Contributions à la prévision statistique

Olivier P. Faugeras

THÈSE DE DOCTORAT

DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Ecole doctorale de Sciences mathématiques de Paris-Centre (ED 386)

Spécialité : Statistique Mathématique

Présentée par: M. Olivier Paul FAUGERAS

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse : Contributions à la prévision statistique

Soutenue le 28 Novembre 2008, devant le jury composé de :

- M. Marc HALLIN, Président du jury
- M. Denis BOSQ, Directeur de thèse
- M. Christian GENEST, Rapporteur
- Mme. Hannelore LIERO, Rapporteur
- M. Patrice BERTAIL, Examinateur
- M. Michel BRONIATOWSKI, Examinateur

Remerciements

Au moment de mettre un point final à cette thèse, je tiens à remercier toutes les personnes qui m'ont aidé à mener à bien ce travail, et en particulier les membres de mon jury. En premier lieu, je voudrais remercier chaleureusement M. Denis Bosq, Professeur Emérite à l'Université Pierre et Marie Curie, d'avoir été mon directeur de thèse. Merci à M. Marc Hallin, Professeur à l'Université Libre de Bruxelles, qui me fait l'honneur de présider le jury de cette thèse. Merci à M. Christian Genest, Professeur à l'Université Laval à Québec, et à Mme Hannelore Liero, Professeur à l'université de Postdam, d'avoir accepté de juger ce travail et d'en avoir été les rapporteurs. Leurs remarques et suggestions ont étés précieuses pour l'amélioration de la qualité de ce manuscrit. Je suis par ailleurs extrêmement reconnaissant à M. Christian Genest de l'accueil qu'il m'a réservé au congrès de la SSC-SFdS à Ottawa et des échanges que nous avons pu avoir. Merci à MM. les Professeurs Patrice Bertail, de l'Université Paris X, et Michel Broniatowski, de l'Université Pierre et Marie Curie, de me faire aussi l'honneur d'avoir accepté de participer à ce jury en tant qu'examinateurs.

Je veux aussi remercier M. Paul Deheuvels, Professeur à l'Université Paris 6, de m'avoir accueilli dans son DEA, puis dans son laboratoire. Je voudrais remercier à ce titre tous les membres du LSTA, ainsi que les personnes que j'ai eu la chance de rencontrer dans les laboratoires Lim& Bio de Paris 13 et Modal'X de Paris 10, Nanterre. Je tiens ainsi à remercier plus particulièrement MM. Emmanuel Guerre, de l'Université Queen Mary de Londres, Gérard Biau et Philippe Saint-Pierre, de l'Université Pierre et Marie Curie

pour leurs encouragements et leurs conseils décisifs à l'accomplissement de ce travail, MM. Francesco Russo et Alain Venot de l'Université Paris 13 pour m'avoir accueilli à Paris 13 et donné l'occasion de faire connaître mon travail, MM. Armelle Guillou de l'Université de Strasbourg, Karine Tribouley de l'Université Paris 10, Taoufik Bouezmarni de HEC Montréal, Alexandre Leblanc de l'Université du Manitoba et Johan Segers de l'Université Catholique de Louvain pour l'intérêt qu'ils ont manifesté dans ma recherche, et bien d'autres qui ont grandement contribué à la réussite de cette thèse.

Merci aux jeunes docteurs, passés ou à venir, notamment François-Xavier Lejeune, Lahcen Douge, Kaouthar El Fassi, Pierre Ribereau, Esterina Masiello, Clara Zelli, Gwladys Toulemonde, Olivier Bouaziz, Claire Coiffard, Rawane Samb entre autres, pour m'avoir soutenu au cours de ces années. Merci bien sûr à Anne Durrande, Louise Lamart et Pascal Epron pour leur dévouement au labo et leur gentillesse. Merci enfin à toutes celles et ceux qui m'ont accompagné pendant cette aventure et dont la place me manque ici pour les énumérer tous.

A mon père, Paul-Etienne Faugeras,

Docteur en Physique, ancien chercheur au CERN,

1938-2008.

Contents

Introduction et Résumé

Les Sciences naturelles et physiques s'attachent traditionnellement à la compréhension du phénomène sous-jacent aux observations, par la création d'un *modèle*, dont la validité et la pertinence peut être remise en cause au cours d'une *expérience*. C'est par ce travail dialectique entre théorisation et expérimentation, que la connaissance scientifique progresse. Pour citer H. Poincaré dans *La Science et l'hypothèse*, mentionnons que si la science se construit à partir de faits tirés de l'expérience, "une accumulation de faits n'est pas plus une science qu'un tas de pierres n'est une maison." Aussi, dans l'élaboration de sa théorie, le scientifique doit soumettre constamment ses hypothèses à la vérification par l'expérience. Et pour accomplir ce travail de validation, "avant tout, le savant doit prévoir."

C'est dire, outre son intérêt éminemment pratique, l'importance du problème de la prévision, notamment pour fonder l'utilisation de la Statistique à des fins scientifiques. Bien évidemment, le propos de cette thèse est bien plus modeste en regard de ces enjeux épistémologiques. Le présent travail traite de la Prévision Statistique. Il aborde ce problème sous deux points de vue, paramétrique et non paramétrique, qui correspondent aux deux parties de cette thèse.

En effet, la plupart des phénomènes physiques dans la Nature ont un élément aléatoire dans leur structure, qui fait que les grandeurs sont variables et ne peuvent être prévues avec certitude. Il est alors naturel d'adopter une approche Statistique. Un modèle probabiliste est alors supposé décrire le comportement du phénomène, qui évolue selon une loi

de probabilité.

- Dans la première partie de ce mémoire, le modèle probabiliste est un *processus stochastique*, et l'on observe une *série temporelle*, c'est-à-dire une collection d'observations effectuées séquentiellement dans le temps, par exemple $x_1, \ldots, x_T$. Des exemples de séries temporelles abondent dans de nombreux domaines, notamment en économie, ingénierie: elles peuvent correspondre à des températures moyennes journalières, à des cours de Bourse, etc... On cherche alors à prédire une valeur *future* x_{T+h} connaissant le *passé* $x_1, \ldots, x_T$. La difficulté du sujet a donné lieu à de multiples façons d'aborder ce problème, voir par exemple Chatfield [34]. On suppose ici que l'on est dans un *cadre paramétrique*: les observations (x_t) sont la réalisation d'un processus stochastique (X_t) dont la loi est indexée par un paramètre θ inconnu. Une approche de type "plug-in" est alors développée dans les chapitres 1 et 2.

- Dans la seconde partie de ce mémoire, l'approche abordée est quelque peu différente. Le modèle probabiliste est *non-paramétrique*, et on observe un échantillon de n couples de variables aléatoires (X_i, Y_i), $i = 1, \ldots, n$, indépendants, identiquement distribués. On cherche alors à prédire, au sens d'expliquer, la variable Y par la variable prédictive X. L'intérêt se porte alors sur l'estimation de paramètres de position conditionnels, construits à partir d'un estimateur de la densité conditionnelle. A cet effet, on propose et on étudie un nouvel estimateur de la densité conditionnelle (chapitres 3 à 5).

Prévision statistique paramétrique des processus (chapitres 1 et 2)

Dans le chapitre 1, nous donnons dans un premier temps quelques éléments de la théorie générale de la prévision statistique paramétrique, telle qu'elle est développée par Bosq

dans [24], chapitre 1. Nous particularisons ensuite le cadre de cette théorie à la prévision statistique d'un processus stochastique. A cet effet, nous nous efforçons de clarifier ce problème et de le distinguer des problèmes liés que sont ceux de la prévision probabiliste et de l'estimation de la régression. Une telle approche amène naturellement à adopter un point de vue asymptotique, qui consiste, à partir d'un estimateur $\hat{\theta}_T$ du paramètre inconnu θ qui gouverne la loi du processus, à construire un prédicteur de type "plug-in" , de la fonction de régression $r_\theta(.)$. On obtient alors le prédicteur statistique

$$\hat{X}_{T+h} := r_{\hat{\theta}_T}(X_1, \ldots, X_T).$$

Cependant, le fait que les mêmes données servent à la fois pour l'estimation du paramètre et pour le calcul du prédicteur rend cette approche difficile. A cet effet, nous présentons un moyen de pallier à cette difficulté en découplant ces deux problèmes de façon asymptotique dans le cadre de processus mélangeants. Plus précisément, nous mettons en oeuvre un procédé de séparation temporelle des données et montrons comment l'erreur de prévision peut être approchée asymptotiquement par une erreur quadratique intégrée.

