, which is much closer to the first change point detected by functional approach of Berkes et al. (2009) at year 1984. This Bayesian approach based on the volume and the peak could not detect the second change point. This is due to the fact that this change does not affect the peak and the volume, but it affects mostly the time of occurring. Functional approach allows detecting this change point because it directly employs all data of a discharge time series, thus containing most available information on shape, timing, and duration, etc. The results obtained with the

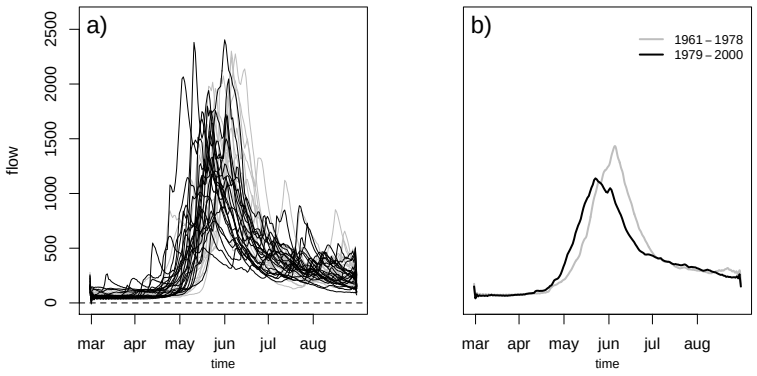


Figure 8.5 – Segments obtained using method of Zhang et al. (2011) for the Romaine river station
In the left: all the curves by segments. In the right: the mean curves of each two segments.

method of Zhang et al. (2011) are not directly compared with those obtained with the method of Seidou and Ouarda (2007) because this last method does not take into account the temporal dependence.

The Moisie river station

For the Moisie river station, we select the first four principal components because it represents 85% of the explained variance. In the hypothesis testing, we fix the error of the first kind to 5%. The method of Berkes et al. (2009) allows to obtain a change point at the year 1981. It means that we can split the set of curves into the two following segments, $1968 - 1981$ and $1982 - 1991$, of size 14 and 10 respectively. We reiterate the procedure on the obtained two segments and we do not detect any change point. For this station, we can conclude that this method allows detecting just one change point at years 1981. Figure 8.6 shows the mean curves of the flood hydrographs corresponding to the obtained segments. This figure shows that floods in the two segments occur at the same date, but those of the first segment last longer and have a larger peak.

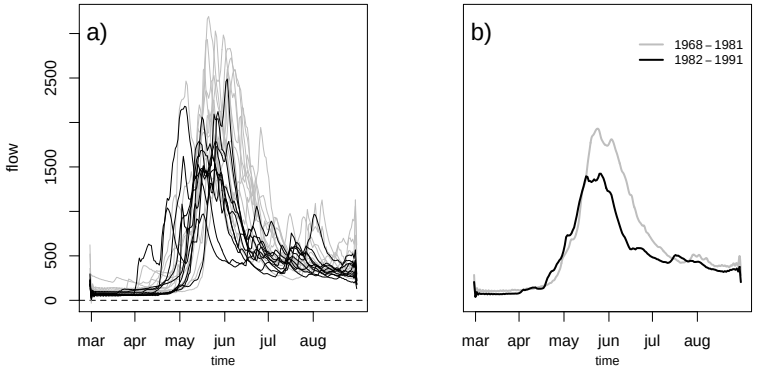


Figure 8.6 – Segments obtained using method of Berkes et al. (2009) for the Moisie river station
In the left: all the curves by segments. In the right: the mean curves of each two segments.

By the method of Zhang et al. (2011), a change point at year 1984 is detected which

produces the two following segments, $1968 - 1984$ and $1985 - 1991$, of respective size 17 and 7. Figure 8.7 shows that maximum floods are different on these two periods: the floods begin at the same time but ends early during the second period.

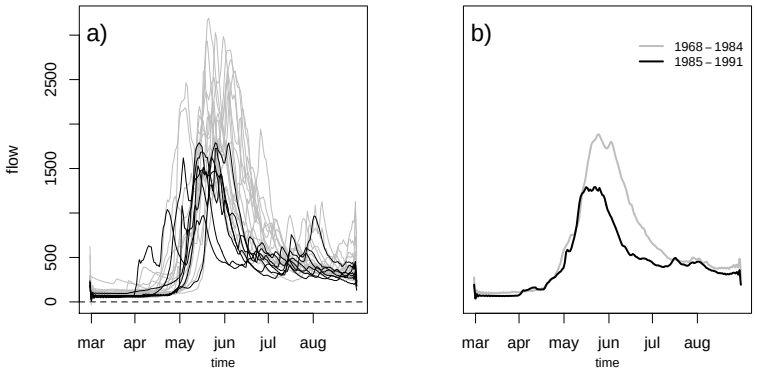


Figure 8.7 – Segments obtained using method of Zhang et al. (2011) for the Moisie river station
In the left: all the curves by segments. In the right: the mean curves of each two segments.

We also test on this station the existence of a change point on the peak, the volume and the duration separately using the method of Seidou and Ouarda (2007). Only, the method based on the peaks detects a change point at year 1978. This change point is much closer than that detected using the method of Berkes et al. (2009)

Discussion

For the Romaine river station, the detected change points are consistent with classes obtained in Chapter 7 with the functional k-means method based on the projected curves. In fact, in this chapter we note a change around year 1981 in the appearance frequency of the curves in each two obtained classes. Note that the aim of the comparison with a classical approach is not to show that functional approaches perform better or worse, but is to check whether functional approaches give results that are consistent with those obtained using a traditional approach. When we are interested in only one characteristic of the flood event, such as the peak, the volume or the duration, it is recommended to use a traditional univariate approach. However, the functional approach takes into account all the characteristics of the flood event simultaneously, so if we do not have a preference over any characteristic on the flood event, it is preferred to use a functional approach.

In hydrological change point analysis, if a significant change is detected, it is important to try to understand the cause. Indeed, change in hydrological characteristics can be caused by climate variability or climate change but many other explanations are possible, such as change caused by man (urbanization, water abstraction etc.), natural catchment changes, and problem linked to data. The best way to improve the understanding of change is to gather as much information as possible, using, e.g., information about change in the catchment. In addition, related variables, like temperature and precipitation, can help to determine whether changes in flow can be explained by climatic factors. Indeed, streamflows depend strongly on the spatial distribution of precipitation in a watershed, and on the interactions between temperature and precipitation which determine whether precipitation falls as rain or snow (Ben Alaya et al. (2014)).

Based on obtained results, for Romaine and Moisie rivers stations, the frequency of late

floods has decreased with time, whereas floods that start early and that are characterized by a low peak and a low volume become more frequent. This behaviour could be explained by climate variability. However, when period of hydrological records is short, climate variability easily gives rise to apparent change. Thus, because of climate variability, records of 30 years or less are certainly too short to confirm the presence of hydrological change. Indeed, oceanic and atmospheric oscillations on a large scale, known as teleconnections, influence the variability and trends in the climate (e.g. Hurrell and Van Loon (1997), Rogers (1997)). The North Atlantic Oscillation (NAO), El Nino-Southern Oscillation (ENSO) and Pacific Decadal Oscillation (PDO) are among the most known atmospheric oscillations. These oscillations can affect the hydrological watershed. Thus, the developed application in this work can be extended by the characterization of different obtained flood period using the climatic oscillations indices and other climatic factors.

8.5 Conclusion

The purpose of the present chapter is to apply change point detection methods on flood hydrographs using functional data framework. Two functional change point approaches are presented and adapted to flood events. Applications are carried out for two hydrological stations in the province of Quebec, Canada. The presented functional approaches are compared to a classical Bayesian approach applied on the peak, the volume and the duration of the flood event separately. Based on this comparison, functional approaches give results that are consistent with those obtained using the traditional univariate approach. Moreover, for the Romaine river, the obtained changes are consistent with some results obtained with a functional classification method in chapter 7. The functional approach could consider not only the duration, volume, peak, and date of the peak, but also the entire information contained in the discharge time series of the flood event. Finally, the presented methodologies are general and can therefore be applied to other hydrological events, for example, to the classes of droughts curves, storms, and rainfall.

Conclusion générale et perspectives

Conclusion

Dans ce travail de thèse, nous avons principalement considéré la modélisation non paramétrique par la méthode à noyau. Plus particulièrement, nous avons modélisé deux catégories de données, à savoir les données spatialement dépendantes et les données fonctionnelles. Nous avons proposé, dans les deux premières parties de la thèse, des estimateurs des fonctions de densité de probabilité, ainsi que de son mode, et de régression en présence de variables α-mélangeantes. La dernière partie de ce travail concerne l'application de l'estimation de la densité à des méthodes de classification pour données spatiales ou fonctionnelles ainsi que de méthodes de détection de rupture pour données fonctionnelles. Notons qu'une partie des méthodes utilisées dans ces applications n'est pas nécessairement basée sur la méthode à noyau.

Nous avons commencé par introduire une nouvelle approche à noyau pour la modélisation de données spatiales. D'une part, cette approche a permis de construire un nouvel estimateur de la fonction de densité de probabilité spatiale. La particularité de l'approche proposée est de tenir compte à la fois des valeurs observées et de la position des sites où ont lieu les observations. Sous certaines hypothèses, la convergence presque sûre, ponctuelle et uniforme, avec vitesse est obtenue. Cet estimateur de la densité est ensuite utilisé dans l'estimation du mode spatial, dont la convergence est également étudiée. Une méthode de classification non supervisée, basée sur l'estimateur du mode spatial, est ensuite proposée.

Dans la continuité, nous avons étendu l'approche proposée précédemment pour l'estimation de la fonction de régression spatiale dans l'objectif de faire de la prédiction spatiale. Une étude des propriétés asymptotiques de cet estimateur de la régression est conduite. Nous obtenons les convergences presque complète et en moyenne d'ordre q, avec vitesse, de l'estimateur. Nous proposons ensuite un prédicteur spatial issu de l'estimation de la régression. Les comportements pratiques de l'estimateur et du prédicteur sont étudiés par leur application sur des données simulées puis environnementales.

Ensuite, nous avons adapté l'estimateur précédent de la fonction de régression spatiale au cadre de données fonctionnelles. Sous certaines conditions, nous obtenons la convergence en moyenne quadratique, avec vitesse, de cet estimateur. Un prédicteur spatial est également proposé. Enfin, nous étudions l'estimateur par le biais de simulations.

Nous poursuivons la modélisation non paramétrique de la régression pour données fonctionnelles dans le contexte des séries temporelles, toujours dans un souci de prise en compte d'une dépendance des données dans l'estimation. Nous supposons que les erreurs du modèle de régression sont autocorrélées. Dans cette situation, l'estimateur à noyau classique ne tient pas compte de l'information contenue dans les erreurs. Dans ce travail, nous proposons une procédure permettant de transformer le modèle original de sorte que

le modèle transformé tienne compte de cette information et dont le terme d'erreur devienne non corrélé. Il s'agit alors d'appliquer l'estimateur à noyau classique sur le modèle transformé. La normalité asymptotique du nouvel estimateur est obtenue sous certaines conditions. Une étude de simulation et l'application sur des données réelles ont permis de mettre en évidence l'apport de notre méthode en présence d'erreurs fortement corrélées.

La dernière partie de ce mémoire est essentiellement consacrée à des applications. Des applications de l'estimateur de la densité spatiale à une méthode de classification, proposée au début de ce travail de thèse, sont établies sur des données simulées et réelles. Ensuite, nous considérons des outils issus de la statistique pour données fonctionnelles pour résoudre des problèmes de classification non supervisée lorsque les objets à traiter sont de nature fonctionnelle. Les données traitées concernent les débits de rivières Québécoises. Plusieurs méthodes ont été appliquées, dont l'une basée sur l'estimation de la densité et du mode. Les résultats obtenus nous ont ensuite amené à considérer des méthodes de détection de rupture sur ces données.

Ainsi, ce travail de thèse nous a permis de contribuer à différentes thématiques de la statistique de manière théorique mais aussi appliquée. Notre apport théorique concerne l'introduction et l'étude de nouvelles approches non paramétriques pour modéliser des données spatiales et/ou fonctionnelles. Les applications effectuées concernent principalement la classification non supervisée pour données spatiales, puis fonctionnelles ainsi que des méthodes de détection de rupture pour données fonctionnelles. Tout au long de ce travail, des questions et remarques sont apparues laissant place à quelques perspectives de recherche que nous développons ci-après.