Dans le chapitre 2, nous mettons en oeuvre le procédé annoncé sur une classe générale de processus qui suivent un modèle additif approximativement markovien. Plus précisément, on se place donc dans le cadre suivant où la fonction de régression $r_\theta(.)$ dépend approximativement des k_T dernières valeurs $(X_{T-i}, i = 1, \cdots, k_T)$ avec $k_T \leq T$, $k_T \to \infty$, c'est-à-dire,

$$X^*_{T+1} := E_\theta \left[X_{T+1} \left| X_{-\infty}^T \right. \right] := \sum_{i=0}^{k_T} r_i(X_{T-i}, \theta) + \eta_{k_T}(\mathbb{X}, \theta),$$

où chaque fonction r_i représente la contribution (additive) de la X_{T-i} valeur, et où $\eta_{k_T}(\mathbb{X}, \theta)$ est une fonction de carré intégrable asymptotiquement négligeable dans un sens à préciser. Ce modèle, qui est une extension d'un modèle étudié par Bosq [23], est suffisamment large pour être applicable à un certain nombre de cas particuliers, comme le modèle autorégressif AR. En estimant θ par $\hat{\theta}_{\phi(T)}$ sur l'intervalle $[0, \varphi(T)]$ avec $\varphi(T) \to \infty$,

on construit alors le prédicteur statistique

$$\hat{X}_{t+1} = r_{\hat{\theta}_{\varphi(T)}}(X_{T-k_T}^T).$$

Sous des conditions de mélangeance et de régularité, on peut séparer le problème d'estimation de θ sur $[0, \varphi(T)]$ du problème probabiliste sur la "mémoire" du processus entre $[T - k_T, T]$. En effet, nous obtenons la consistance asymptotique dans le théorème suivant,

Théorème 2.5 *Si les hypothèses* $\mathbf{H_0}$,$\mathbf{H_1}$,$\mathbf{H_2}$ *sont vérifiées, alors*

$$\limsup_{T\to\infty} E_\theta(\hat{X}_{T+1} - X_{T+1}^*)^2 = 0$$

et la loi limite du prédicteur statistique, dans le théorème suivant.

Théorème 2.10 *Si les hypothèses* $\mathbf{H}'_0$,$\mathbf{H}'_1$,$\mathbf{H}'_2$ *sont vérifiées, alors*

$$\sqrt{\varphi(T)}(\hat{X}_{T+1} - X_{T+1}^*) \overset{d}{\rightsquigarrow} < U, V >$$

où U *et* V *sont deux variables indépendantes,* U *de loi* $\mathcal{N}(0, \sigma^2(\theta))$ *et* V *la loi limite de* $\sum_{i=0}^{+\infty} \partial_\theta r_i(X_{T-i}; \theta)$.

Ces résultats sont ensuite appliqués sur des exemples.

Estimation non-paramétrique de la densité conditionnelle et application à la prédiction (chapitres 3 à 6)

Dans la première partie de cette thèse, nous avons cherché comment découpler le problème d'estimation et le problème de prévision pure dans le problème mixte de la prévision statistique. Par suite, dans la prévision de Y par X, ceci nous a amené à nous intéresser à étudier la structure de dépendance entre X et Y. Suites à des considérations de symétrie, entendue au sens d'invariance, ces investigations se sont concrétisées dans la proposition d'un nouvel estimateur de la densité conditionnelle.

Plus précisément, dans le chapitre 3, on se donne un échantillon de n couples de variables aléatoires (X_i, Y_i) indépendants, identiquement distribués et à valeurs dans $\mathbb{R} \times \mathbb{R}$, et on cherche à construire un estimateur non paramétrique de la densité conditionnelle $f(y|x)$ de Y sachant que $X = x$. Une première approche, exploitée dans la littérature notamment par Rosenblatt [114], Bosq [20] et Roussas [115], consiste à *estimer les densités* f_{XY} de (X, Y) et f de X par des estimateurs non paramétriques, notamment de type Parzen [102] Rosenblatt [113], pour former un estimateur de forme quotient, à partir de la définition de la densité conditionnelle, comme

$$f(y|x) = \frac{f_{XY}(x, y)}{f(x)}.$$

Une seconde approche consiste à *transformer* les données Y_i en pseudo-données,

$$Y_i' := K_h(Y_i - y) := \frac{1}{h} K\left(\frac{Y_i - y}{h}\right)$$

où K est un noyau de Parzen-Rosenblatt, et exploiter le fait que, par un théorème de type Bochner,

$$E\left(Y'|X = x\right) \approx f(y|x), \quad h \to 0,$$

pour effectuer la régression des pseudo-données Y_i' sur les X_i par des techniques non-paramétriques telles que Nadaraya-Watson [99, 144], polynômes locaux, projections, etc.... L'approche que nous proposons consiste en quelque sorte à combiner des éléments de ces deux approches: en transformant les données en X et Y par leurs fonctions de répartition marginales respectives F et G, l'expression de la densité conditionnelle s'écrit sous la forme du produit suivant

$$f(y|x) = g(y)c(F(x), G(y))$$

où g est la densité de Y et c est la densité de *copule*, i.e. la densité jointe du couple de variable transformées $(U, V) := (F(X), G(Y))$. Cette notion de copule, introduite par Sklar [122] au travers du théorème qui porte maintenant son nom, permet de séparer l'aléa qui dépend uniquement des marges du vecteur (X, Y), de l'aléa qui dépend uniquement

de la structure de dépendance entre X et Y. L'estimateur est alors construit à partir des fonctions de répartition empirique et d'estimateurs à noyau des densités, et s'écrit sous la forme suivante,

$$\begin{aligned}\hat{f}_n(y|x) :=& \left[\frac{1}{nh_n}\sum_{i=1}^{n} K_0\left(\frac{y-Y_i}{h_n}\right)\right] \cdot \left[\frac{1}{na_n^2}\sum_{i=1}^{n} K_1\left(\frac{F_n(x)-F_n(X_i)}{a_n}\right)\right. \\ & \left. K_2\left(\frac{G_n(y)-G_n(Y_i)}{a_n}\right)\right] \\ :=& \hat{g}_n(y)\hat{c}_n(F_n(x), G_n(y)).\end{aligned}$$

Nous étudions ses propriétés asymptotiques et obtenons, à partir des résultats classiques de convergence des estimateurs à noyau de la densité et sous les conditions usuelles de régularité sur les densités et les noyaux, les résultats suivants:

- Consistance ponctuelle en probabilité,

  Théorème 3.5 *Sous les conditions de régularité sur les densités et les noyaux, si h_n et a_n tendent vers zéro quand $n \to \infty$ de façon que $nh_n \to \infty$, $na_n^4 \to \infty$, $\frac{\sqrt{\ln\ln n}}{na_n^3} \to 0$, alors*

  $$\hat{f}_n(y|x) = f(y|x) + O_P\left(\frac{1}{\sqrt{nh_n}} + h_n^2 + a_n^2 + \frac{1}{\sqrt{na_n^2}} + \frac{1}{na_n^4} + \frac{\sqrt{\ln\ln n}}{na_n^3}\right).$$

  Un choix de fenêtres satisfaisant $a_n \simeq n^{-1/6}$ et $h_n \simeq n^{-1/5}$ donne les vitesses de convergences optimales pour les régularités considérées, ici $n^{-1/3}$.