Perspectives

A la fin de chaque chapitre de ce manuscrit, des questions apparaissent et alimentent nos perspectives de recherche que nous pouvons résumer de la manière suivante :

- Tout d'abord, la nouvelle approche proposée au chapitre 3 pour estimer la fonction de régression spatiale r du modèle $Y_{\mathbf{i}} = r(X_{\mathbf{i}}) + \epsilon_{\mathbf{i}}$ est basée sur l'hypothèse que $r(x) = \mathbb{E}[Y|X = x]$. Cette approche qui combine le noyau sur les observations avec celui sur les sites pourrait être étendue au cas où la fonction de régression représente un élément de la distribution conditionnelle autre que l'espérance conditionnelle. Par exemple, l'estimation du quantile conditionnel est motivée par son intérêt comme outil de prévision alternatif à l'espérance conditionnelle. Nous considérons un champ aléatoire $\{(X_{\mathbf{i}}, Y_{\mathbf{i}})\}$ à valeurs dans $\mathbb{R}^d \times \mathbb{R}$ où les sites $\mathbf{i}$ sont tels que $\mathbf{i} \in \mathcal{I}_{\mathbf{n}} \subset \mathbb{Z}^N$. Pour $\alpha \in]0,1[$, le quantile conditionnel d'ordre α, dénoté $q_\alpha(x)$ est une solution de l'équation $F^x(q_\alpha(x)) = \alpha$. En adaptant l'approche développée au chapitre 3, on envisage un estimateur de la distribution conditionnelle F^x défini par

$$\widehat{F}^{x_{\mathbf{j}}}(y_{\mathbf{j}}) = \begin{cases} \dfrac{\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_1\left(\frac{x_{\mathbf{j}}-X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_2\left(\frac{\mathbf{j}-\mathbf{i}}{\rho_{\mathbf{n}}}\right) K_3\left(\frac{y_{\mathbf{j}}-Y_{\mathbf{i}}}{h_{\mathbf{n}}}\right)}{\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_1\left(\frac{x_{\mathbf{j}}-X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_2\left(\frac{\mathbf{j}-\mathbf{i}}{\rho_{\mathbf{n}}}\right)} & \text{si } \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_1\left(\frac{x_{\mathbf{j}}-X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_2\left(\frac{\mathbf{j}-\mathbf{i}}{\rho_{\mathbf{n}}}\right) \neq 0 \\ 0 & \text{sinon} \end{cases}$$

  où K_1 et K_2 sont des noyaux et K_3 est une fonction de distribution. De plus, $b_{\mathbf{n}}$, $\rho_{\mathbf{n}}$ et $h_{\mathbf{n}}$ sont des séquences de nombres positifs qui tendent vers 0. L'estimateur à noyau du quantile conditionnel $\widehat{q}_\alpha$ est lié à l'estimateur de la distribution conditionnelle de la manière suivante $\widehat{F}^{x_{\mathbf{j}}}(\widehat{q}_\alpha(x_{\mathbf{j}})) = \alpha$.
- Les modèles spatiaux que nous avons considérés supposent que la position des sites où ont lieu les observations est fixe. Cependant, dans la réalité cette hypothèse n'est

pas vérifiée dans toutes les disciplines. Nous souhaitons donc étendre l'étude de notre approche au cas où la position des sites est également aléatoire, dans un cadre spatial ou spatio-temporel, étendant ainsi les travaux de Menezes et al. (2010). Dans ce cas, la construction des estimateurs étudiés dans le cadre de cette thèse reste valide. Les changements surviennent dans les hypothèses faites sur les noyaux ainsi que sur les vitesses de convergence des estimateurs qui devraient inclure un terme lié à la fenêtre du noyau K_2. En effet, dans ce manuscrit, les quantités liées au noyau K_2 sont déterministes et ainsi les termes de biais des vitesses de convergence ne dépendent que du paramètre de lissage de K_1.

- Notons également que nous avons supposé que la région où ont lieu les observations est discrète. Nous envisageons dans un travail futur de généraliser notre approche au cas où l'ensemble spatial étudié est une région continue. Des travaux de la littérature concernent l'étude non paramétrique spatiale des fonctions de densité et de régression lorsque les sites sont à valeurs dans un espace continu (e.g. Biau (2003) ainsi que Dabo-Niang et Yao (2007) étudient les estimateurs à noyau classiques). Nous considérons ici que les sites d'observations $\mathbf{i}$ sont à valeurs dans $\mathbb{R}^N$ et pour $\mathbf{T} = (T_1, \ldots, T_N) \in \mathbb{R}_+^{*N}$, nous définissons la région rectangulaire $\mathcal{I}_{\mathbf{T}}$ par $\mathcal{I}_{\mathbf{T}} = \{\mathbf{i} = (i_1, \ldots, i_N) \in \mathbb{R}_+^N : 0 \leq i_k \leq T_k, k = 1, \ldots, N\}$. De plus, nous écrivons $\mathbf{T} \to \infty$ si $\min_{k=1,\ldots,N} T_k \to \infty$ et nous posons $\widehat{\mathbf{T}} = T_1 \times \cdots \times T_N$. Une extension possible de l'estimateur de la fonction de densité proposée au chapitre 2 pour un champ aléatoire $(X_{\mathbf{i}})_{\mathbf{i}}$ à valeurs dans $\mathbb{R}^d$ et continûment indexé est la suivante

$$f_{\mathbf{T}}(x_{\mathbf{j}}) = \frac{1}{\widehat{\mathbf{T}} b_{\mathbf{T}}^d \rho_{\mathbf{T}}^N} \int_{\mathcal{I}_{\mathbf{T}}} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{T}}}\right) K_2\left(\frac{\mathbf{j} - \mathbf{i}}{\rho_{\mathbf{T}}}\right) d\mathbf{i}$$

où K_1 et K_2 sont des noyaux sur $\mathbb{R}^d$ et $\mathbb{R}^N$ respectivement, $b_{\mathbf{T}}$ et $\rho_{\mathbf{T}}$ sont des fenêtres qui tendent vers 0 lorsque $\mathbf{T} \to \infty$. De la même manière, l'estimateur de la fonction de régression de $X_{\mathbf{i}} \in \mathbb{R}^d$ sur $Y_{\mathbf{i}} \in \mathbb{R}$ étudiée au chapitre 3 peut être étendu pour des champs aléatoires continûment indexés et défini par

$$r_{\mathbf{T}}(x_{\mathbf{j}}) = \frac{\frac{1}{\widehat{\mathbf{T}} b_{\mathbf{T}}^d \rho_{\mathbf{T}}^N} \int_{\mathcal{I}_{\mathbf{T}}} Y_{\mathbf{i}} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{T}}}\right) K_2\left(\frac{\mathbf{j}-\mathbf{i}}{\rho_{\mathbf{T}}}\right) d\mathbf{i}}{f_{\mathbf{T}}(x_{\mathbf{j}})} \quad \text{si } f_{\mathbf{T}}(x_{\mathbf{j}}) \neq 0$$

et $r_{\mathbf{T}}(x_{\mathbf{j}}) = \frac{1}{\widehat{\mathbf{T}}} \int_{\mathcal{I}_{\mathbf{T}}} Y_{\mathbf{i}} d\mathbf{i}$ si $f_{\mathbf{T}}(x_{\mathbf{j}}) = 0$.

- Une autre question est apparue au moment de l'application présentée au chapitre 3 sur les données du Jura, sur la concentration en métaux lourds dans le sol. En effet, ce jeu de données est constitué également de variables catégorielles comme, par exemple, la nature du sol. Actuellement, ce genre de données n'est pas pris en compte dans nos modèles. En s'inspirant des travaux de Li et Racine (2003), Racine et Li (2004), ..., nous souhaitons étendre nos modèles spatiaux aux données catégorielles et continues. En effet, en ignorant ces variables nous nous sommes privés d'information qui pourrait améliorer nos prédictions. Soit $(X_{\mathbf{i}})_{\mathbf{i}}$ un champ aléatoire pour lequel les observations sont constituées de variables catégorielles (discrètes), notées $X_{\mathbf{i}}^d$, et de variables continues, notées $X_{\mathbf{i}}^c$, ainsi on note $X_{\mathbf{i}} = (X_{\mathbf{i}}^d, X_{\mathbf{i}}^c)$. Dans la suite, nous adaptons les notations pour données non spatialisées de Racine et Li (2004) au cadre spatial qui nous intéresse. Nous supposons que $X_{\mathbf{i}}^d$ est un vecteur de taille $k \times 1$ et $X_{t,\mathbf{i}}^d$ dénote le t-*ème* élément de $X_{\mathbf{i}}^d$ pour $t = 1, \ldots, k$. Nous considérons que les valeurs que peuvent prendre $X_{t,\mathbf{i}}^d$ sont $0, 1, \ldots c_t - 1$ avec $c_t \geq 2$ et qu'il n'y

a pas d'ordre naturel dans $X_{\mathbf{i}}^d$. On pose $X_{\mathbf{i}}^d \in \mathcal{D} = \Pi_{t=1}^k \{0, 1, \ldots, c_t - 1\}$. Dans la suite, on considère la fonction suivante

$$l(X_{t,\mathbf{i}}^d, x_{t,\mathbf{j}}^d, \lambda) = \begin{cases} 1 & \text{si } X_{t,\mathbf{i}}^d = x_{t,\mathbf{j}}^d \\ \lambda & \text{si } X_{t,\mathbf{i}}^d \neq x_{t,\mathbf{j}}^d. \end{cases}$$

On définit également $d_{x_\mathbf{i}, x_\mathbf{j}} = \sum_{t=1}^k \mathbf{1}_{[X_{t,\mathbf{i}}^d \neq x_{t,\mathbf{j}}^d]}$ où $\mathbf{1}_{[\cdot]}$ est la fonction indicatrice. De plus, on a

$$L(X_{\mathbf{i}}^d, x_{\mathbf{j}}^d, \lambda) = \prod_{t=1}^k l(X_{t,\mathbf{i}}^d, x_{t,\mathbf{j}}^d, \lambda) = 1^{k - d_{x_\mathbf{i}, x_\mathbf{j}}} \lambda^{d_{x_\mathbf{i}, x_\mathbf{j}}}.$$

Par ailleurs, les variables $X_{\mathbf{i}}^c$ sont à valeurs dans $\mathbb{R}^p$ et nous utilisons $W(\cdot)$ la fonction noyau associée aux variables continues $X_{\mathbf{i}}^c$ et h la fenêtre correspondante. Avec ces notations, le noyau K_1 (sur les observations), considéré dans les chapitres 2 et 3, s'écrit

$$K_1 = L(X_{\mathbf{i}}^d, x_{\mathbf{j}}^d, \lambda) W\left(\frac{X_{\mathbf{i}}^c - x_{\mathbf{j}}^c}{h}\right)$$

Nous souhaitons donc adapter les estimateurs proposés aux chapitres 2 et 3 avec cette écriture de K_1 et ensuite étudier les nouveaux estimateurs obtenus.

- Un inconvénient avec la nouvelle approche proposée (chapitres 2, 3 et 4) pour l'estimation spatiale est la sélection simultanée des deux fenêtres. En effet, dans les applications proposées cette sélection se fait par validation croisée et requiert un temps assez important avant de déterminer les fenêtres optimales. Nous pouvons, par exemple, décider de fixer la fenêtre du noyau K_2 relatif aux sites comme cela est fait dans Kelejian et Prucha (2007) et que nous avons testé dans les applications du chapitre 4. Dans un travail futur, la fenêtre peut être définie comme une proportion du nombre de sites dont la proportion dépend de la dépendance spatiale présente dans l'échantillon. Pour cela, nous pouvons ajouter une étape préliminaire qui consisterait à mesurer la dépendance spatiale présente dans les données puis étudier le lien entre cette dépendance et la taille de la fenêtre.
- La nouvelle approche proposée (chapitres 2, 3 et 4) pour estimer les fonctions de densité de probabilité et de régression spatiales ne tient pas compte des *effets de bord* (voir le chapitre 5 page 154 dans Gaetan et Guyon (2008)) qui sont plus importants en statistique spatiale qu'en statistique temporelle. Une solution pourrait être de donner moins d'importance aux observations enregistrées sur les bords du domaine. Pour cela, une possibilité serait de définir des fenêtres différentes pour le noyau K_2 selon que le site étudié soit sur le bord du domaine ou non.
- Dans le chapitre 5, nous avons proposé une procédure permettant d'améliorer l'estimateur à noyau de la régression lorsque les variables explicatives sont fonctionnelles et le terme d'erreur est autocorrélé. Nous envisageons d'adapter cette procédure au cadre des données spatiales avec autocorrélation des erreurs. Le modèle de régression que nous souhaitons étudier est $Y_\mathbf{i} = r(X_\mathbf{i}) + u_\mathbf{i}$ lorsque le champ aléatoire étudié $(X_\mathbf{i}, Y_\mathbf{i})$ est à valeurs dans $\mathbb{R}^d \times \mathbb{R}$, le processus des erreurs $u_\mathbf{i}$ est autocorrélé, et les sites sont à valeurs dans $\mathbb{Z}^N$. Plus précisément, on admet que

$$u_\mathbf{i} = \lambda \omega_\mathbf{i}' U + \epsilon_\mathbf{i}$$

où le processus $\{\epsilon_\mathbf{i}\}$ est un bruit blanc, λ est un paramètre inconnu, $U = (u_\mathbf{1}, \ldots, u_\mathbf{n})'$ et $\omega_\mathbf{i} = (\omega_{\mathbf{i},\mathbf{j}})_{\mathbf{i},\mathbf{j} \in \mathcal{I}_\mathbf{n}}$ est un vecteur connu de taille $\widehat{\mathbf{n}} \times 1$ permettant de donner des

poids différents selon la proximité des sites **j** et **i** (voir Wang et al. (2012)). On envisage d'adapter la procédure développée au chapitre 5 à la situation précédemment exposée, pour estimer la fonction de régression r.

- Jusqu'à présent, les travaux développés dans le cadre du projet de collaboration Franco-Québécoise ont consisté à appliquer des méthodes de statistique pour données fonctionnelles sur les débits des rivières de différentes stations. A l'avenir, nous souhaitons intégrer l'information spatiale des différentes stations dans nos études dans un cadre régional. En effet, les stations voisines d'une station étudiée peuvent avoir de l'influence sur cette dernière.
- Dans la première partie de cette thèse, nous nous sommes essentiellement intéressés aux données multivariées alors que la seconde partie concerne plutôt des données fonctionnelles. Nous projetons de nous intéresser à un genre de données plus récent, à savoir les données multivariées fonctionnelles, étudiées par exemple dans Jacques et Preda (2014) dans le contexte de la classification non supervisée. A l'avenir, nous aimerions étudier un modèle de régression pour données spatiales dans le cas où la variable explicative $\mathbf{X} = \{\mathbf{X}(t)\}_{t\in[0,T]}$ est composée de p courbes $X^1(t), \ldots, X^p(t)$ et la variable réponse Y est réelle ou fonctionnelle.
- Pour terminer, une perspective future concerne l'extension de nos travaux à un autre cadre de dépendance spatiale, en particulier la m-dépendance (voir e.g. les travaux de Wang et Woodroofe (2014)).