- Consistance ponctuelle presque sûre,

  Théorème 3.7 *Sous les conditions de régularité sur les densités et les noyaux, si $h_n \to 0$ et $a_n \to 0$ tels que $\frac{nh_n}{\ln\ln n} \to \infty$, $\frac{(\ln n)^{1/2}(\ln\ln n)^{1/2}}{na_n^3} \to 0$, $\frac{\ln\ln n}{na_n^4} \to 0$, alors*

  $$\begin{aligned}\hat{f}_n(y|x) =& f(y|x) \\ & + O_{a.s.}\left(h_n^2 + \sqrt{\frac{\ln\ln n}{nh_n}} + a_n^2 + \sqrt{\frac{\ln\ln n}{na_n^2}} + \frac{\ln\ln n}{na_n^4} + \frac{(\ln n)^{1/2}(\ln\ln n)^{1/2}}{na_n^3}\right).\end{aligned}$$

Un choix de fenêtres satisfaisant $a_n \simeq \left(\frac{\ln\ln n}{n}\right)^{1/6}$ et $h_n \simeq ((\ln\ln n/n))^{1/5}$ donne les vitesses de convergences optimales pour les régularités considérées, ici $(\ln\ln n/n)^{1/3}$.

- Convergence en loi,

  Théorème 3.9 *Sous les conditions de régularité sur les densités et les noyaux, si* $h_n \to 0$, $a_n \to 0$, *tels que*

  $$nh_n \to \infty, \quad \frac{\sqrt{\ln\ln n}}{na_n^3} \to 0, \quad na_n^4 \to \infty, \quad na_n^6 \to 0,$$

  alors

  $$\sqrt{na_n^2}\left(\hat{f}_n(y|x) - f(y|x)\right) \overset{d}{\rightsquigarrow} \mathcal{N}\left(0, g(y)f(y|x)||K||_2^2\right).$$

Ces résultats sont ensuite étendus sur des compacts de $\mathbb{R}$, dans les théorèmes suivants:

- consistance uniforme en y sur un compact en probabilité,

  Théorème 3.18 *Sous les conditions de régularité sur les densités et les noyaux, si* $h_n \simeq (\ln n/n)^{1/5}$ *et* $a_n \simeq (\ln n/n)^{1/6}$, *alors, pour* x *appartenant à l'intérieur du support de* f *et un intervalle* $[a,b]$ *inclus dans l'intérieur du support de* g,

  $$\sup_{y\in[a,b]} |\hat{f}_n(y|x) - f(y|x)| = O_p\left(\left(\frac{\ln n}{n}\right)^{1/3}\right).$$

- consistance uniforme en y sur un compact presque sûrement,

  Théorème 3.18 *Sous les conditions de régularité sur les densités et les noyaux, si* $h_n \simeq (\ln n/n)^{1/5}$ *et* $a_n \simeq (\ln n/n)^{1/6}$, *alors, pour* x *appartenant à l'intérieur du support de* f *et un intervalle* $[a,b]$ *inclus dans l'intérieur du support de* g,

  $$\sup_{y\in[a,b]} \left|\hat{f}_n(y|x) - f(y|x)\right| = O_{a.s.}\left(\left(\frac{\ln n}{n}\right)^{1/3}\right).$$

De même que pour le cas ponctuel, le choix de fenêtres précédent donne une vitesse de convergence uniforme optimale pour les régularités considérées.

Dans le chapitre 4, les propriétés asymptotiques théoriques obtenues dans le chapitre précédent sont complétées par des discussions dans plusieurs directions. Dans un premier temps, nous nous intéressons à l'efficacité des transformations empiriques des données par F_n et G_n. En nous appuyant notamment sur les travaux de Reiss [109] et la notion de déficience introduite par Hodges et Lehmann [75], nous suggérons une modification de notre estimateur initial, où les fonctions de répartitions empiriques sont remplacées par des estimateurs lissés $\hat{F}$, $\hat{G}$, à noyau. De façon similaire, lorsque le support des densités marginales est borné, nous suggérons d'utiliser les estimateurs basés sur les polynômes de Bernstein introduits par Vitale [139]. Nous en profitons alors pour établir une connection heuristique avec les estimateurs de la densité à fenêtre locale. La consistance de ces estimateurs modifiés est montrée dans le corollaire 4.3. Enfin, nous présentons une extension de l'estimateur dans un cadre semi-paramétrique, lorsque de l'information supplémentaire concernant la distribution de la variable explicative X peut-être incorporée dans un modèle paramétrique (proposition 4.5). Dans un deuxième temps, nous nous intéressons à l'implantation numérique de l'estimateur proposé. D'autres considérations sur la transformation des données nous amènent à recommander aussi l'utilisation d'autres estimateurs que les fonctions de répartition empirique. Nous recensons ensuite quelques techniques de réduction du biais de l'estimation de la densité de copule, notamment l'utilisation des noyaux Beta proposés par Chen [35], ceci afin d'améliorer sensiblement les performances à échantillon fini de l'estimateur. Enfin, nous esquissons une stratégie de sélection des fenêtres. Dans un troisième temps, nous effectuons une brève simulation numérique afin de comparer en pratique notre estimateur à ses concurrents. Après une comparaison théorique des biais et variances asymptotiques où nous montrons que notre estimateur a une variance plus petite dès que le produit des densités marginales est inférieur à l'unité, nous mettons en évidence sur un modèle la différence de comportement liée à la structure produit de notre estimateur face aux estimateurs alternatifs basés sur des structures de quotients. Plus précisément, les résultats numériques semblent montrer

un comportement prometteur de l'estimateur lorsque l'on s'intéresse à l'estimation de la densité conditionnelle pour de grandes valeurs de x, i.e. dans un domaine où l'on dispose de peu de données X_i, ce qui pourrait s'avérer potentiellement intéressant pour l'inférence d'événements extrêmes.

Dans le chapitre 5, nous montrons comment utiliser l'estimation de la densité conditionnelle comme une première étape pour construire des prédicteurs. On s'intéresse notamment aux prédicteurs ponctuels que sont le mode et la moyenne conditionnelle, et aux intervalles de prévision couvrant une probabilité α donnée, comme les régions de plus grande densité, i.e. l'ensemble des valeurs de y telles que la densité conditionnellle dépasse un seuil f_α. En particulier, en ce qui concerne le mode conditionnel, nous montrons, à partir des résultats de convergence uniforme de la densité du chapitre 3, sa consistance dans la proposition suivante,

Proposition 5.2 *Sous des hypothèses adéquates, l'estimateur du mode conditionnel* $\hat{\theta}(x) = \arg\sup_{y\in S'} \hat{f}_n(y|x)$ *converge presque sûrement vers le mode conditionnel* $\theta(x) = \arg\sup_{y\in S'} f(y|x)$.

et nous donnons quelques remarques sur son implémentation pratique. De façon similaire, nous établissons la convergence de l'intervalle de prévision $[\hat{y}_\alpha, \hat{y}^\alpha]$ empirique, obtenu à partir de l'estimateur de la densité conditionnelle, vers l'intervalle de prévision théorique $[y_\alpha, y^\alpha]$ dans la proposition suivante,

Proposition 5.4 *Si le seuil empirique* $\hat{f}_\alpha$ *converge p.s. vers le seuil théorique* f_α*, alors* $\hat{y}_\alpha \overset{a.s}{\to} y_\alpha$ *et* $\hat{y}^\alpha \overset{a.s}{\to} y^\alpha$*,*

et indiquons aussi des méthodes pour sa détermination pratique. Enfin, nous nous intéressons à la moyenne conditionnelle, i.e. la fonction de régression, et établissons un résultat de consistance dans le cas simple où le support de Y est bornée (Proposition 5.5) ainsi qu'une connection asymptotique heuristique dans le lemme 5.7 avec l'estimateur des plus proches voisins en rangs de Yang [145] et Stute [128] similaire à celle qui existe entre l'estimateur de la densité conditionnelle à double noyau de Rosenblatt et Roussas

[114, 115] et l'estimateur de la régression de Nadaraya et Watson [99, 144].

En conclusion de cette partie, nous dressons brièvement dans le chapitre 6 des perspectives de recherche et d'applications possibles de cet estimateur, notamment pour l'inférence et l'estimation d'évènements rares ou extrêmes.