General conclusion and perspectives

Conclusion

In this thesis, we mainly considered nonparametric modeling with the kernel method. Specifically, we modeled two kinds of data, namely spatially dependent data and functional data. In the two first parts of the thesis, we estimated probability density function, as well as its mode, and regression function, in presence of α-mixing variables. The last part of this work concerns the application of density estimation to classification methods for spatial or functional data as well as change point detection methods for functional data. We notice that certain of these methods used are not necessarily based on the kernel method.

We started by introducing a new kernel approach to modeling spatial data. On one hand, this approach led us to construct a new estimate of the spatial probability density function. The specificity of the proposed approach is to take into account both observed values and spatial locations where the observations are made. Under some assumptions, the pointwise and uniform almost sure convergences with rate were obtained. Then, this density estimate is used to estimate the spatial mode, whose consistency is also studied. An unsupervised classification method, based on spatial mode estimate, is then proposed.

In the continuity, we extended the previous proposed approach to spatial regression function estimation in order to make spatial forecasting. A study of the asymptotic properties of this regression estimate is carried out. We get almost complete and mean of order q consistencies, with rate, of the estimate. Then, we proposed a spatial predictor from the regression estimation. The practical behaviors of the estimate and of the predictor were studied by applying them on simulated and environmental data.

Then, we adapted the previous spatial regression function estimate for functional data. Under some assumptions, we get the quadratic mean consistency with rate, of this estimate. A spatial predictor is also proposed. Finally, we studied this estimate through some simulations.

We continue the nonparametric modeling of the regression function for functional data in the time series context, always concerned with taking into account the dependence of the data in the estimation. We suppose that the errors of the regression model are autocorrelated. In this situation, the classical kernel estimate does not take into account the information contained in the errors. In this work, we propose a transformation procedure for the original model so that the transformed model takes into account this information and with an uncorrelated error term. The question is then to apply the classical kernel estimate on this transformed model. Asymptotic normality of this new estimate is obtained under some conditions. A simulation study and application on real data have highlighted the benefits of our method in presence of highly correlated errors.

The last part of this manuscript is mainly devoted to applications. Some applications

of the spatial density estimate to a classification method, proposed at the beginning of this thesis, are made on simulated and real data. Then, we consider functional data analysis tools to solve unsupervised classification problems when the studied objects are of functional nature. The studied data concern rivers flows from Quebec. Several methods were applied, with one based on the estimation of the density and of the mode. The obtained results led us to consider change point detection methods on these data.

Thus, this thesis allowed us to contribute to several statistical issues in a theoretical but also applied way. Our theoretical contribution concerns the introduction and the study of new nonparametric approaches to modeling spatial and/or functional data. The applications made mainly concern unsupervised classification for spatial data, then functional as well as change point detection methods for functional data. During this work, some issues and remarks appeared leading to some research perspectives that we develop hereafter.

Perspectives

At the end of each chapter, some issues have emerged and enriches our research perspectives which can be summed up as follows:

- First of all, the new approach proposed in chapter 3 to estimate the spatial regression function r from model $Y_{\mathbf{i}} = r(X_{\mathbf{i}}) + \epsilon_{\mathbf{i}}$ is based on the hypothesis that $r(x) = \mathbb{E}[Y|X = x]$. This approach that combines a kernel on the observations and that on the sites could be extended to the case where the regression function represents an element of the conditional distribution other than the conditional expectation. For example, the conditional quantile estimation is motivated by its interest as alternative prevision tool to conditional expectation. We consider a random field $\{(X_{\mathbf{i}}, Y_{\mathbf{i}})\}$ valued in $\mathbb{R}^d \times \mathbb{R}$ where sites $\mathbf{i}$ are such that $\mathbf{i} \in \mathcal{I}_{\mathbf{n}} \subset \mathbb{Z}^N$. For $\alpha \in]0,1[$, the conditional quantile of order α, denoted $q_\alpha(x)$ is a solution of the equation $F^x(q_\alpha(x)) = \alpha$. By adapting, the approach developed in chapter 3, we consider an estimate of the conditional distribution F^x defined by

$$\widehat{F}^{x_{\mathbf{j}}}(y_{\mathbf{j}}) = \begin{cases} \dfrac{\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_1\left(\frac{x_{\mathbf{j}}-X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_2\left(\frac{\mathbf{j}-\mathbf{i}}{\rho_{\mathbf{n}}}\right) K_3\left(\frac{y_{\mathbf{j}}-Y_{\mathbf{i}}}{h_{\mathbf{n}}}\right)}{\sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_1\left(\frac{x_{\mathbf{j}}-X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_2\left(\frac{\mathbf{j}-\mathbf{i}}{\rho_{\mathbf{n}}}\right)} & \text{if } \sum_{\mathbf{i}\in\mathcal{I}_{\mathbf{n}}} K_1\left(\frac{x_{\mathbf{j}}-X_{\mathbf{i}}}{b_{\mathbf{n}}}\right) K_2\left(\frac{\mathbf{j}-\mathbf{i}}{\rho_{\mathbf{n}}}\right) \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

  where K_1 and K_2 are two kernels and K_3 is a distribution function. Moreover, $b_{\mathbf{n}}$, $\rho_{\mathbf{n}}$ and $h_{\mathbf{n}}$ are sequences of positive numbers tending to 0. The conditional quantile kernel estimate $\widehat{q}_\alpha$ is linked to the conditional distribution estimate in the following way $\widehat{F}^{x_{\mathbf{j}}}(\widehat{q}_\alpha(x_{\mathbf{j}})) = \alpha$.
- Spatial models that we have considered suppose spatial locations where observations are made to be fixed. However, in realty, this hypothesis is not verified in all disciplines. Therefore, we would like to extend the study of our approach to the case where the position of the sites is also random, in a spatial or spatio-temporal framework, thus extending the work of Menezes et al. (2010). In this case, the construction of studied estimators in this thesis remains valid. The changes appear in hypotheses made on the kernels as well as on the rate of convergence of the estimators which should include a term linked to the bandwidth of kernel K_2. Indeed, in this manuscript, quantities linked to the kernel K_2 are deterministic and thus bias terms of convergence rates depend only on the smoothing parameter of K_1.

- We also notice that we have supposed the domain where the observations are made to be discrete. In a future work, we will consider the generalization of our approach to the case where the studied spatial domain is a continuous region. Some works of the literature concern the nonparametric spatial study of density and regression functions when sites are valued in a continuous space (e.g. Biau (2003) as well as Dabo-Niang and Yao (2007) study classical kernel estimators). Here we consider that sites $\mathbf{i}$ are valued in $\mathbb{R}^N$ and for $\mathbf{T} = (T_1, \ldots, T_N) \in \mathbb{R}_+^{*N}$, we define the rectangular region $\mathcal{I}_{\mathbf{T}}$ by $\mathcal{I}_{\mathbf{T}} = \{\mathbf{i} = (i_1, \ldots, i_N) \in \mathbb{R}_+^N : 0 \leq i_k \leq T_k, k = 1, \ldots, N\}$. Moreover, we write $\mathbf{T} \to \infty$ if $\min_{k=1,\ldots,N} T_k \to \infty$ and let $\widehat{\mathbf{T}} = T_1 \times \cdots \times T_N$. A possible extension of the density function estimator, proposed in chapter 2, for a random field $(X_{\mathbf{i}})_{\mathbf{i}}$ valued in $\mathbb{R}^d$ and continuously indexed is

$$f_{\mathbf{T}}(x_{\mathbf{j}}) = \frac{1}{\widehat{\mathbf{T}} b_{\mathbf{T}}^d \rho_{\mathbf{T}}^N} \int_{\mathcal{I}_{\mathbf{T}}} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{T}}}\right) K_2\left(\frac{\mathbf{j} - \mathbf{i}}{\rho_{\mathbf{T}}}\right) d\mathbf{i}$$

where K_1 and K_2 are kernels on $\mathbb{R}^d$ and $\mathbb{R}^N$ respectively, $b_{\mathbf{T}}$ and $\rho_{\mathbf{T}}$ are bandwidths tending to 0 when $\mathbf{T} \to \infty$. In the same way, the regression function estimate of $X_{\mathbf{i}} \in \mathbb{R}^d$ on $Y_{\mathbf{i}} \in \mathbb{R}$ studied in chapter 3 can be extended for continuously indexed random fields and defined by

$$r_{\mathbf{T}}(x_{\mathbf{j}}) = \frac{\frac{1}{\widehat{\mathbf{T}} b_{\mathbf{T}}^d \rho_{\mathbf{T}}^N} \int_{\mathcal{I}_{\mathbf{T}}} Y_{\mathbf{i}} K_1\left(\frac{x_{\mathbf{j}} - X_{\mathbf{i}}}{b_{\mathbf{T}}}\right) K_2\left(\frac{\mathbf{j}-\mathbf{i}}{\rho_{\mathbf{T}}}\right) d\mathbf{i}}{f_{\mathbf{T}}(x_{\mathbf{j}})} \quad \text{if } f_{\mathbf{T}}(x_{\mathbf{j}}) \neq 0$$

and $r_{\mathbf{T}}(x_{\mathbf{j}}) = \frac{1}{\widehat{\mathbf{T}}} \int_{\mathcal{I}_{\mathbf{T}}} Y_{\mathbf{i}} d\mathbf{i}$ if $f_{\mathbf{T}}(x_{\mathbf{j}}) = 0$.

- There is another question that has arisen during the application presented in chapter 3 on the Jura data set, about heavy metal concentrations in the soil. Indeed, the data set is also constituted of categorical variables as, for instance, the nature of the soil. Currently, this kind of data is not taken into account in our models. Following the work of Li and Racine (2003)), Racine and Li (2004), ..., we would like to extend our spatial models to categorical and continuous data. Indeed, by ignoring these variables we forgot information that could improve our predictions. Let $(X_{\mathbf{i}})_{\mathbf{i}}$ a random field for which observations are constituted of categorical (discrete) variables, denoted $X_{\mathbf{i}}^d$, and of continuous variables, denoted $X_{\mathbf{i}}^c$, thus we write $X_{\mathbf{i}} = (X_{\mathbf{i}}^d, X_{\mathbf{i}}^c)$. In the following, we adapt the notations for non spatialized data of Racine and Li (2004) to the spatial framework we are interested in. We suppose that $X_{\mathbf{i}}^d$ is a vector of size $k \times 1$ and $X_{t,\mathbf{i}}^d$ denotes the *t-th* element of $X_{\mathbf{i}}^d$ for $t = 1, \ldots, k$. We consider that the possible values of $X_{t,\mathbf{i}}^d$ are $0, 1, \ldots c_t - 1$ with $c_t \geq 2$ and that there is no natural order in $X_{\mathbf{i}}^d$. Let $X_{\mathbf{i}}^d \in \mathcal{D} = \Pi_{t=1}^k \{0, 1, \ldots, c_t - 1\}$. In the following, we consider the following function

$$l(X_{t,\mathbf{i}}^d, x_{t,\mathbf{j}}^d, \lambda) = \begin{cases} 1 & \text{if } X_{t,\mathbf{i}}^d = x_{t,\mathbf{j}}^d \\ \lambda & \text{if } X_{t,\mathbf{i}}^d \neq x_{t,\mathbf{j}}^d. \end{cases}$$

We also define $d_{x_{\mathbf{i}}, x_{\mathbf{j}}} = \sum_{t=1}^k \mathbf{1}_{[X_{t,\mathbf{i}}^d \neq x_{t,\mathbf{j}}^d]}$ where $\mathbf{1}_{[\cdot]}$ is the indicator function. Moreover, we have

$$L(X_{\mathbf{i}}^d, x_{\mathbf{j}}^d, \lambda) = \prod_{t=1}^k l(X_{t,\mathbf{i}}^d, x_{t,\mathbf{j}}^d, \lambda) = 1^{k - d_{x_{\mathbf{i}}, x_{\mathbf{j}}}} \lambda^{d_{x_{\mathbf{i}}, x_{\mathbf{j}}}}.$$

Furthermore, variables $X_{\mathbf{i}}^c$ are valued in $\mathbb{R}^p$ and we utilize the kernel function $W(\cdot)$ associated with the continuous variables $X_{\mathbf{i}}^c$ and the corresponding bandwidth h.

With these notations, the kernel K_1 (on the observations), considered in chapters 2 and 3, is written

$$K_1 = L(X_{\mathbf{i}}^d, x_{\mathbf{j}}^d, \lambda) W\left(\frac{X_{\mathbf{i}}^c - x_{\mathbf{j}}^c}{h}\right)$$

Therefore, we would like to adapt the proposed estimate in chapters 2 and 3 with this writing of K_1 and then to study the new obtained estimators.

- A disadvantage with the new proposed approach (chapters 2, 3 and 4) for the spatial estimation, is the simultaneous selection of the two bandwidths. Indeed, in the proposed applications, this selection is made by cross validation and is quite time-consuming before determining optimal bandwidths. We can, for example, decide to fix the bandwidth of kernel K_2, relating to the sites, as it is done in Kelejian and Prucha (2007) and that we have tested in applications of chapter 4. In a future work, this bandwidth could be defined as a proportion of the number of sites whose the proportion depends on the spatial dependence presents in the sample. For this purpose, we can add a preliminary step which would consist in measuring the spatial dependence represented in the data and then studying the link between this dependence and the size of the bandwidth.
- The proposed approach (chapters 2, 3 and 4) for studying spatial probability density and regression functions does not take into account the *edge effect* (see chapter 5 page 154 in Gaetan and Guyon (2008)) which are more important in spatial statistics than in time series. One solution would be to attach less importance to observations recorded on the edges of the domain. For this purpose, one possibility could be to define different bandwidths for the kernel K_2 depending on whether or not the studied site is on the edge of the domain.
- In chapter 5, we proposed a procedure improving kernel estimate of the regression function when explanatory variables are functional and the error term is uncorrelated. We are looking at adapting this procedure to spatial data with autocorrelated errors. The regression model we want to study is $Y_{\mathbf{i}} = r(X_{\mathbf{i}}) + u_{\mathbf{i}}$ when the studied random field $(X_{\mathbf{i}}, Y_{\mathbf{i}})$ is valued in $\mathbb{R}^d \times \mathbb{R}$, the error process $u_{\mathbf{i}}$ is autocorrelated, and the sites are valued in $\mathbb{Z}^N$. Specifically, we assume that

$$u_{\mathbf{i}} = \lambda \omega_{\mathbf{i}}' U + \epsilon_{\mathbf{i}}$$

where the process $\{\epsilon_{\mathbf{i}}\}$ is a white noise, λ is an unknown parameter, $U = (u_{\mathbf{1}}, \ldots, u_{\mathbf{n}})'$ and $\omega_{\mathbf{i}} = (\omega_{\mathbf{i},\mathbf{j}})_{\mathbf{i},\mathbf{j} \in \mathcal{I}_{\mathbf{n}}}$ is a known vector of size $\widehat{\mathbf{n}} \times 1$ allowing to give different weights according to the proximity between sites $\mathbf{j}$ and $\mathbf{i}$. We are concerned by adapting the procedure developed in chapter 5 to the previous exposed situation, in order to estimate the regression function r.
- Up to now, works carried out in the framework of the French-Quebec collaborative project have consisted in applying statistical methods for functional data on rivers flows from different stations. In the near future, we want to include spatial information of the different stations in our studies in a regional framework. Indeed, neighbor stations of a studied station could influence this latter.
- In the first part of this thesis, we are mainly interested in multivariate data whereas the second part concerns rather functional data. We intend to deal with a more recent kind of data, namely multivariate functional data, studied for instance in Jacques and Preda (2014) in the context of clustering. In the future, we would like to study a regression model for spatial data in the case where the explanatory

variable $\mathbf{X} = \{\mathbf{X}(t)\}_{t\in[0,T]}$ is composed of p curves $X^1(t), \ldots, X^p(t)$ and the response variable is real or functional.

- Finally, a future perpective concerns the extension of our works to another spatial dependence framework, in particular the m-dependence (see e.g. the work of Wang and Woodroofe (2014)).

Annexe A

Rappels

A.1 Lemmas

This section regroups lemmas used in proofs of different chapters.

Lemma A.1. From Carbon et al. (1997)
Denote by $\mathcal{L}_r(\mathcal{F})$ the class of $\mathcal{F}$-measurable random variables X that satisfy : $\|X\|_r = (\mathbb{E}|X|^r)^{1/r} < \infty$. Suppose that $X \in \mathcal{L}_r(\mathcal{B}(E))$, $Y \in \mathcal{L}_r(\mathcal{B}(E'))$, $1 \leq r, s, t < \infty$ and $\frac{1}{r} + \frac{1}{s} + \frac{1}{t} = 1$. Then,

$$|\mathbb{E}XY - \mathbb{E}X\mathbb{E}Y| \leq C\|X\|_r\|Y\|_s\{\psi\left(\mathrm{Card}(E), \mathrm{Card}(E')\right)\varphi(\mathrm{dist}(E, E'))\}^{1/t}.$$

For bounded random variables with probability 1, we have :

$$|\mathbb{E}XY - \mathbb{E}X\mathbb{E}Y| \leq C\{\psi\left(\mathrm{Card}(E), \mathrm{Card}(E')\right)\varphi(\mathrm{dist}(E, E'))\}.$$

Lemma A.2. From Carbon et al. (2007)
Let the sets $S_1, S_2, \ldots, S_k$ containing each m sites and such that, for all $i \neq j$, and for $1 \leq i, j \leq k$, $\mathrm{dist}(S_i, S_j) \geq \delta_0$. Let $W_1, W_2, \ldots, W_k$ be a sequence of random variables with real values and measurable respectively with respect to $\mathcal{B}(S_1), \ldots, \mathcal{B}(S_k)$ and W_l with values in $[a, b]$. There exists a sequence of independent random variables $W_1^*, W_2^*, \ldots, W_k^*$ such that W_l^* has the same distribution as W_l and satisfies :

$$\sum_{l=1}^{k} \mathbb{E}|W_l - W_l^*| \leq 2k(b-a)\psi((k-1)m, m)\varphi(\delta_0).$$

A.2 Notions de statistique asymptotique

A.2.1 Mode de convergence

Convergence presque complète On dit que $(X_n)_{n\in\mathbb{N}}$ converge presque complètement vers une variable aléatoire réelle X, si et seulement si, $\forall \epsilon > 0$:

$$\sum_{n\in\mathbb{N}} \mathbb{P}[|X_n - X| > \epsilon] \quad < \quad \infty$$

et la convergence presque complète de $(X_n)_{n\in\mathbb{N}}$ vers X est notée par :

$$\lim_{n\to\infty} X_n \quad = \quad X, \quad p.c.$$

Vitesse de convergence presque complète La vitesse de convergence presque complète de $(X_n)_{n\in\mathbb{N}}$ vers X est de l'ordre de u_n, si et seulement si, $\forall \epsilon_0 > 0$:

$$\sum_{n\in\mathbb{N}} \mathbb{P}[|X_n - X| > \epsilon_0 u_n] \quad < \quad \infty$$

et nous écrivons :

$$X_n - X \quad = \quad O_{p.c.}(u_n)$$

Convergence presque sûre On dit que X_n converge presque sûrement vers X (ou X_n converge vers X avec probabilité 1) si

$$\mathbb{P}\left[\omega : \lim_{n\to\infty} X_n(\omega) \to X(\omega)\right] \quad = \quad 1.$$

On écrit $X_n \xrightarrow{p.s.} X$. Un caractérisation de la convergence presque sûre est que, pour $\epsilon > 0$ et $\forall n \geq m$ avec $m \to \infty$

$$\lim_{n\to\infty} \mathbb{P}[|X_n - X| \leq \epsilon] = 1.$$

Convergence en moyenne d'ordre q, $q \geq 0$ On dit que X_n converge en moyenne d'ordre q vers X et l'on écrit $X_n \xrightarrow{m.q.} X$ si

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X|^q) \quad = \quad 0$$

Convergence en probabilité On dit que $(X_n)_{n\in\mathbb{N}}$ converge en probabilité vers une variable aléatoire réelle X et on écrit $X_n \xrightarrow{p} X$, si et seulement si, $\forall \epsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}[|X_n - X| > \epsilon] \quad = \quad 0$$

Convergence en loi On dit que X_n converge en distribution (en loi) vers X et on écrit $X_n \xrightarrow{\mathcal{L}} X$ si $\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$ quand $n \to \infty$ en tout point de continuité x de F (fonction de répartition de X).

Quelques relations entre les différents modes de convergence

- La convergence presque complète implique la convergence presque sûre (ainsi que la convergence en probabilité).
- La convergence presque sûre implique la convergence en probabilité.
- La convergence en moyenne d'ordre q implique la convergence en probabilité.

A.2.2 Ordre de grandeur

Pour comparer les vitesses de convergence asymptotique de différents estimateurs, on utilise la famille de notation de Landau. Nous considérons en premier des ordres impliquant des séquences déterministes que l'on exprime avec les notions de "petit o" et de "grand O".

Définition A.3. Soit a_n une séquence déterministe (non stochastique) indexée par un entier positif $n = 1, 2, \ldots$ et C une constante positive. On écrit

- $a_n = O(1)$ si, quand $n \to \infty$, a_n reste bornée, i.e. $|a_n| \leq C$ (on dit alors que a_n est "bornée") ;

- $a_n = o(1)$ si $a_n \to 0$ quand $n \to \infty$;
- $a_n = O(b_n)$ si $a_n/b_n = O(1)$, ou de manière équivalente $a_n \leq Cb_n$;
- $a_n = o(b_n)$ si $a_n/b_n \to 0$ quand $n \to \infty$.

Considérons maintenant des séquences stochastiques que l'on exprimera avec les notions de "petit o_p" et de "grand O_p".

Définition A.4. Soit X_n une séquence de variables aléatoires réelles indexée par un entier positif $n = 1, 2, \ldots$ et C une constante positive. On dit que X_n est bornée en probabilité, si pour tout $\epsilon > 0$, il existe une constante positive C est un entier positif N tel que

$$\mathbb{P}[|X_n| > C] \leq \epsilon$$

pour tout $n \geq N$, où ϵ est un petit nombre positif arbitraire. On écrit

- $X_n = O_p(1)$ pour indiquer que X_n est bornée en probabilité ;
- $X_n = o_p(1)$ si $X_n \xrightarrow{p} 0$, où $\xrightarrow{p}$ dénote la convergence en probabilité (i.e. $\mathbb{P}[|X_n| > \epsilon] \to 0$ quand $n \to \infty$) ;
- $X_n = O_p(Y_n)$ si $X_n/Y_n = O_p(1)$, et $X_n = o_p(Y_n)$ si $X_n/Y_n = o_p(1)$.

Remark A.5. Notons que si $X_n = o_p(1)$ alors $X_n = O_p(1)$ (mais $X_n = O_p(1)$ n'implique pas $X_n = o_p(1)$). De plus, si $X_n = O(1)$ (bornée) alors $X_n = O_p(1)$ (mais $X_n = O_p(1)$ n'implique pas $X_n = O(1)$).

Theorem A.6. Soit $\{X_n\}_{n=1}^{\infty}$ une séquence de variables aléatoires réelles et soient a_n et b_n des séquences déterministes de nombres non négatifs. Alors

- Si $\mathbb{E}|X_n| = O(a_n)$ alors $X_n = O_p(a_n)$
- Si $\mathbb{E}(X^2) = O(b_n)$ alors $X_n = O_p(b_n^{1/2})$

A.3 Quelques inégalités

Pour prouver les résultats asymptotiques des estimateurs proposés dans le cadre de cette thèse, nous utilisons certaines inégalités que nous présentons ci-après.

A.3.1 L'inégalité de Markov

Soit X une variable aléatoire non-négative et supposons que $\mathbb{E}(X)$ existe. Pour $\varepsilon > 0$,

$$\mathbb{P}(X > \varepsilon) \quad \leq \quad \frac{\mathbb{E}(X)}{\varepsilon}$$

De plus, soit X une variable aléatoire réelle telle que $\mathbb{E}|X|^r < \infty$, alors

$$\mathbb{P}(|X| > \varepsilon) \quad \leq \quad \frac{\mathbb{E}|X|^r}{\varepsilon^r}$$

A.3.2 L'inégalité de Tchebychev (Chebyshev)

Cette inégalité permet d'évaluer la distance entre les valeurs prises par une variable aléatoire X et son espérance. Soit X une variable aléatoire définie sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ et qui admet une espérance $\mathbb{E}(X)$ et une variance $V(X)$. Alors pour tout $a > 0$

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \quad \leq \quad \frac{V(X)}{a^2}$$

Lemma A.7. Soit Y une variable aléatoire positive, définie sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$. On note $\mathbb{E}(Y)$ l'espérance de Y. Alors, pour tout $a > 0$,

$$\mathbb{P}(Y \geq a) \quad \leq \quad \frac{\mathbb{E}(Y)}{a}$$

A.3.3 L'inégalité de Hölder

Soit X une variable aléatoire dans L^p, $p \geq 1$ c'est à dire telle que $\mathbb{E}|X|^p < \infty$ et $\|X\|_p = (\mathbb{E}|X|^p)^{1/p}$. Soit Y une variable aléatoire dans L^q, $q \geq 1$. Si $p \geq 1$ et $q \geq 1$ tels que $\frac{1}{p} + \frac{1}{q} = 1$, alors

$$\|XY\|_1 \quad \leq \quad \|X\|_p \|Y\|_q$$

A.3.4 L'inégalité de Minkowski (triangulaire)

Soit X une variable aléatoire X dans L^q, $q \geq 1$. On a

$$\|X + Y\|_q \quad \leq \quad \|X\|_q + \|Y\|_q$$

A.4 Processus fortement mélangeants

Introduites par Rosenblatt (1956a), les conditions de mélange sont souvent considérées dans la littérature pour mesurer la dépendance faible entre les variables, permettant d'obtenir des vitesse de convergence des estimateurs. Il existe différents types de mélange : α-mélange, ϕ-mélange, ψ-mélange, ρ-mélange, β-mélange. Dans le cadre de cette thèse, on s'intéresse essentiellement aux processus α-mélangeants (ou fortement mélangeants), dont la définition et quelques propriétés sont données ci-après.

Définition A.8. Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé, $\mathcal{B}$ et $\mathcal{C}$ deux sous-tribus de $\mathcal{A}$. Le coefficient de α-mélange de $\mathcal{B}$ et $\mathcal{C}$ s'écrit

$$\alpha = \alpha(\mathcal{B}, \mathcal{C}) = \sup_{B \in \mathcal{B}, C \in \mathcal{C}} |\mathbb{P}(B \cap C) - \mathbb{P}(B)\mathbb{P}(C)| \tag{A.1}$$

On peut remarquer que $0 \leq \alpha(\mathcal{B}, \mathcal{C}) \leq 1/4$. De plus, si $\mathcal{B}$ et $\mathcal{C}$ sont indépendantes alors $\alpha(\mathcal{B}, \mathcal{C}) = 0$.

Définition A.9. Les coefficients de forte mélangeance d'un processus $X = (X_t)$ sont définis par

$$\alpha_X(h) = \sup_t \alpha\{\sigma(X_u, u \leq t), \sigma(X_u, u \geq t + h)\} \tag{A.2}$$

Nous allons définir deux sous-classes de séquences mélangeantes.