Chapter 1

Some generalities on statistical prediction

Abstract : This chapter is of introductory nature. Its aim is to introduce the subject of statistical prediction by formulating the statistical prediction problem of predicting an unobserved value X_{T+h}, with $h > 0$, of a stochastic process $(X_t)_{t\in\mathbb{T}}$ given an observed sample path $(X_t)_{0\leq t\leq T}$, discussing its connections with other related topics, and preparing the method of prediction developed in chapter 2. Starting from general considerations on statistical prediction in section 1.1, we progressively clarify this problem from the closely related ones of probabilistic prediction and regression estimation in section 1.2. In particular, in the case where the law of the process is governed by an unknown parameter θ, we show that solving the problem amounts to calculating a plugged version of the estimated conditional expectation. From a statistical standpoint, these two issues, estimating the parameter and calculating the conditional expectation, are coupled, making the study of the statistical predictor difficult. To that end, we propose in the mixing framework a data-splitting device to separate these two issues in section 1.3. More precisely, we show that an approximation of the prediction error is asymptotically close to the true prediction

error. A concrete application of this device to a class of processes is developed in chapter 2.

1.1 Some elements of statistical prediction theory

In this section, we give some elements of the general parametric statistical prediction theory, as developed by Bosq in [24], chapter 1.

1.1.1 The general prediction model

Let $(\Omega, \mathcal{A}, P)$ a probability space and note $\mathcal{X}$ and $\mathcal{Y}$ two sub-algebras of $\mathcal{A}$, standing for the collection of observed and non-observed events, respectively. To be more specific, assume the probability law is indexed by a parameter $\theta \in \Theta$, where Θ is a parameter space, so that we are given the parametric statistical model $(\Omega, \mathcal{A}, P_\theta, \theta \in \Theta)$. Moreover, assume that $\mathcal{X} = \sigma(X)$, where X stands for the collection of observed random variables and takes its values in a measurable space $(\mathbb{X}, \mathscr{X})$.

In the prediction of a $\mathcal{Y}$-measurable real-valued random variable Y by a $\mathcal{X}$-measurable random variable X, one is interested in finding a function p, which be $\mathcal{X}$-measurable, such that $p(X)$ represents a good approximation of the unobserved Y. More generally, one can also consider the prediction of a given known function of X, Y and θ. To that purpose, define the *predictand* $g(X, Y, \theta) \in \cap_{\theta\in\Theta}\mathbb{L}^2(P_\theta)$, where g is a known function, and the *predictor* $p(X)$ to be any measurable function of X, where p is known, assumed also to be such that $p(X) \in \cap_{\theta\in\Theta}\mathbb{L}^2(P_\theta)$.

The quality of this approximation is evaluated by the *quadratic prediction error*

$$R_\theta(p, g) := E_\theta(p(x) - g(X, Y, \theta))^2, \quad \theta \in \Theta. \tag{1.1}$$

Such a criteria induces the following preference relation:

Definition 1.1 *The predictor p_1 is preferred to the predictor p_2 for predicting g, if and*

only if $R_\theta(p_1, g) \leq R_\theta(p_2, g), \quad \theta \in \Theta$. *If so, one notes* $p_1 \prec p_2$.

The prediction problem of $g(X, Y, \theta)$ can be split into several cases, depending on the structure of the predictand considered:

- if g is a function of Y only, one has a *pure prediction* problem;
- if g is a function of θ only, one has a *pure estimation* problem;
- in any other case, one has a *mixed* problem.

1.1.2 Decomposition of the prediction error

The simple lemma below, which belongs to folklore, gives a fundamental decomposition of the quadratic prediction error, (see e.g. lemma 1.1 of [24]):

Lemma 1.2 *The following decomposition of the Quadratic Prediction Error holds,*

$$\begin{aligned} R_\theta(p, g) &= E_\theta[(g(X, Y, \theta) - E_\theta[g(X, Y, \theta)|X])^2] \\ &\quad + E_\theta[(E_\theta[g(X, Y, \theta)|X] - p(X))^2] \\ R_\theta(p, g) &:= PPE_\theta(g) + SPE_\theta(p, g) \end{aligned} \tag{1.2}$$

Hence, $p_1 \prec p_2$ *for predicting* g *if and only* $p_1 \prec p_2$ *for predicting* $E_\theta[g|X]$.

Proof *It simply follows from the Pythagorean property of the conditional expectation.*

The consequences of this lemma are twofold:

- A lower bound on the prediction error is given by the *Probabilistic Prediction Error* $PPE_\theta(g) := E_\theta[(g(X, Y, \theta) - E_\theta[g(X, Y, \theta)|X])^2]$, which is not controllable by the Statistician. Therefore, the Statistician can only try to minimise the *Statistical Prediction Error* $SPE_\theta(p, g) := E_\theta[(E_\theta[g(X, Y, \theta)|X] - p(X))^2]$.

- For the case where g is a known function of Y only, i.e. when $g(X, Y, \theta) = g(Y)$, predicting $g(Y)$ is therefore equivalent to predicting $E_\theta[g(Y)|X]$. The latter being a function of X and θ, one sees that, in general, a non degenerate prediction problem is mixed.

1.1.3 Statistical prediction theory and statistical estimation theory

Parallelling the theory of parametric statistical estimation theory as exposed, e.g. in Lehmann and Casella [92], an analogue theory can be constructed as developed in Bosq and Blanke [24] and Bosq [22], which we briefly sketch below. In the context of prediction, the counterpart of sufficient statistics are prediction sufficient statistics, which add a conditional independence condition to the sufficiency one . Thanks to such prediction sufficient statistics, a Rao-Blackwell type theorem for prediction can be formulated. Optimality can be investigated in a similar fashion to that of UMVU estimation. Admissibility considerations implies that the search for optimal predictors is restricted to the class of unbiased ones. The unique optimal unbiased predictor is then characterised by a Lehmann-Scheffé theorem. Under regularity conditions, Cramér-Rao bounds can be obtained. We refer the interested reader to the detailed exposition of Bosq and Blanke [24].

Remark 1.3 *Other criteria than the quadratic prediction error can also be considered. In a decision theoretical framework à la Wald [141], a risk is defined through the expectation of a positive loss function L by $E_\theta[L(p, g)]$. Special interest lie in loss functions associated with a location parameter $\mu = \mu_\theta$. In the real valued case, for a random variable Z, μ is defined as $E_\theta L(Z, \mu) = \min_{a\in\mathbb{R}} E_\theta L(Z, a)$. Since $E_\theta L(g, p) = E_\theta[E_\theta[L(g, p)|X]]$ is minimum for $p_\theta(X) = \mu_\theta(X)$, the following classical loss functions allow to define the corresponding point predictors,*

- *for the square loss* $L(u,v) = (u-v)^2$, $\mu_\theta(X)$ *is the conditional mean* $E_\theta[g|X]$;
- *for the absolute value loss* $L(u,v) = |u-v|$, $\mu_\theta(X)$ *is the median of the conditional distribution of* g *given* X;
- *for the* $0-1$ *loss* $L(u,v) = \mathbb{1}_{|u-v|>\epsilon}$ *for* $\epsilon > 0$, $\mu_\theta(X)$ *is the mode of the conditional distribution of* g *given* X;

The advantage of the square loss is that it simplifies the mathematical treatment of the problem as shown, e.g., in lemma 1.2 above. Other choices such are possible such as the $0-1$ *loss, especially for random variables taking their values in a discrete set, or the absolute value loss or the Huber loss (See Huber [79]), often motivated by robustness considerations, which we will not pursue here. A different approach in a non parametric framework, by estimation of these conditional locations parameters is pursued in chapter 5.*

1.2 Asymptotic prediction for a stochastic process

In this section, we particularise the above framework to the context of a time series, and intend to clarify it with the related problems of probabilistic prediction and regression estimation.

1.2.1 Formulation of the problem for a stochastic process

To that purpose, let $(X_t)_{t\in\mathbb{T}}$ be a real-valued square integrable stochastic process defined on a probability space $(\Omega, \mathcal{A}, P_\theta)$, with $\theta \in \Theta$. For $\mathbb{T} = \mathbb{Z}$, $(X_t)_{t\in\mathbb{Z}}$ is a discrete time stochastic process, and one observes the past $X = (X_1, \ldots, X_T)$ and intends to predict the future $Y = X_{T+h}$, where $h \in \mathbb{N}$ is the *horizon*. For $\mathbb{T} = \mathbb{R}$, $(X_t)_{t\in\mathbb{R}}$ is a continuous time stochastic process, and one observes $X = (X_t, 0 \leq t \leq T)$ and intends to predict $Y = X_{T+h}$, where $h > 0$ is the horizon.

In various situations, no optimal predictor may exist (see e.g. lemma 1.2 of Bosq and Blanke [24]) or may be extremely difficult to compute (see subsections 1.2.2 below). As is the case for statistical estimation theory, we will show below that it is therefore natural to adopt an asymptotic point of view, with sample size tending to infinity. To that purpose, we index the data X from which a statistical predictor has to be built by time T. To prevent confusion, we rename the X vector as D_T, i.e. we define the observed past $D_T := (X_1, \ldots, X_T)$ or $D_T := (X_t, 0 \leq t \leq T)$ for discrete or continuous time, respectively. The statistical predictor will be noted $\hat{X}_{T+h}$ and thus is a random variable which is $\sigma(D_T)$ measurable, i.e. such as there exists a function $p \in \mathbb{L}^2(P)$ such that $\hat{X}_{T+h} = p(D_T)$. We are bound to make the following distinction in order to separate the problems.

1.2.2 Probabilistic and Statistical prediction

Probabilistic Prediction

In the present context, lemma 1.2 writes

$$E_\theta(X_{T+h} - p(D_T))^2 = E_\theta[X_{T+h} - E_\theta(X_{T+h}|D_T)]^2 + E_\theta[E_\theta(X_{T+h}|D_T) - p(D_T)]^2 \tag{1.3}$$

and the theoretical answer to the minimisation problem of the statistical prediction error is given by the so-called *Bayes* or *Probabilistic* predictor, defined as

$$X^*_{T+h} := E_\theta(X_{T+h}|D_T). \tag{1.4}$$

From a probabilistic standpoint, i.e. assuming the knowledge of the parameter θ, the prediction problem thus reduces to the calculation of the conditional expectation 1.4.

Examples of probabilistic prediction

Depending on the assumed model on the process, the calculation of the conditional expectation may be more or less difficult. This problem has been tackled by numerous people and a huge literature is devoted to the subject. We sketch below only a glimpse of the topic.

- **Linear prediction** Assume X_t is a Gaussian process in discrete time. Then, since (X_{T+h}, D_T) is a Gaussian vector, the conditional expectation 1.4 reduces to a linear function of the D_T, and the search for the Bayes predictor reduces to the search for the best linear predictor.

- **Kolmogorov-Wiener theory** For weakly stationary square integrable time series and predictors restricted to the class of linear predictors, a complete solution is provided by the work of Kolmogorov and Wiener, cf. e.g. [29].

- **Filtering and Control for Diffusion processes** For a diffusion process satisfying the stochastic differential equation,

$$dX_t = a(X_t, t)dt + \sigma(X_t, t)dW_t$$

  where $(W_t)_{t\in\mathbb{T}}$ is a Wiener process, and the functions a and b are assumed to be known, the conditional expectation calculation was solved by Kallianpur and Zakai. We refer the reader to the vast literature (see e.g. [84] and the references therein) on the related problems of stochastic control and filtering of stochastic processes.

- **Markov** If the process is assumed to be Markovian of order m, then the conditional expectation $E_\theta(X_{T+h}|D_T)$ reduces to $E_\theta(X_{T+h}|X_T, \ldots, X_{T-m+1})$.

Statistical prediction

However, from a statistical standpoint, the solution 1.4 is not satisfactory, since we assumed the knowledge of the underlying model of the process. Indeed, this probabilistic

predictor is not a genuine statistical predictor depending only on the data D_T, since it is also a function of the *unknown* parameter θ. As a consequence, it can not used by the statistician to make a practical prediction.

The mixed nature of the statistical prediction problem, as announced in section 1.1, clearly appears here. One both has

1. a purely statistical problem of estimation of the unknown law of the process built from the data D_T,

2. a purely probabilistic problem of the calculation of the conditional expectation $E_\theta(X_{T+h}|D_T)$, i.e. of calculation of the regression function $r_\theta(d_T) := E_\theta(X_{T+h}|D_T = d_T)$.

As the unknown θ has to be estimated, it is therefore natural to adopt an *asymptotic* point of view. A possible solution to overcome this mixed difficulty may consist in

1. estimating θ by an estimator $\hat{\theta}_T$ from the data D_T,

2. building a plug-in type statistical predictor from the regression function $r_\theta(d_T) := E_\theta(X_{T+h}|D_T = d_T)$, by
$$\hat{X}_{T+h} := r_{\hat{\theta}_T}(D_T).$$

However, the fact that the same data D_T is involved in both problems, renders the asymptotic behaviour of this statistical predictor difficult to study. Before presenting a partial remedy for this issue in section 1.3, we sketch below some clarifications between this prediction problem and related approaches.

1.2.3 Delineation of the asymptotic prediction problem

Note that in the remainder of this chapter, we temporally omit the parameter θ indexing the law of the process, in order to simplify notations.

Taking into account asymptotics, there are several other distinctions we would like to clarify, which depend on the features the Statistician wants to incorporate in the formulation of his problem:

Probabilistic versus Empirical Error

The quadratic prediction error 1.1 writes here as

$$R_T(\hat{X}_{T+h}, X_{T+h}) := E(X_{T+h} - p(D_T))^2. \tag{1.5}$$

Note that this risk, defined as an expected loss, although it gives the theoretical error, is not observable by the Statistician, since the distribution of the process is usually unknown. Consequently, the Statistician can also choose to measure the prediction error by an empirical criteria. To that end, the problem is cast in a sequential fashion. In the discrete time case and when the goal is to predict the next value (i.e. $h = 1$), at each time instant $t = 1, 2, \ldots$, a sequence of predictors $p_t(D_{t-1})$ is constructed, based on the values of $(X_1, X_2, \ldots, X_{t-1})$. After T time instants, the normalised cumulative empirical prediction error of the strategy $\mathbf{p}$ consisting of the sequence of predictors $\{p_t\}$, is

$$R_T^e(\mathbf{p}, X) := \frac{1}{T}\sum_{t=1}^{T}(p_t(D_t) - X_t)^2 \tag{1.6}$$

which is termed the Cesáro loss by [66]. The connections between these probabilistic and empirical sequential errors 1.5 and 1.6 are investigated by Algoet [6, 7], Györfi et al. [66] chapter 27. See also the discussion below on the static and dynamic forecasting problem. This empirical measure of the error allows to reformulate the problem in a repeated game-theoretic framework, building on the pioneering work of Blackwell [19] on approachability theory, as exemplified by the recent monograph of Cesa-Bianchi and Lugosi [32] on the prediction of individual sequences. Since the error is now observable, it can be used as side information in order to find solutions of the minimizing of 1.6 in a recursive way. Such an approach, which has the advantage of not making assumptions on the underlying

process governing the time series, consists in combining several base predictors according to their past performances, mirroring the gradient optimisation algorithms of the dynamic programming paradigm (See Cesa-Bianchi and Lugosi [32] and the references therein).

Assumptions on the process and the predictors

Parallelling the approach of statistical estimation where the Statistician can decide to restrict the search of estimators to the family of unbiased ones - which leads to the development of UMVU estimation, the Statistician may also decide to set limitations on the space of possible predictors. One can distinguish mainly between two kind of limitations,

- Shape constraints: the Statistician can make structural assumptions on the shape of the possible functions p, e.g. to be linear in the predictands, i.e. in the discrete time case, $p(D_T) = \sum_{i=1}^{T} a_i X_i$, where $a_i \in \mathbb{R}$, $i = 1, \ldots, T$.

- Memory size: instead of taking into account all of the possible past to make his prediction, he can decide to limit himself to finite memory predictors, i.e. functions of the m-proximity past, $p(D_T) = p(X_T, X_{T-1} \ldots, X_{T-m+1})$ in the discrete time case.

In a dual manner, the Statistician can make assumptions on the stochastic process governing the observed time series. Among others, he can impose some structure such as a Gaussian or a Markov one. These limitations in the predictors space are strongly connected with the assumptions one is willing to make on the stochastic process. Gaussian and Markov hypothesis lead naturally to these limitations, as was mentioned in the examples of subsection 1.2.2 above.