Définition A.10. On dit que la séquence $(\xi_n)_{n\in\mathbb{Z}}$ est arithmétiquement (ou algébriquement) α-mélangeante à une vitesse $a > 0$ si

$$\exists C > 0, \alpha(n) \leq Cn^{-a} \tag{A.3}$$

Cette séquence est dite géométriquement α-mélangeante si

$$\exists C > 0, \exists t \in (0,1), \alpha(n) \leq Ct^n \tag{A.4}$$

Le processus X est dit fortement mélangeant ou α-mélangeant, si $\alpha_X(h) \to 0$ quand $h \to \infty$. Si $\alpha_X(h)$ tend vers 0 à vitesse exponentielle, alors X est dit géométriquement fortement mélangeant. Enfin, on a les inégalités suivantes :

Inégalité de Billingsley : Si X et Y sont deux variables aléatoires bornées, on a

$$|cov(X,Y)| \leq 4\|X\|_\infty \|Y\|_\infty \alpha(\sigma(X), \sigma(Y)) \tag{A.5}$$

Inégalité de Davydov : Si $X \in L^q(P)$, $Y \in L^r(P)$ avec $q > 1$, $r > 1$ et $\frac{1}{q} + \frac{1}{r} < 1$, on a

$$|cov(X,Y)| \leq 2p(2\alpha)^{\frac{1}{p}} (\mathbb{E}|X|^q)^{\frac{1}{q}} (\mathbb{E}|Y|^r)^{\frac{1}{r}} \tag{A.6}$$

où $\dfrac{1}{p} + \dfrac{1}{q} + \dfrac{1}{r} = 1$.

Dans ce travail, afin d'obtenir des convergences presque sûres (et presque complètes) de nos estimateurs, nous avons également utilisé des inégalités de type exponentiel relatives à des sommes de variables aléatoires. Tout d'abord, pour des variables aléatoires indépendantes, nous avons le théorème suivant :

Théorème A.11. *(Bosq (1998)) Posons $X_1, \ldots, X_n$ des variables aléatoires indépendantes, réelles et de moyenne nulle et $S_n = \sum_{i=1}^n X_i$. Nous avons les inégalités suivantes :*

1. *Si $a_i \leq X_i \leq b_i$, $i = 1 \ldots, n$ où $a_1, b_1, \ldots, a_n, b_n$ sont des constantes, alors*

$$\mathbb{P}(|S_n| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \qquad t > 0 \tag{A.7}$$

(Inégalité de Hoeffding)

2. *S'il existe une constante $c > 0$ telle que (Condition de Cramer)*

$$\mathbb{E}|x_i|^k \leq c^{k-2} k! \mathbb{E}X_i^2 < +\infty \tag{A.8}$$

$i = 1, \ldots, n$ et $k = 3, 4, \ldots$ alors

$$\mathbb{P}(|S_n| \geq t) \leq 2\exp\left(-\frac{t^2}{4\sum_{i=1}^n X_i^2 + 2ct}\right), \qquad t > 0 \tag{A.9}$$

(Inégalité de Bernstein)

Revenons maintenant au cas des variables α-mélangeantes, le théorème suivant concernent les processus stochastiques bornés.

Théorème A.12. *(Bosq (1998)) Soit $(X_t, t \in \mathbb{Z})$ un processus à valeurs réelles, de moyenne nulle tel que $\sup_{1\leq t\leq n} \|X_t\|_\infty \leq b$. Alors pour chaque entier $s \in [1, \frac{n}{2}]$ et chaque $\epsilon > 0$*

$$\mathbb{P}(|S_n| \geq n\epsilon) \leq 4\exp\left(-\frac{\epsilon^2}{8b^2}s\right) + 22\left(1 + \frac{4b}{\epsilon}\right)^{1/2} s\alpha\left(\left\lfloor \frac{n}{2s} \right\rfloor\right) \tag{A.10}$$

A.5 Les fonctions noyaux

Lorsque l'on s'intéresse aux propriétés théoriques d'un estimateur à noyau, on énonce les hypothèses que doivent satisfaire les fonctions de poids (noyaux) $K(\cdot)$ afin d'obtenir ces propriétés. Ainsi les résultats asymptotiques restent valident pour toute une classe de noyaux. Par exemple, la fonction $K(\cdot)$ peut être supposée bornée et à support compact. En pratique, l'utilisateur doit faire un choix plus précis : il doit choisir une fonction particulière $K(\cdot)$ parmi tous les noyaux satisfaisant les hypothèses émises. En général, le choix du noyau n'a pas d'influence majeure (s'il est choisi dans une classe raisonnable d'estimateurs) sur le résultat final contrairement au choix du paramètre de lissage qui est crucial. Cependant, pour chaque application considérée, nous avons testé plusieurs noyaux avant de faire notre choix. Il s'avère que, selon les données disponibles, ce ne sont pas toujours les mêmes fonctions $K(\cdot)$ qui permettent d'obtenir les meilleurs résultats. L'objectif ici est de rassembler la majorité des noyaux rencontrés dans la littérature. Notons qu'une même fonction $K(\cdot)$ peut avoir plusieurs appellations. Nous allons d'abord définir la notion de noyau.

Définition A.13. On appelle noyau une fonction $K : \mathbb{R} \to \mathbb{R}$ telle que $= K(u) \geq 0$, $\int K^2(u)du < +\infty$ et $\int K(u)du = 1$.

Notons que pour deux observations x_i et x_0, $K\left(\frac{x_i - x_0}{h}\right)$ atteint son maximum en 0 lorsque $x_i = x_0$ et décroît avec la distance $|x_0 - x_i|$.

Une variété de fonctions noyaux sont possibles en général, mais des considérations pratiques et théoriques limitent le choix. Par exemple, des fonctions de poids prenant de très petites valeurs peuvent causer des problèmes numériques. Ainsi on se restreint à des fonctions qui valent 0 à l'extérieur d'un certain interval fixé. Une fonction couramment utilisée, qui vérifient des propriétés d'optimalité est de forme parabolique. Dans la suite, nous décrivons des noyaux rencontrés dans la littérature (Deheuvels (1977), Hardle (1990), Tsybakov (2009), etc.)

A.5.1 Exemples de noyaux à support compact :

Le noyau uniforme ou naïf ou rectangulaire ou box(car)

$$K(x) = \frac{1}{2}\mathbf{1}_{[-1;+1]}(x)$$

Le noyau triangulaire ou Bartlett

$$K(x) = (1 - |x|)\mathbf{1}_{[-1;+1]}(x)$$

Le noyau circulaire ou cosinus

$$K(x) = \frac{\pi}{4}\cos\left(\frac{\pi}{2}x\right)\mathbf{1}_{[-1;+1]}(x)$$

Le noyau tricube

$$K(x) = \frac{70}{81}(1 - |x|^3)^3\mathbf{1}_{[-1;+1]}(x)$$

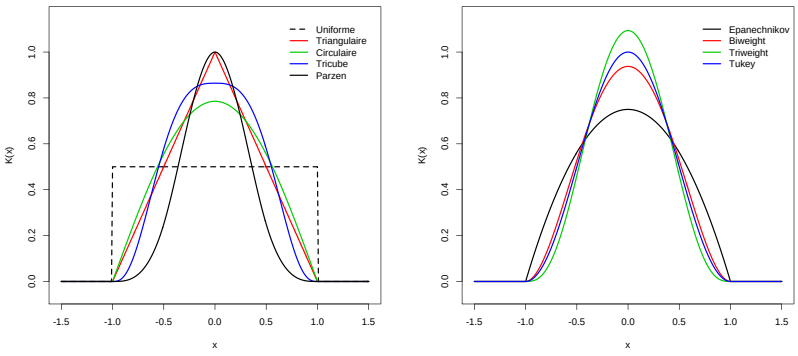


Figure A.1 – Quelques exemples de fonctions noyaux à support compact

Le noyau Parzen

$$K(x) = \begin{cases} 1 - 6|x|^2 + 6|x|^3 & \text{si} \quad |x| \leq \frac{1}{2}, \\ 2(1-|x|)^3, & \text{si} \quad \frac{1}{2} \leq |x| \leq 1 \\ 0 & \text{sinon}, \end{cases}$$

Le noyau parabolique d'Epanechnikov ou quadratique

$$K(x) = \frac{3}{4}\left(1 - x^2\right)\mathbf{1}_{[-1;+1]}(x)$$

Le noyau Biweight de Tukey ou quartique

$$K(x) = \frac{15}{16}(1-x^2)^2\mathbf{1}_{[-1;+1]}(x)$$

Le noyau Triweight ou cubique

$$K(x) = \frac{35}{32}(1-x^2)^3\mathbf{1}_{[-1;+1]}(x)$$

Le noyau de Tukey-Hanning

$$K(x) = \frac{1+\cos(\pi x))}{2}\mathbf{1}_{[-1;+1]}(x)$$

A.5.2 Exemples de noyaux à support non compact :

Le noyau normal ou Gaussien

$$K(x) = \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-x^2}{2}\right)$$

Le noyau de Silverman

$$K(x) = \frac{1}{2}\exp\left(\frac{-|x|}{\sqrt{2}}\right)\sin\left(\frac{|x|}{\sqrt{2}} + \frac{\pi}{4}\right)$$

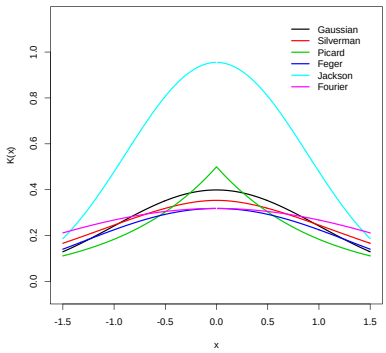


FIGURE A.2 – Quelques exemples de fonctions noyaux à support non-compact

Le noyau de Picard

$$K(x) = \frac{1}{2}\exp\left(-|x|\right)$$

Le noyau de Fejer-de la Vallée Poussin

$$K(x) = \frac{1}{\pi}\left(\frac{1}{x}\sin(x)\right)^2$$

Le noyau de Jackson-de la Vallée Poussin

$$K(x) = \frac{3}{\pi}\left(\frac{1}{x}\sin(x)\right)^4$$

Le noyau de Fourier

$$K(x) = \frac{1}{\pi}\left(\frac{1}{x}\sin(x)\right)$$

A.5.3 Autres

Les noyaux présentés dans la suite donnent des valeurs négatives pour certaines valeurs de x.

Le noyau de Legendre d'ordre 1

$$K(x) = \frac{3}{8}(3 - 5x^2)\mathbf{1}_{[-1;+1]}(x)$$

Le noyau de Legendre d'ordre 2

$$K(x) = \frac{15}{128}(15 - 70x^2 + 63x^4)\mathbf{1}_{[-1;+1]}(x)$$

Le noyau de Gram-Charlier d'ordre 1

$$K(x) = \frac{1}{2}(3 - x^2)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

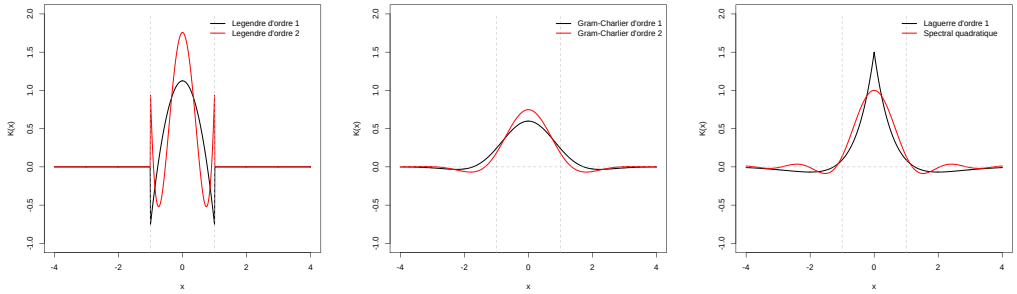


Figure A.3 – Quelques exemples de fonctions noyaux

Le noyau de Gram-Charlier d'ordre 2

$$K(x) = \frac{1}{8}(x^4 - 10x^2 + 15)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

Le noyau de Gram-Charlier d'ordre 2

$$K(x) = \frac{1}{8}(x^4 - 10x^2 + 15)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

Le noyau spectral quadratique

$$K(x) = \frac{25}{12\pi^2 x^2}\left(\frac{\sin(6\pi x/5)}{6\pi x/5} - \cos(6\pi x/5)\right)$$

A.5.4 Noyaux multivariés

Dans les sous-sections précédentes, tous les noyaux définis sont valables lorsque x est univariée ($x \in \mathbb{R}$). Lorsque les données à traiter sont multivariées ($X = (x_1, \ldots, x_d) \in \mathbb{R}^d$), on a

$$K(X) = K(x_1, \ldots, x_d)$$

et il existe plusieurs possibilités pour adapter les noyaux existants au cadre multivarié. Tout d'abord, on peut considérer le noyau multiplicatif, qui s'écrit de la manière suivante

$$K(X) = K(x_1) \times \ldots \times K(x_d).$$

On peut également considérer le noyau radialement symétrique défini par

$$K(X) \propto K(\|X\|), \qquad \text{avec } \|x\| = \sqrt{X'X}$$

Ainsi, par exemple, le noyau multivarié d'Epanechnikov est défini par

$$K(X) = \left(\frac{3}{4}\right)^d (1 - x_1^2)\mathbf{1}_{[|x_1|\leq 1]} \ldots (1 - x_d^2)\mathbf{1}_{[|x_d|\leq 1]} \qquad \textit{(Multiplicatif)}$$

$$K(X) = \frac{(d+2)}{2c_d}(1 - X'X)\mathbf{1}_{[X'X\leq 1]} \qquad \textit{(Radialement symétrique)}$$

où c_d est le volume d'une boule de dimension d (par exemple, $c_1 = 2$, $c_2 = \pi$, $c_3 = 4\pi/3$,...). On peut se référer, par exemple, au chapitre 4 de Silverman (1986), au chapitre 3 de Härdle (2004).