Static versus Dynamic forecasting.

Taking into account asymptotics, there are two more distinctions to be made for the prediction of stationary ergodic processes, depending on the way one goes to infinity, as formulated by Cover [38].

- Dynamic forecasting: One fixes the beginning of the observed series, and search for a prediction of the process in the infinite future. In other words, $D_T = (X_1, \ldots, X_T)$ is the data, X_{T+1} is the value to be predicted, and one looks for a sequence of predictors $\mathbf{p} = (p_T)$, such that

$$\lim_{T\to\infty} |p_T(D_T) - E[X_{T+1}|D_T]| = 0 \text{ a.s.}$$

- Static forecasting: One fixes the time of prediction, and look back in the increasing past $D'_T = (X_{-T}, \ldots, X_{-1}, X_0)$ to make a prediction of X_1. In other words, one look for a predictor p such that

$$\lim_{T\to\infty} p(D'_T) = E[X_1|X_0, X_{-1}, \ldots, X_{-\infty}] \text{ a.s.}$$

We refer the reader to [66] chapter 27 and the references therein for a detailed discussion of these topics. In particular, there are some negative findings about a universal solution for the Dynamic forecasting problem, and the performance of the normalised cumulative prediction error 1.6 is intimately linked to the Static forecasting problem, in the sense that for any prediction strategy $\mathbf{p} = \{p_T(D_T)\}$ and stationary ergodic process (X_t),

$$\liminf_{T\to\infty} R^e_T(p) \geq R^*$$

where

$$R^* = E[(X_1 - E(X_1|X_0, X_{-1}, \ldots, X_{-\infty}))^2].$$

1.2.4 Statistical Prediction is not regression estimation

To complement the above distinctions and introduce the approach of the next section, we discuss the relation between statistical prediction and regression estimation, inspired by Györfi et al [66].

Regression estimation with i.i.d. data

In the classical regression in discrete time setting, assume that we have an i.i.d. sample $D_n := (X_i, Y_i)_{t=1,\dots,T}$ from variables (X, Y). One wants to predict Y by X. In a first stage, the regression function $r(x) := E[Y|X = x]$ is estimated by a function $\hat{r}(x, D_T)$ from this data. In a second stage, assume we have a new observation $(\tilde{X}, \tilde{Y})$, which has the same law as (X, Y) and is independent of the data D_T. The $\tilde{Y}$ value is predicted by $\hat{r}(\tilde{X}, D_T)$. In that case, since $\tilde{X}$ is independent of D_n, by conditioning on $\tilde{X}$, the statistical prediction error becomes

$$\begin{aligned} E[(r(\tilde{X}) - \hat{r}(\tilde{X}, D_T))^2] &= \int E[(r(\tilde{X}) - \hat{r}(\tilde{X}, D_T))^2 | \tilde{X} = x] dP_{\tilde{X}}(x) \\ &= \int E[(r(x) - \hat{r}(x, D_T))^2] dP_X(x) \end{aligned}$$

as $E[\tilde{Y}|\tilde{X}, D_T] = E[\tilde{Y}|\tilde{X}]$, and the prediction error is the same as the Mean Integrated Square Error (MISE) of the regression. Therefore, the prediction problem reduces, in the i.i.d case, to the estimation of the regression function.

Regression estimation for dependent data

Now, assume the data is no longer i.i.d. but is a time series. In other words, consider for example that $(\zeta_t)_{t\in\mathbb{N}}$ is a strictly stationary Markov chain. The data D_T is now made of $(X_t, Y_t) = (\zeta_t, \zeta_{t+1})$ for $t = 0, \dots, T-1$. The (auto)-regression function $r(x) = E[Y|X = x] = E[\zeta_1|\zeta_0 = x]$ can still be estimated from the data by a function $\hat{r}(x, D_T)$. Prediction of the next outcome means that one is now interested in $(\tilde{X}, \tilde{Y}) = (\zeta_T, \zeta_{T+1})$. In that

situation, $\tilde{X}$ is no longer independent of the data D_T, which entails that $E[\tilde{Y}|\tilde{X}, D_T] \neq E[\tilde{Y}|\tilde{X}]$. Therefore, one may have that

$$\min_p E\left[\tilde{Y} - p(\tilde{X}, D_T)\right]^2 = E\left[\tilde{Y} - E(\tilde{Y}|\tilde{X}, D_T)\right]^2$$
$$< \min_p E\left[Y - p(X)\right]^2 = E\left[Y - E[Y|X]\right]^2$$

where the inequality is strict. That is to say that one can find theoretically a statistical predictor which has a lower prediction error than that of the regression error, and the predictor obtained from plugging $\tilde{X}$ into the regression estimator $\hat{r}(x, D_T)$ is no longer optimal.

Towards independence

Assume now that $(\tilde{X}, \tilde{Y})$ is distributed as (X_0, Y_0) and be independent from D_T. Then $E[\tilde{Y}|\tilde{X}, D_T] = E[\tilde{Y}|\tilde{X}]$ and consequently

$$\min_p E\left[Y - p(\tilde{X}, D_T)\right]^2 = E\left[Y - E(\tilde{Y}|\tilde{X}, D_T)\right]^2$$
$$= E\left[\tilde{Y} - E(\tilde{Y}|\tilde{X})\right]^2 = \min_p E\left[\tilde{Y} - p(\tilde{X})\right]^2$$

Therefore, the prediction of $\tilde{Y}$ by $\tilde{X}$ reduces to the estimation of the regression function $r(x) = E[Y|X = x]$ and the predictor is directly obtained by plugging $\tilde{X}$ in the regression function, $\hat{X}_{T+1} = \hat{r}(\tilde{X}, D_T)$.

Another look of this phenomenon is through the statistical prediction error,

$$E\left[r(\tilde{X}) - \hat{r}(\tilde{X}, D_T)\right]^2 = \int E\left[\left(r(\tilde{X}) - \hat{r}(\tilde{X}, D_T)\right)^2 |\tilde{X} = x\right] dP_{\tilde{X}}(x)$$

If D_T and $\tilde{X}$ are independent, then

$$E\left[\left(r(\tilde{X}) - \hat{r}(\tilde{X}, D_T)\right)^2 |\tilde{X} = x\right] = E\left[\left(r(x) - \hat{r}(x, D_T)\right)^2\right]$$

and therefore,

$$E\left[r(\tilde{X}) - \hat{r}(\tilde{X}, D_T)\right]^2 = \int E\left[\left(r(x) - \hat{r}(x, D_T)\right)^2\right] dP_X(x)$$

and as in the i.i.d. case, the prediction error is the same as the MISE error of the regression function. As a consequence, the same rate of convergence as those of the regression estimation would be obtained for the prediction.

However, it is difficult to assume in practice that the statistician may have at his disposal such independent auxiliary random variables $(\tilde{X}, \tilde{Y})$, since usually $(\tilde{X}, \tilde{Y}) = (X_T, X_{T+h})$. Nonetheless, it is shown in the next section how to implement a substitute of this idea in a mixing context.

1.3 Asymptotic decoupling by temporal separation

The discussion of the preceding subsection has shown that it would be desirable to have at our disposal, an extra sample $(\tilde{X}, \tilde{Y})$ that be distributed as (X, Y) but such that $\tilde{X}$ be independent of the observed data. In this section, we substantiate this approach in the mixing context, by setting up a data-splitting device and show how the prediction error can thus be approximated.

1.3.1 Time splitting

The data-splitting device consists in making the data D_T and the predictive variable $\tilde{X}$ asymptotically independent by splitting the sample in two subsamples separated by an increasing gap: in the mixing context, a way to achieve this result is

- to estimate the regression function $r(x)$ on the data D_{T-k_T}, where $k_T \to \infty$ and $k_T = o(T)$,
- and to take as predictive variables the closest portion of the past $\tilde{X} = (X_{T-\pi_t}, \ldots, X_T)$, with $\pi_T \geq k_T$ in such a way that $\pi_T - k_T \to \infty$.