A.6 Les semi-métriques

Les estimateurs proposés tout au long de cette thèse sont composés de fonctions à noyau $K(\cdot)$. Ces dernières donnent un poids plus ou moins important aux données selon leur proximité. Dans le cas où les données sont réelles (comme dans les chapitres 2, 3 et 6), la proximité des données est évaluée à partir de normes classiques comme la norme Euclidienne. Cependant, lorsque nous considérons des données fonctionnelles (comme dans les chapitres 4, 5, 7 et 8) les distances classiques ne sont plus adaptées et le problème du fléau de la dimension apparait. Le fléau de la dimension survient notamment dans le cadre du modèle non paramétrique de régression multivariée, dans lequel les vitesses de convergence ralentissent lorsque la dimension des données augmente. De la même manière nous pouvons penser que les vitesses de convergence deviennent très lentes en présence de variables de dimension infinie.

Dans ce travail, nous avons utilisé des semi-métriques (voir Ferraty et Vieu (2006)) qui permettent de mesurer la proximité entre des variables fonctionnelles. Dans un premier temps, nous définissons la notion de semi-métrique puis nous explicitons quelques familles de semi-métriques rencontrées dans le littérature. En effet, selon les données étudiées la semi-métrique à utiliser varie et influe sur la qualité de l'estimation.

Définition A.14. On dit que d est une semi-métrique sur un espace $\mathcal{E}$ dès que

1. $\forall x \in \mathcal{E},\ d(x,x) = 0$
2. $\forall (x,y,z) \in \mathcal{E} \times \mathcal{E} \times \mathcal{E},\ d(x,y) \leq d(x,z) + d(z,y)$

Notons que $d(x,y) = 0$ n'implique pas que $x = y$.

Considérons un échantillon de n courbes $X_1, \ldots, X_n$ identiquement distribuées selon la variable fonctionnelle $X = \{X(t), t \in T\}$. Le choix de la semi-métrique va dépendre de la nature des données, de la forme des courbes (par exemple, les courbes peuvent être lisses, rugueuses, etc), de considérations pratiques et du problème à traiter. Notons que ce choix est très important puisqu'il joue un rôle important sur la qualité des estimations. Nous présentons quelques familles de semi-métriques, les plus rencontrées dans la littérature, mais d'autres existent. Parmi les familles existantes, l'une est basée sur le calcul de distances entre les dérivées d'ordre d en utilisant une approximation par β-spline. Cette famille est adaptée au traitement de courbes lisses et ne dépend pas du pas de discrétisation. Par conséquent, cette famille convient également lorsque les données ne sont pas enregistrées aux mêmes points de discrétisation des courbes. Il existe également une famille similaire, basée sur le calcul des distances entre les dérivées d'ordre d, mais en utilisant une approximation de Fourier. Des semi-métriques basées sur une analyse en composante principale fonctionnelle ou sur la méthode des moindres carrés partiels fonctionnels peuvent également être utilisées. Mais ces deux dernières familles de semi-métriques peuvent seulement être utilisées lorsque les courbes sont observées aux mêmes points de discrétisation et si le pas de discrétisation est assez fin. La première donne de bons résultats lorsque les courbes sont rugueuses alors que la seconde servira plutôt en présence de variables réponses multivariées. Enfin, une famille calcule les proximités entre les courbes en tenant compte de l'effet de décalage horizontal entre deux objets fonctionnels. Plus de détails sur ces sémi-métriques sont donnés, par exemple, dans Ferraty et Vieu (2006).

Les familles de semi-métriques présentées précédemment sont disponibles dans le package *fda.usc* du logiciel *R*.

A.7 Les probabilités de petites boules

Les probabilités de petites boules ("small ball probabilities" en anglais) servent à décrire le comportement local des données fonctionnelles. Elles sont très importantes pour l'étude théorique des estimateurs présentés aux chapitres 4 et 5, puisqu'elles apparaissent dans les vitesses de convergence. On peut définir la probabilité de petite boule comme la probabilité pour une variable aléatoire fonctionnelle X' à valeur dans un espace semi-métrique $\mathcal{E}$ d'appartenir à la boule $\mathcal{B}(X,h)$ de centre X à valeur dans $\mathcal{E}$ et de rayon h. Pour plus de détails sur cette notion, le lecteur intéressé peut se référer, par exemple, au travaux suivants : Dabo-Niang (2004), Ferraty et Vieu (2006). On peut également l'écrire sous la forme suivante :

$$\varphi_X(h) = \mathbb{P}(X' \in \mathcal{B}(X,h))$$

Lorsque la taille de l'échantillon étudié tend vers l'infini, la taille de la fenêtre h diminue et ainsi cette probabilité tend à être petite, d'où le nom de "probabilités de petites boules". Notons que cet outil dépend de la semi-métrique utilisée et du paramètre de fenêtre.

On peut noter aussi que les probabilités de petites boules utilisées dans la littérature fonctionnelle sont l'équivalent des termes h_n^d rencontrés dans la littérature non paramétrique pour données de dimension finie d.

A.8 Divers

A.8.1 Espérance conditionnelle

Soient X et Y deux variables aléatoires absolument continues de densité conjointe $f_{X,Y}$ et de densités marginales respectives f_X et f_Y, on définit l'espérance de Y conditionnellement à X par

$$\begin{aligned}\mathbb{E}[Y|X=x] &= \int_{\mathbb{R}} y f_{X|Y}(x|y)dy \\ &= \int_{\mathbb{R}} y \frac{f_{X,Y}(x,y)}{f_X(x)}dy \\ &= \frac{\int_{\mathbb{R}} y f_{X,Y}(x,y)dy}{f_X(x)}\end{aligned}$$

Table des figures

Liste des tableaux

Bibliographie

Abraham, C., Cornillon, P. A., Matzner-Løber, E., and Molinari, N. (2003). Unsupervised curve clustering using b-splines. *Scandinavian Journal of Statistics*, 30(3) :581–595.

Allard, D., D'Or, D., and Froidevaux, R. (2011). An efficient maximum entropy approach for categorical variable prediction. *European Journal of Soil Science*, 62(3) :381–393.

Andrews, J. L. and McNicholas, P. D. (2013). Variable selection for clustering and classification. *Journal of Classification*, pages 1–18.

Aneiros-Pérez, G., Cao, R., and Vilar-Fernández, J. M. (2011). Functional methods for time series prediction : a nonparametric approach. *Journal of Forecasting*, 30(4) :377–392.

Anselin, L. and Florax, R. J. G. M. (1995). *New Directions in Spatial Econometrics*. Advances in Spatial Science. Springer.

Antoniadis, A., Brossat, X., Cugliari, J., and Poggi, J.-M. (2014). Une approche fonctionnelle pour la prévision non-paramétrique de la consommation d'électricité. *Journal de la Société Française de Statistique*, 155(2) :202–219.

Aspirot, L., Bertin, K., and Perera, G. (2009). Asymptotic normality of the Nadaraya-Watson estimator for nonstationary functional data and applications to telecommunications. *Journal of Nonparametric Statistics*, 21(5) :535–551.

Assani, A. A. and Tardif, S. (2005). Classification, caractérisation et facteurs de variabilité spatiale des régimes hydrologiques naturels au Québec (Canada), approche écogéographique. *Revue des sciences de l'eau*, 18(2) :247–266.

Aston, J. A., Chiou, J.-M., and Evans, J. P. (2010). Linguistic pitch analysis using functional principal component mixed effect models. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 59(2) :297–317.

Aston, J. A. D. and Kirch, C. (2011). Detecting and estimating epidemic changes in dependent functional data. Working paper, Coventry : University of Warwick. Centre for Research in Statistical Methodology.

Atteia, O., Dubois, J.-P., and Webster, R. (1994). Geostatistical analysis of soil contamination in the Swiss Jura. *Environmental Pollution*, 86(3) :315–327.

Atteia, O., Thélin, P., Pfeifer, H. R., Dubois, J. P., and Hunziker, J. C. (1995). A search for the origin of cadmium in the soil of the Swiss Jura. *Geoderma*, 68(3) :149–172.

Attouch, M., Laksaci, A., and Ould-Said, E. (2009). Asymptotic distribution of robust estimator for functional nonparametric models. *Communications in Statistics - Theory and Methods*, 38(8) :1317–1335.

Attouch, M., Laksaci, A., and Saïd, E. O. (2010). Asymptotic normality of a robust estimator of the regression function for functional time series data. *Journal of the Korean Statistical Society*, 39(4) :489–500.

Attouch, M. K., Gheriballah, A., and Laksaci, A. (2011). Robust nonparametric estimation for functional spatial regression. In Ferraty, F., editor, *Recent Advances in Functional Data Analysis and Related Topics*, Contributions to Statistics, pages 27–31. Physica-Verlag HD.

Auder, B. and Fischer, A. (2012). Projection-based curve clustering. *Journal of Statistical Computation and Simulation*, 82(8) :1145–1168.

Aue, A., Norinho, D. D., and Hormann, S. (2014). On the prediction of stationary functional time series. *Journal of the American Statistical Association*. in press.

Azzedine, N., Laksaci, A., and Ould-Saïd, E. (2008). On robust nonparametric regression estimation for a functional regressor. *Statistics & Probability Letters*, 78(18) :3216–3221.

Baíllo, A. and Grané, A. (2009). Local linear regression for functional predictor and scalar response. *Journal of Multivariate Analysis*, 100(1) :102–111.

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. (2005). Clustering with Bregman divergences. *The Journal of Machine Learning Research*, 6 :1705–1749.

Barrientos-Marin, J., Ferraty, F., and Vieu, P. (2010). Locally modelled regression and functional data. *Journal of Nonparametric Statistics*, 22(5) :617–632.

Basse, M., Dabo-Niang, S., and Diop, A. (2008). Mean square properties of a class of kernel density estimates for spatial functional random variables. *Annales De L'ISUP Publications de l'Institut de Statistique de l'Université de Paris*, pages 91–108.

Bel, L., Allard, D., Laurent, J. M., Cheddadi, R., and Bar-Hen, A. (2009). CART algorithm for spatial data : Application to environmental and ecological data. *Computational Statistics & Data Analysis*, 53(8) :3082–3093.

Belmar, O., Velasco, J., and Martinez-Capel, F. (2011). Hydrological classification of natural flow regimes to support environmental flow assessments in intensively regulated Mediterranean rivers, Segura River Basin (Spain). *Environmental Management*, 47(5) :992–1004.

Ben Alaya, M. A., Chebana, F., and Ouarda, T. B. M. J. (2014). Probabilistic gaussian copula regression model for multisite and multivariable downscaling. *Journal of Climate*, 27(9) :3331–3347.

Berkes, I., Gabrys, R., Horváth, L., and Kokoszka, P. (2009). Detecting changes in the mean of functional observations. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 71(5) :927–946.

Berlinet, A., Elamine, A., and Mas, A. (2011). Local linear regression for functional data. *Annals of the Institute of Statistical Mathematics*, 63(5) :1047–1075.

Biau, G. (2002). Optimal asymptotic quadratic errors of density estimators on random fields. *Statistics & probability letters*, 60(3) :297–307.

Biau, G. (2003). Spatial kernel density estimation. *Mathematical methods of Statistics*, 12(4) :371–390.

Biau, G. and Cadre, B. (2004). Nonparametric spatial prediction. *Statistical Inference for Stochastic Processes*, 7(3) :327–349.

Bosq, D. (1991). Modelization, nonparametric estimation and prediction for continuous time processes. In *Nonparametric functional estimation and related topics*, pages 509–529. Springer.

Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes, Estimation and Prediction*. Springer, second edition.

Bosq, D. (2000). *Linear Processes in Function Spaces : Theory and Applications*, volume 149 of *Lecture Notes in Statistics*. Springer-Verlag New York Inc.

Bosq, D. and Blanke, D. (2008). *Inference and prediction in large dimensions*, volume 754. John Wiley & Sons.

Bower, D., Hannah, D. M., and McGregor, G. R. (2004). Techniques for assessing the climatic sensitivity of river flow regimes. *Hydrological processes*, 18(13) :2515–2543.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3) :200–217.

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to time series and forecasting*, volume 1. Taylor & Francis.

Brockwell, P. J. and Davis, R. A. (2009). *Time series : theory and methods*. Springer.

Cadre, B. and Paris, Q. (2010). On Hölder fields clustering. *TEST*, 21(2) :301–316.

Carbon, M. (2006). Polygone des fréquences pour des champs aléatoires. *Comptes Rendus Mathematique*, 342(9) :693–696.

Carbon, M., Francq, C., and Tran, L. T. (2007). Kernel regression estimation for random fields. *Journal of Statistical Planning and Inference*, 137(3) :778–798.

Carbon, M., Hallin, M., and Tat Tran, L. (1996). Kernel density estimation for random fields : the L_1 theory. *Journal of Nonparametric Statistics*, 6(2-3) :157–170.

Carbon, M., Tran, L. T., and Wu, B. (1997). Kernel density estimation for random fields (density estimation for random fields). *Statistics & Probability Letters*, 36(2) :115–125.

Chamizo, L. F. and Iwaniec, H. (1995). On the sphere problem. *Revista Matemática Iberoamericana*, 11(2) :417–430.

Chebana, F., Dabo-Niang, S., and Ouarda, T. B. M. J. (2012). Exploratory functional flood frequency analysis and outlier detection. *Water Resources Research*, 48(4) :W04514.

Chebana, F. and Ouarda, T. B. M. J. (2011). Multivariate quantiles in hydrological frequency analysis. *Environmetrics*, 22(1) :63–78.

Chilès, J.-P. and Delfiner, P. (1999). *Geostatistics : Modeling Spatial Uncertainty*. Wiley Series in Applied Probability and Statistics. John Wiley & Sons, Inc.

Cook, E. R., Anchukaitis, K. J., Buckley, B. M., D'Arrigo, R. D., Jacoby, G. C., and Wright, W. E. (2010). Asian monsoon failure and megadrought during the last millennium. *Science*, 328(5977) :486–489.

Cox, D. D. and Kim, T. Y. (1995). Moment bounds for mixing random variables useful in nonparametric function estimation. *Stochastic Processes and their Applications*, 56(1) :151–158.