With that device, the fluctuation induced by the predictive variable $\tilde{X} = (X_{T-\pi_T}, \ldots, X_T)$ in the plugging in the regression function is asymptotically separated from the variability

of the estimation of the regression function, based on the data $(X_1, \ldots, X_{T-k_T})$. In a purely Markovian setup, i.e. when the process has a fixed amount of "memory", the trick reduces to setting $\tilde{X} = X_T$ and changing the estimation data from D_T to D_{T-k_T}. A more general setup is considered in chapter 2.

1.3.2 Coupling in β-mixing

Before exemplifying this device, we recall some definitions and properties related to the β-mixing coefficients [140].

Definition 1.4 (Volkonski and Rozanov) *Let $(\Omega, \mathcal{A}, P)$ a probability space. For any two σ-algebras $\mathcal{U}, \mathcal{V}$ of $\mathcal{A}$, the β-mixing coefficient between $\mathcal{U}$ and $\mathcal{V}$ is defined by*

$$\beta(\mathcal{U}, \mathcal{V}) = \frac{1}{2} \sup \left(\sum_{i \in I} \sum_{j \in J} |P(U_i \cap V_j) - P(U_i)P(V_j)| \right)$$

where the supremum is taken over all the partitions $(U_i)_{i \in I}$ and $(V_j)_{j \in J}$ of Ω, with $U_i \in \mathcal{U}$ and $V_j \in \mathcal{V}$.

According to Delyon [44], quoted by Viennet [138], one has the following lemma:

Lemma 1.5 (Delyon) *Let X and X' two random variables with values in the separable metric spaces E and E' respectively. Then, there exists two positive functions $b : E \to [0,1]$ and $b' : E' \to [0,1]$ such that:*

- $b \in L_1(P_X)$ *and* $b' \in L_1(P'_X)$
- $\beta(X, X') = \int b(X) P_X(dx) = E_X(b(X))$
- $\beta(X, X') = \int b'(X') P_{X'}(dx') = E_{X'}(b(X'))$

and such that, for every positive bounded function g and g' measurable with respect to X and X' respectively, one also has:

$$\begin{aligned} E_X(g(X)b(X)) &:= \int g(X)b(X)P_X(dx) \\ &= \frac{1}{2} \iint g(x) \left| P_{X,X'} - P_X \otimes P_{X'} \right| (dx, dx') \end{aligned}$$

$$\begin{aligned} E_{X'}(g'(X')b'(X')) &:= \int g'(X')b'(X')P_{X'}(dx') \\ &= \frac{1}{2}\iint g'(x')\left|P_{X,X'} - P_X \otimes P_{X'}\right|(dx,dx') \end{aligned}$$

The functions b et b' are the Radon-Nikodyn derivative of the measure $\frac{1}{2}|P_{X,X'} - P_X \otimes P_{X'}|$ with respect to P_X and $P_{X'}$ respectively.

1.3.3 Equivalent risks

We show below, that the temporal separation device allows to asymptotically get equivalent risks, as is shown in Bosq and Blanke [24].

To that end, assume for simplicity that $(X_T)_{T\in\mathbb{Z}}$ is a strictly stationary real-valued square-integrable Markov process such that for every $T \geq 1$, $(X_1,\ldots,X_T)$ has a joint density $f_{X_1,\ldots,X_T}$ with respect to Lebesgue measure $\lambda^{\otimes T}$. Note $f_{D_{T-k_T}}(z)$ the density of $(X_1,\ldots,X_{T-k_T})$ for $(x_1,\ldots,x_{T-k_T}) := z_T$. In that case the data is $D_T = (X_1,\ldots,X_T)$ and the predictive variable $\tilde{X} = X_T$. Moreover, assume the process is β-mixing, in the sense that $\beta_{k_T} = \beta(\sigma(D_{T-k_T}),\sigma(X_T)) \to 0$ as $k_T \to \infty$. We note $r(.) = E[X_1|X_0 = .]$ the regression function for which we assume we have an estimator $\hat{r}(.,D_T)$. We want to approximate the quadratic statistical predictive risk

$$\begin{aligned} I_{T-k_T} &:= E\left[r(X_T) - \hat{r}(X_T, D_{T-k_T})\right]^2 \\ &= \iint (\hat{r}(x,z_T) - r(x))^2 f_{X,D_{T-k_T}}(x,z_n)dxdz_T \end{aligned}$$

by its counterpart we would have had, under the proviso of an independent $\tilde{X}_T$:

$$\begin{aligned} J_{T-k_T} &:= \int E\left[\hat{r}(x,D_{T-k_T}) - r(x)\right]^2 f_X(x)dx \\ &= \iint \left[\hat{r}(x,z_T) - r(x))\right]^2 f_X(x) f_{D_{T-k_T}}(z_T)dxdz_T \end{aligned}$$

One has the following proposition, see lemma 2.1 of [24],

Proposition 1.6 *(Dedecker) Assume,* $g(D_{T-k_T}) = \sup_{x\in\mathbb{R}}\left(\hat{r}(x,D_{T-k_T}) - r(x)\right)^2 < \infty$ *a.s. Then,*

1. *if* $Eg^p(D_{T-k_T}) < \infty$ *for a* $p > 1$, *then* $|I_{T-k_T} - J_{T-k_T}| \leq 2\beta_{k_T}^{1-1/p}||g(.)||_p$

2. *if* $Eg(D_{T-k_T}) < \infty$, *then* $|I_{T-k_T} - J_{T-k_T}| \leq 2\phi_{k_T}||g(.)||_1$

3. *if* $||g(D_{T-k_T})||_\infty < \infty$, *then* $|I_{T-k_T} - J_{T-k_T}| \leq 2\beta_{k_T}||g(.)||_\infty$

Proof *Note that the boundedness assumption on* g *is fulfilled, if e.g.* r *and* $x \to \hat{r}(x, z_T)$ *are bounded. We have,*

$$\begin{aligned}|I_{T-k_T} - J_{T-k_T}| &\leq \iint \left[r(x) - \hat{r}(x, z_T)\right]^2 \left| f_{X,D_{T-k_T}}(x, z_T) - f_{D_{T-k_T}}(z_T) f_X(x) \right| dx dz_T \\ &\leq \iint g(z_T) \left| f_{X,D_{T-k_T}}(x, z_T) - f_{D_{T-k_T}}(z_T) f_X(x) \right| dx dz_T\end{aligned}$$

By using Delyon's result (lemma 1.5), there exists a $b(z_T)$ *function such that*

$$|I_{T-k_T} - J_{T-k_T}| \leq 2 \int g(z_T) b(z_T) f_{D_{T-k_T}}(z_T) dz_T$$

and such that $||b||_\infty = \varphi_{k_T}$, $||b||_1 = \beta_{k_T}$, *where*

$$\varphi(k_T) = sup_{B \in \sigma(D_{T-k_T}), P(C)>0, C \in \sigma(X_T)} |P(C) - P(C|B)|$$

is the φ *mixing coefficient.*

1. *If* $E(g^p(D_{T-k_T})) < \infty$, *then, by Hölder's inequality,*

$$\begin{aligned}|I_{T-k_T} - J_{T-k_T}| &\leqslant 2E^{1/p}\left(g^p(D_{T-k_T})\right) E^{1-1/p}\left(b^{p/(p-1)}(D_{T-k_T})\right) \\ &\leqslant 2\left\|g\right\|_p \left\|b\right\|_q\end{aligned}$$

with $1/p + 1/q = 1$. *Since* b *takes its values in* $[0, 1]$, *one has* $E\left(\left|b^{p/(p-1)}(D_{T-k_T})\right|\right) \leqslant E\left(b(D_{T-k_T})\right)$, *thus* $\left\|b(D_{T-k_T})\right\|_q \leqslant \beta(k_T)^{1/q}$. *Therefore,*

$$|I_{T-k_T} - J_{T-k_T}| \leqslant 2\left\|g\right\|_p \beta(k_T)^{1-1/p}$$

2. One also have, if $\|b(D_{T-k_T})\|_\infty = \varphi(k_T)$*, that,*

$$|I_{T-k_T} - J_{T-k_T}| \leqslant 2 \|g\|_1 \varphi(k_T)$$

3. If g *is bounded, i.e. if* $\|g\|_\infty < \infty$*, then*

$$|I_{T-k_T} - J_{T-k_T}| \leqslant 2 \|g\|_\infty \beta(k_T)$$

As a consequence, this proposition shows that the integrated quadratic error J_T is asymptotically equivalent to the quadratic prediction error I_T as $T \to \infty$. An application to an additive model is developed in the following chapter 2.