Crambes, C., Delsol, L., and Laksaci, A. (2008). Robust nonparametric estimation for functional data. *Journal of Nonparametric Statistics*, 20(7) :573–598.

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley Series in Probability and Statistics. John Wiley & Sons.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*, volume 110 of *Wiley Series in Probability and Statistics*. Wiley-Interscience, revised edition.

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147 :1–23.

Cullen, H. M., Kaplan, A., Arkin, P. A., and deMenocal P B (2002). Impact of the North Atlantic Oscillation on Middle Eastern climate and streamflow. *Climatic Change*, 55(3) :315–338.

Dabo-Niang, S. (2002). Estimation de la densité dans un espace de dimension infinie : Application aux diffusions. *Comptes Rendus Mathématique*, 334(3) :213–216.

Dabo-Niang, S. (2004). Kernel density estimator in an infinite-dimensional space with a rate of convergence in the case of diffusion process. *Applied Mathematics Letters*, 17(4) :381–386.

Dabo-Niang, S., Ferraty, F., and Vieu, P. (2004). Nonparametric unsupervised classification of satellite wave altimeter forms. *Proceedings in Computational Statistics, Ed. J. Antoch, Physica-Verlag, Heidelberg New-York*, pages 879–886.

Dabo-Niang, S., Ferraty, F., and Vieu, P. (2006). Mode estimation for functional random variable and its application for curves classification. *Far East Journal of Theoretical Statistics*, 18(1) :93–119.

Dabo-Niang, S., Ferraty, F., and Vieu, P. (2007). On the using of modal curves for radar waveforms classification. *Computational Statistics & Data Analysis*, 51(10) :4878–4890.

Dabo-Niang, S., Hamdad, L., Ternynck, C., and Yao, A.-F. (2014a). A kernel spatial density estimation allowing for the analysis of spatial clustering : application to Monsoon Asia Drought Atlas data. *Stochastic Environmental Research and Risk Assessment*, Accepted, Online :1–25.

Dabo-Niang, S., Kaid, Z., and Laksaci, A. (2011a). Sur la régression quantile pour variable explicative fonctionnelle : Cas des données spatiales. *Comptes Rendus Mathematique*, 349(23) :1287–1291.

Dabo-Niang, S., Kaid, Z., and Laksaci, A. (2012a). On spatial conditional mode estimation for a functional regressor. *Statistics & Probability Letters*, 82(7) :1413–1421.

Dabo-Niang, S., Kaid, Z., and Laksaci, A. (2012b). Spatial conditional quantile regression : Weak consistency of a kernel estimate. *Romanian Journal of Pure and Applied Mathematics*, 57(4) :311–339.

Dabo-Niang, S. and Laksaci, A. (2007). Estimation non paramétrique du mode conditionnel pour variable explicative fonctionnelle. *Comptes Rendus Mathematique*, 344(1) :49–52.

Dabo-Niang, S. and Laksaci, A. (2012). Nonparametric quantile regression estimation for functional dependent data. *Communications in Statistics-Theory and Methods*, 41(7) :1254–1268.

Dabo-Niang, S., Ould-Abdi, S. A., Ould-Abdi, A., and Diop, A. (2014b). Consistency of a nonparametric conditional mode estimator for random fields. *Statistical Methods & Applications*, 23(1) :1–39.

Dabo-Niang, S., Rachdi, M., and Yao, A.-F. (2011b). Kernel regression estimation for spatial functional random variables. *Far East Journal of Theoretical Statistics*, 37(2) :77–113.

Dabo-Niang, S. and Rhomari, N. (2003). Estimation non paramétrique de la régression avec variable explicative dans un espace métrique. *Comptes Rendus Mathematique*, 336(1) :75–80.

Dabo-Niang, S. and Rhomari, N. (2009). Kernel regression estimation in a Banach space. *Journal of Statistical Planning and Inference*, 139(4) :1421–1434.

Dabo-Niang, S., Ternynck, C., and Yao, A.-F. (2014c). Spatial regression for multivariate data taking spatial characteristics into consideration and applications. *Preprint*.

Dabo-Niang, S. and Thiam, B. (2010). Robust quantile estimation and prediction for spatial processes. *Statistics & Probability Letters*, 80(17) :1447–1458.

Dabo-Niang, S. and Yao, A.-F. (2007). Kernel regression estimation for continuous spatial processes. *Mathematical Methods of Statistics*, 16(4) :298–317.

Dabo-Niang, S. and Yao, A.-F. (2013). Kernel spatial density estimation in infinite dimension space. *Metrika*, 76(1) :19–52.

Dabo-Niang, S., Yao, A.-F., Pischedda, L., Cuny, P., and Gilbert, F. (2010). Spatial mode estimation for functional random fields with application to bioturbation problem. *Stochastic Environmental Research and Risk Assessment*, 24(4) :487–497.

Davies, T. M., Hazelton, M. L., and Marshall, J. C. (2011). Sparr : analyzing spatial relative risk using fixed and adaptive kernel density estimation in R. *Journal of Statistical Software*, 39(1) :1–14.

Deheuvels, P. (1977). Estimation non paramétrique de la densité par histogrammes généralisés. *Revue de Statistique Appliquée*, 25(3) :5–42.

Deistler, M. and Hannan, E. J. (1988). *The Statistical Theory of Linear Systems*. Wiley, New York.

Delaigle, A. and Gijbels, I. (2004). Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. *Annals of the Institute of Statistical Mathematics*, 56(1) :19–47.

Delaigle, A., Hall, P., and Bathia, N. (2012). Componentwise classification and clustering of functional data. *Biometrika*, page ass003.

Delicado, P., Giraldo, R., Comas, C., and Mateu, J. (2010). Statistics for spatial functional data : some recent contributions. *Environmetrics*, 21(3-4) :224–239.

Delsol, L. (2007a). CLT and L_q errors in nonparametric functional regression. *Comptes Rendus Mathematique*, 345(7) :411–414.

Delsol, L. (2007b). Régression non-paramétrique fonctionnelle : Expressions asymptotiques des moments. *Annales de l'ISUP*, 51(3) :43–67.

Delsol, L. (2008). *Régression sur variable fonctionnelle : Estimation, tests de structure et Applications.* PhD thesis, Université Paul Sabatier-Toulouse III.

Delsol, L. (2009). Advances on asymptotic normality in non-parametric functional time series analysis. *Statistics*, 43(1) :13–33.

Demongeot, J., Laksaci, A., Madani, F., and Rachdi, M. (2013). Functional data : local linear estimation of the conditional density and its application. *Statistics*, 47(1) :26–44.

Easterling, D. R. and Peterson, T. C. (1995). A new method for detecting undocumented discontinuities in climatological time series. *International journal of climatology*, 15(4) :369–377.

El Machkouri, M. (2007). Nonparametric regression estimation for random fields in a fixed-design. *Statistical Inference for Stochastic Processes*, 10(1) :29–47.

El Machkouri, M. (2011). Asymptotic normality of the Parzen–Rosenblatt density estimator for strongly mixing random fields. *Statistical Inference for Stochastic Processes*, 14(1) :73–84.

El Machkouri, M. and Stoica, R. (2010). Asymptotic normality of kernel estimates in a regression model for random fields. *Journal of Nonparametric Statistics*, 22(8) :955–971.

Ettinger, B., Guillas, S., and Lai, M.-J. (2012). Bivariate splines for ozone concentration forecasting. *Environmetrics*, 23(4) :317–328.

Ezzahrioui, M. and Ould-Saïd, E. (2008). Asymptotic results of a nonparametric conditional quantile estimator for functional time series. *Communications in Statistics ?Theory and Methods*, 37(17) :2735–2759.

Fan, J. and Yao, Q. (2003). *Nonlinear time series*, volume 2. Springer.

Fazekas, I. and Chuprunov, A. (2006). Asymptotic normality of kernel type density estimators for random fields. *Statistical inference for stochastic processes*, 9(2) :161–178.

Febrero, M., Galeano, P., and González-Manteiga, W. (2007). A functional analysis of NOx levels : location and scale estimation and outlier detection. *Computational Statistics*, 22(3) :411–427.

Febrero, M., Galeano, P., and González-Manteiga, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. *Environmetrics*, 19(4) :331–345.

Ferraty, F., Goia, A., and Vieu, P. (2002a). Functional nonparametric model for time series : a fractal approach for dimension reduction. *Test*, 11(2) :317–344.

Ferraty, F., Goia, A., and Vieu, P. (2002b). Régression non-paramétrique pour des variables aléatoires fonctionnelles mélangeantes. *C. R. Math. Acad. Sci. Paris*, 334(3) :217–220.

Ferraty, F., Laksaci, A., Tadj, A., and Vieu, P. (2010). Rate of uniform consistency for nonparametric estimates with functional variables. *Journal of Statistical Planning and Inference*, 140(2) :335–352.

Ferraty, F., Laksaci, A., Tadj, A., and Vieu, P. (2012a). Estimation de la fonction de régression pour variable explicative et réponse fonctionnelles dépendantes. *Comptes Rendus Mathematique*, 350(13) :717–720.

Ferraty, F., Laksaci, A., Tadj, A., Vieu, P., et al. (2011). Kernel regression with functional response. *Electronic Journal of Statistics*, 5 :159–171.

Ferraty, F., Laksaci, A., and Vieu, P. (2005a). Functional time series prediction via conditional mode estimation. *Comptes Rendus Mathematique*, 340(5) :389–392.

Ferraty, F., Laksaci, A., and Vieu, P. (2006). Estimating some characteristics of the conditional distribution in nonparametric functional models. *Statistical Inference for Stochastic Processes*, 9(1) :47–76.

Ferraty, F., Mas, A., and Vieu, P. (2007). Nonparametric regression on functional data : inference and practical aspects. *Australian & New Zealand Journal of Statistics*, 49(3) :267–286.

Ferraty, F., Rabhi, A., and Vieu, P. (2005b). Conditional quantiles for dependent functional data with application to the climatic" el niño" phenomenon. *Sankhyā : The Indian Journal of Statistics*, pages 378–398.

Ferraty, F., Van Keilegom, I., and Vieu, P. (2012b). Regression when both response and predictor are functions. *Journal of Multivariate Analysis*, 109 :10–28.

Ferraty, F. and Vieu, P. (2000). Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés. *Compte Rendus de l'Académie des Sciences, Series I, Mathematics*, 330 :403–406.

Ferraty, F. and Vieu, P. (2002). The functional nonparametric model and application to spectrometric data. *Computational Statistics*, 17(4) :545–564.

Ferraty, F. and Vieu, P. (2004). Nonparametric models for functional data, with application in regression, time series prediction and curve discrimination. *Nonparametric Statistics*, 16(1-2) :111–125.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis : theory and practice*. Springer.

Fischer, A. (2010). Quantization and clustering with Bregman divergences. *Journal of Multivariate Analysis*, 101(9) :2207–2221.

Fraiman, R., Justel, A., and Svarc, M. (2008). Selection of variables for cluster analysis and classification rules. *Journal of the American Statistical Association*, 103(483).

Gabrys, R., Horváth, L., and Kokoszka, P. (2010). Tests for error correlation in the functional linear model. *Journal of the American Statistical Association*, 105(491) :1113–1125.

Gabrys, R. and Kokoszka, P. (2007). Portmanteau test of independence for functional observations. *Journal of the American Statistical Association*, 102(480) :1338–1348.

Gaetan, C. and Guyon, X. (2008). *Modélisation et statistique spatiales*, volume 63. Springer.

Ganora, D., Claps, P., Laio, F., and Viglione, A. (2009). An approach to estimate nonparametric flow duration curves in ungauged basins. *Water Resources Research*, 45(10).

Gao, J., Lu, Z., and Tjøstheim, D. (2008). Moment inequalities for spatial processes. *Statistics & Probability Letters*, 78(6) :687–697.

García-Soidán, P. and Menezes, R. (2012). Estimation of the spatial distribution through the kernel indicator variogram. *Environmetrics*, 23(6) :535–548.

Gasser, T., Hall, P., and Presnell, B. (1998). Nonparametric estimation of the mode of a distribution of random curves. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 60(4) :681–691.

Gheriballah, A., Laksaci, A., and Rouane, R. (2010). Robust nonparametric estimation for spatial regression. *Journal of Statistical Planning and Inference*, 140(7) :1656 – 1670.

Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*. Oxford University Press on Demand.

Goovaerts, P. (1998). Ordinary cokriging revisited. *Mathematical Geology*, 30(1) :21–42.

Goudie, A. S. (2006). Global warming and fluvial geomorphology. *Geomorphology*, 79(3) :384–394.

Guillas, S. and Lai, M.-J. (2010). Bivariate splines for spatial functional regression models. *Journal of Nonparametric Statistics*, 22(4) :477–497.

Guyon, X. (1995). *Random Fields on a Network : Modeling, Statistics, and Applications*. Probability and its Applications. Springer-Verlag.

Hall, P. (1982). Cross-validation in density estimation. *Biometrika*, 69(2) :383–390.

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 68(1) :109–126.

Hall, P., Müller, H.-G., and Wu, P.-S. (2006). Real-time density and mode estimation with application to time-dynamic mode tracking. *Journal of Computational and Graphical Statistics*, 15(1).

Hallin, M., Lu, Z., and Tran, L. T. (2001). Density estimation for spatial linear processes. *Bernoulli*, pages 657–668.

Hallin, M., Lu, Z., and Tran, L. T. (2004a). Kernel density estimation for spatial processes : the L_1 theory. *Journal of Multivariate Analysis*, 88(1) :61–75.

Hallin, M., Lu, Z., and Tran, L. T. (2004b). Local linear spatial regression. *The Annals of Statistics*, 32(6) :2469–2500.

Hallin, M., Lu, Z., and Yu, K. (2009). Local linear spatial quantile regression. *Bernoulli*, 15(3) :659–686.

Hannah, D. M., Smith, B. P. G., Gurnell, A. M., and McGregor, G. R. (2000). An approach to hydrograph classification. *Hydrological Processes*, 14(2) :317–338.

Hardle, W. (1990). *Applied nonparametric regression*, volume 27. Cambridge Univ Press.

Härdle, W. (2004). *Nonparametric and semiparametric models*. Springer.

Harris, N. M., Gurnell, A. M., Hannah, D. M., Petts, G. E., et al. (2000). Classification of river regimes : a context for hydroecology. *Hydrological Processes*, 14(16-17) :2831–2848.

Hartigan, J. A. (1975). *Cluster algorithms*. New York : John Wiley & Sons, Inc.

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics : The *np* package. *Journal of statistical software*, 27(5) :1–32.

Hörmann, S. and Kokoszka, P. (2010). Weakly dependent functional data. *The Annals of Statistics*, 38(3) :1845–1884.

Horváth, L., Hušková, M., and Kokoszka, P. (2010). Testing the stability of the functional autoregressive process. *Journal of Multivariate Analysis*, 101(2) :352–367.

Horváth, L., Hušková, M., and Rice, G. (2013). Test of independence for functional data. *Journal of Multivariate Analysis*, 117 :100–119.

Horváth, L. and Kokoszka, P. (2012). *Inference for functional data with applications*, volume 200. Springer.

Hurrell, J. W. and Van Loon, H. (1997). Decadal variations in climate associated with the North Atlantic oscillation. *Climatic Change*, 36(3-4) :301–326.

Hyndman, R. J. and Shang, H. L. (2010). Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*, 19(1).

Jacques, J. and Preda, C. (2014). Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, 71 :92–106.

James, G. M. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98(462) :397–408.

Jiang, D., Eick, C. F., and Chen, C.-S. (2007). On supervised density estimation techniques and their application to spatial data mining. In *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems*, GIS '07, pages 65 :1–65 :4, New York, NY, USA. ACM.

Journel, A. G. (1983). Nonparametric estimation of spatial distributions. *Journal of the International Association for Mathematical Geology*, 15(3) :445–468.

Karácsony, Z. and Filzmoser, P. (2010). Asymptotic normality of kernel type regression estimators for random fields. *Journal of Statistical Planning and Inference*, 140(4) :872–886.

Kaufman, L. and Rousseeuw, P. (1990). *Finding groups in data : An introduction to cluster analysis*. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics, New York.

Kelejian, H. H. and Prucha, I. R. (2007). HAC estimation in a spatial framework. *Journal of Econometrics*, 140(1) :131–154.

Khan, S. S. and Ahmad, A. (2004). Cluster center initialization algorithm for K-means clustering. *Pattern Recognition Letters*, 25(11) :1293–1302.

Kingston, D. G., Thompson, J. R., and Kite, G. (2011). Uncertainty in climate change projections of discharge for the Mekong River Basin. *Hydrology and Earth System Sciences*, 15(5) :1459–1471.

Klemelä, J. (2008). Density estimation with locally identically distributed data and with locally stationary data. *Journal of Time Series Analysis*, 29(1) :125–141.

Krzanowski, W. J. and Lai, Y. T. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, 44 :23–34.

Kundzewicz, Z. W. and Robson, A. J. (2004). Change detection in hydrological records : a review of the methodology. Revue méthodologique de la détection de changements dans les chroniques hydrologiques. *Hydrological Sciences Journal*, 49(1).

Laksaci, A., Lemdani, M., and Ould-Saïd, E. (2009). A generalized L_1-approach for a kernel estimator of conditional quantile with functional regressors : consistency and asymptotic normality. *Statistics & Probability Letters*, 79(8) :1065–1073.

Laksaci, A. and Maref, F. (2009). Estimation non paramétrique de quantiles conditionnels pour des variables fonctionnelles spatialement dépendantes. *Comptes Rendus Mathématique*, 347(17-18) :1075–1080.

Laksaci, A. and Mechab, B. (2010). Estimation non paramétrique de la fonction de hasard avec variable explicative fonctionnelle : cas des données spatiales. *Revue Roumaine de Mathématiques Pures et Appliquées*, 55(1) :35–51.

Li, J. and Tran, L. T. (2009). Nonparametric estimation of conditional expectation. *Journal of Statistical Planning and Inference*, 139(2) :164 – 175.

Li, Q. and Racine, J. (2003). Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis*, 86(2) :266–292.

Lin, X. and Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American statistical Association*, 95(450) :520–534.

Lu, Z. and Chen, X. (2002). Spatial nonparametric regression estimation : non-isotropic case. *Acta Mathematicae Applicatae Sinica*, 18(4) :641–656.

Lu, Z. and Chen, X. (2004). Spatial kernel regression estimation : weak consistency. *Statistics & Probability Letters*, 68(2) :125–136.

Masry, E. (2005). Nonparametric regression estimation for dependent functional data : asymptotic normality. *Stochastic Processes and their Applications*, 115(1) :155–177.

Matheron, G. (1962). *Traité de géostatistique appliquée. 1 (1962)*, volume 1. Editions Technip.

Menezes, R., Garcia-Soidán, P., and Febrero-Bande, M. (2008). A kernel variogram estimator for clustered data. *Scandinavian Journal of Statistics*, 35(1) :18–37.

Menezes, R., García-Soidán, P., and Ferreira, C. (2010). Nonparametric spatial prediction under stochastic sampling design. *Journal of Nonparametric Statistics*, 22(3) :363–377.

Meyer, A. (2011). On the number of lattice points in a small sphere. In *WCC 2011 - Workshop on coding and cryptography*, pages 463–472.

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2) :159–179.

Mitchell, W. C. (1966). The number of lattice points in a k-dimensional hypersphere. *Mathematics of Computation*, 20(94) :300–310.

Modarres, R. and Ouarda, T. B. M. J. (2013). Testing and modelling the volatility change in ENSO. *Atmosphere-Ocean*, 51(5) :561–570.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1) :141–142.

Nazemosadat, M. J., Samani, N., Barry, D. A., and Molaii Niko, M. (2006). ENSO forcing on climate change in Iran : Precipitation analysis. *Iranian Journal of Science and Technology*, 30(B4) :555–565.

Neaderhouser, C. C. (1980). Convergence of block spins defined by a random field. *J. Statist. Phys.*, 22(6) :673–684.

Nerini, D., Monestiez, P., and Manté, C. (2010). Cokriging for spatial functional data. *Journal of Multivariate Analysis*, 101(2) :409–418.

Ogden, R. T. and Greene, E. (2010). Wavelet modeling of functional random effects with application to human vision data. *Journal of Statistical Planning and Inference*, 140(12) :3797–3808.

Ouachani, R., Bargaoui, Z., and Ouarda, T. B. M. J. (2013). Power of teleconnection patterns on precipitation and streamflow variability of upper Medjerda Basin. *International Journal of Climatology*, 33 :58–76.

Ouarda, T. B. M. J., Rasmussen, P. F., Cantin, J. F., Bobée, B., Laurence, R., and Hoang, V. D. (1999). Identification d'un réseau hydrométrique pour le suivi des modifications climatiques dans la province de Québec. *Revue des Sciences de l'Eau*, 12(2) :425–448.

Ould-Abdi, A., Diop, A., Dabo-Niang, S., and Ould-Abdi, S. A. (2010a). Estimation non paramétrique du mode conditionnel dans le cas spatial. *Comptes Rendus Mathematique*, 348(13) :815–819.

Ould-Abdi, S., Dabo-Niang, S., Diop, A., and Ould Abdi, A. (2010b). Consistency of a nonparametric conditional quantile estimator for random fields. *Mathematical Methods of Statistics*, 19(1) :1–21.

Ould-Abdi, S. A., Dabo-Niang, S., Diop, A., and Ould-Abdi, A. (2011). Asymptotic normality of a nonparametric conditional quantile estimator for random fields. *Advances in Decision Sciences*.

Pappenberger, F. and Beven, K. J. (2004). Functional classification and evaluation of hydrographs based on Multicomponent Mapping (mx). *International Journal of River Basin Management*, 2(2) :89–100.

Park, A., Guillas, S., and Petropavlovskikh, I. (2013). Trends in stratospheric ozone profiles using functional mixed models. *Atmospheric Chemistry and Physics*, 13(22) :11473–11501.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3) :1065–1076.

Racine, J. and Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, 119(1) :99–130.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional data analysis*. Springer-Verlag, New York.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis : methods and case studies*, volume 77. Springer.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*. Springer-Verlag, New York, 2nd edition.

Richter, B. D., Baumgartner, J. V., Powell, J., and Braun, D. P. (1996). A method for assessing hydrologic alteration within ecosystems. *Conservation Biology*, 10(4) :1163–1174.

Ripley, B. D. (1981). *Spatial Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc.

Robinson, P. M. (2011). Asymptotic theory for nonparametric regression with spatial data. *Journal of Econometrics*, 165(1) :5–19.

Rogers, J. C. (1997). North Atlantic storm track variability and its association to the North Atlantic oscillation and climate variability of Northern Europe. *Journal of Climate*, 10(7) :1635–1647.

Rosenblatt, M. (1956a). A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences of the United States of America*, 42(1) :43.

Rosenblatt, M. (1956b). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3) :832–837.

Rosenblatt, M. (1985). *Stationary sequences and random fields*. Birkhauser, Boston.

Ruckstuhl, A., Welsh, A., and Carroll, R. J. (2000). Nonparametric function estimation of the relationship between two repeatedly measured variables. *Statist. Sinica*, 10(1) :51–71.

Sangalli, L. M., Ramsay, J. O., and Ramsay, T. O. (2013). Spatial spline regression models. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*.

Seidou, O. and Ouarda, T. B. M. J. (2007). Recursion-based multiple change point detection in multiple linear regression and application to river streamflows. *Water Resources Research*, 43(7).

Shang, H. L. (2014). A survey of functional principal component analysis. *AStA Advances in Statistical Analysis*, 98(2) :121–142.

Shao, X. and Zhang, X. (2010). Testing for change points in time series. *Journal of the American Statistical Association*, 105(491) :1228–1240.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Solow, A. R. (1987). Testing for climate change : An application of the two-phase regression model. *Journal of Climate and Applied Meteorology*, 26(10) :1401–1405.

Stein, M. L. (1999). *Interpolation of Spatial Data : Some Theory for Kriging*. Springer Series in Statistics. Springer-Verlag New York, Inc.

Takahata, H. (1983). On the rates in the central limit theorem for weakly dependent random fields. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 64(4) :445–456.

Ternynck, C. (2014). Spatial regression estimation for functional data with spatial dependency. *Journal de la Société Française de Statistique*, 155(2) :138–160.

Ternynck, C., Ben Alaya, M. A., Chebana, F., Dabo-Niang, S., and Ouarda, T. B. M. J. (2014). Streamflow hydrograph classification using functional data analysis. Submitted, in revision.

Tjøstheim, D. (1987). Spatial series and time series : similarities and differences. In Droesbeke, F., editor, *Spatial Processes and Spatial Time Series Analysis*, Proceedings of the 6th Franco-Belgian Meeting of Statisticians, pages 217–228. Brussels : Publications des Facultés Universitaires Saint- Louis.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, pages 234–240.

Tran, L. T. (1990). Kernel density estimation on random fields. *Journal of Multivariate Analysis*, 34(1) :37–53.

Tran, L. T. and Yakowitz, S. (1993). Nearest neighbor estimators for random fields. *Journal of Multivariate Analysis*, 44(1) :23–46.

Tsang, K.-M. (2000). Counting lattice points in the sphere. *Bulletin of the London Mathematical Society*, 32(6) :679–688.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer.

Vincent, L. A. (1998). A technique for the identification of inhomogeneities in Canadian temperature series. *Journal of Climate*, 11(5) :1094–1104.

Wackernagel, H. (2003). *Multivariate Geostatistics : An Introduction with Applications*. Springer-Verlag Berlin Heidelberg, third edition.

Wang, H. and Wang, J. (2009). Estimation of the trend function for spatio-temporal models. *Journal of Nonparametric Statistics*, 21(5) :567–588.

Wang, H., Wang, J., and Huang, B. (2012). Prediction for spatio-temporal models with autoregression in errors. *Journal of Nonparametric Statistics*, 24(1) :217–244.

Wang, Y. and Woodroofe, M. (2014). On the asymptotic normality of kernel density estimators for causal linear random fields. *Journal of Multivariate Analysis*, 123 :201–213.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā : The Indian Journal of Statistics, Series A*, pages 359–372.

Webster, R., Atteia, O., and Dubois, J.-P. (1994). Coregionalization of trace metals in the soil in the Swiss Jura. *European Journal of Soil Science*, 45(2) :205–218.

Weijs, S. V., van de Giesen, N., and Parlange, M. B. (2013). Data compression to define information content of hydrological time series. *Hydrology and Earth System Sciences*, 17(8) :3171–3187.

Wolfowitz, J. (1942). Additive partition functions and a class of statistical hypotheses. *The Annals of Mathematical Statistics*, 13(3) :247–279.

Wong, H., Hu, B., Ip, W., and Xia, J. (2006). Change-point analysis of hydrological time series using grey relational method. *Journal of Hydrology*, 324(1) :323–338.

Xiao, Z., Linton, O. B., Carroll, R. J., and Mammen, E. (2003). More efficient local polynomial estimation in nonparametric regression with autocorrelated errors. *Journal of the American Statistical Association*, 98(464) :980–992.

Yue, S., Ouarda, T. B. M. J., Bobée, B., Legendre, P., and Bruneau, P. (2002). Approach for describing statistical properties of flood hydrograph. *Journal of Hydrologic Engineering*, 7(2) :147–153.

Zhang, X., Shao, X., Hayhoe, K., and Wuebbles, D. J. (2011). Testing the structural stability of temporally dependent functional observations and application to climate projections. *Electronic Journal of Statistics*, 5 :1765–1796.