Chapter 2

Asymptotic Statistical prediction for a parametric additive model

Abstract : We show below how to implement the temporal separation device presented in chapter 1 to a general parametric additive model. We show its asymptotic consistency in the mean square sense and derive its limit law. Illustrations for several examples of time series are provided. This section is a reprint from the article "Prévision paramétrique par séparation temporelle", accepted by *Annales de l'ISUP*. Therefore, we warn the reader of some possible slight repetitions with chapter 1.

English summary : Let $\mathbb{X} = \{X_t, t \in \mathbb{Z}\}$ be a real-valued weakly stationary square integrable process, with law indexed by a parameter θ, observed on a time interval $0 \leq t \leq T$. We are interested in forecasting the unobserved random variable X_{T+1} by a function $\hat{X}_{T+1}$ of the observations $(X_i, i = 0, \cdots, T)$, with the quadratic error criteria $E_\theta(\hat{X}_{T+1} - X_{T+1})^2$. It is well known that the conditional expectation $X^*_{T+1} := E_\theta(X_{T+1} \left| X_0^T \right.) := r_\theta(X_0^T)$ is a solution to this minimisation problem. Nonetheless, this probabilistic forecaster is not a genuine statistical one, since it depends on the unknown value of the parameter θ, which has to be estimated by an estimator $\hat{\theta}_T$. The plug-in statistical forecaster induced

$r_{\hat{\theta}_T}(X_0^T)$ is then a difficult object to study. In this paper, we propose to deal with the case where the probabilistic forecaster depends approximately only on the last k_T values of the time series $(X_{T-i}, i = 1, \cdots, k_T)$. By estimating θ by $\hat{\theta}_{\phi(T)}$ on the interval $[0, \varphi(T)]$, we build a statistical predictor $r_{\hat{\theta}_{\varphi(T)}}(X_{T-k_T}^T)$ and show its consistency and derive its limit in distribution under regularity, mixing, and assumptions on k_T and $\varphi(T)$.

2.1 Introduction

2.1.1 Motivation

Soit $\mathbb{X} = \{X_t, t \in \mathbb{Z}\}$ un processus à valeurs réelles faiblement stationnaire de carré intégrable, défini sur $(\Omega, \mathcal{A}, \mathbb{P})$, de loi $\mathbb{P}$ indexée par un paramètre θ à valeurs dans $\mathbb{R}^d$, observé sur $0 \leq t \leq T$. On cherche à prédire la variable aléatoire X_{T+1} non observée par une statistique $\hat{X}_{T+1}$ qui soit $\sigma(X_t, 0 \leq t \leq T)$ mesurable, de carré intégrable et qui minimise l'erreur quadratique $E_\theta(\hat{X}_{T+1} - X_{T+1})^2$. En notant X_a^b la σ-algèbre engendrée par $(X_t, a \leq t \leq b)$ et $r_\theta(X_0^T)$ l'espérance conditionnelle $E_\theta(X_{T+1}|X_0^T)$, rappelons alors le lemme évident suivant:

Lemma 2.1 *L'erreur de prévision se décompose en un terme probabiliste et un terme d'approximation statistique :*

$$E_\theta(X_{T+1} - \hat{X}_{T+1}(X_0^T))^2 = E_\theta(X_{T+1} - r_\theta(X_0^T))^2 + E_\theta(r_\theta(X_0^T) - \hat{X}_{T+1}(X_0^T))^2$$

Le premier terme s'appelle erreur probabiliste et ne dépend que du processus et le second terme s'appelle erreur statistique de prévision et résulte du choix de $\hat{X}_{T+1}$ par le statisticien. L'erreur de prévision est donc minimisée pour le choix de $\hat{X}_{T+1}(X_0^T) = E_\theta(X_{T+1}|X_0^T) := r_\theta(X_0^T)$.

Néanmoins le choix de ce prédicteur, que l'on qualifiera de probabiliste, n'est pas satisfaisant d'un point de vue statistique car le paramètre θ étant inconnu, l'espérance

conditionnelle n'est pas accessible au statisticien. On est donc naturellement amené à construire un estimateur $\hat{r}_T$ de cette espérance conditionnelle $r_\theta(.)$ basé sur l'échantillon $(X_i, i = 0, \cdots, T)$ pour obtenir le prédicteur statistique $\hat{r}_T(X_0^T)$. Dans un cadre paramétrique où l'on suppose la forme de la fonction de régression r_θ connue, cela se traduit par estimer le paramètre θ par $\hat{\theta}_T$ et construire le prédicteur statistique plug-in $r_{\hat{\theta}_T}(X_0^T)$.

Cependant, le fait que les variables (dépendantes) $(X_0, \cdots, X_T)$ servent à la fois dans le problème (statistique) d'estimation de θ et comme valeurs d'entrée dans le calcul (probabiliste) de la fonction de régression, rend l'étude de l'erreur de prévision statistique malaisée. Une manière usuelle de procéder dans la littérature est d'introduire une hypothèse supplémentaire sur la structure de dépendance du processus (voir par exemple Caires et Ferreira [30] pour une discussion), typiquement markovien d'ordre k, afin de simplifier la fonction de régression $r_\theta(X_0^T)$ en $r_\theta(X_{T-k+1}^T)$, ce qui revient à considérer le problème de la prévision à "passé" fini. Dans le cadre simplifié d'un processus ARMA ayant une structure linéaire, la méthode de Box-Jenkins ou du filtre de Kalman (voir par exemple Box et Jenkins [28] ou Brockwell et Davis [29]) permet de traiter ce problème.

2.1.2 Séparation temporelle

On se propose ici de ne pas faire cette hypothèse mais de séparer les problèmes probabiliste et statistique de façon temporelle. On se place dans le cadre où la fonction de régression $r_\theta(.)$ dépend approximativement des k_T dernières valeurs $(X_{T-i}, i = 1, \cdots, k_T)$ avec $k_T \leq T$, $k_T \to \infty$.

$$X_{T+1}^* := E_\theta \left[X_{T+1} \middle| X_{-\infty}^T \right] := \sum_{i=0}^{k_T} r_i(X_{T-i}, \theta) + \eta_{k_T}(\mathbb{X}, \theta)$$

où chaque fonction r_i représente la contribution (additive) de la X_{T-i} valeur, et où $\eta_{k_T}(\mathbb{X}, \theta)$ est une fonction de carré intégrable asymptotiquement négligeable dans un sens à préciser. Ce modèle additif est une extension d'un cas particulier étudié par Bosq et

Blanke [24] et Bosq [23]. Pour plus de détails concernant les modèles additifs, on pourra se référer à, par exemple, Härdle et al. [71].

Dans le cas étudié par Bosq et Blanke [24] chapitre 2 et Bosq [23], le prédicteur probabiliste a pour structure

$$X^*_{T+h} := r_{T,h}(Y_T, \theta)$$

où Y_T est une variable $\sigma(X^T_{T-k_T})$ mesurable, telle que $0 \leq k_T < T$ et $\lim_{T\to\infty} k_T/T = 0$, i.e. qui représente le proche passé. Le prédicteur statistique est alors construit à partir d'un estimateur $\hat{\theta}_{\varphi(T)}$ du paramètre,

$$\hat{X}_{T+h} = r_{T+h}(\hat{\theta}_{\varphi(T)}, Y_T)$$

avec $0 < \varphi(T) < T$, $T - k_T - \varphi(T) \to \infty$ et $\varphi(T)/T \to 1$. La consistance, la vitesse de convergence et la loi limite du prédicteur statistique sont alors obtenues.

Dans cet article, on suppose qu'on dispose d'un estimateur consistant $\hat{\theta}_T$ de θ. En estimant θ par $\hat{\theta}_{\varphi(T)}$ sur l'intervalle $[0, \varphi(T)]$ avec $\varphi(T) \to \infty$, on construit alors le prédicteur statistique

$$\hat{X}_{t+1} = r_{\hat{\theta}_{\varphi(T)}}(X^T_{T-k_T})$$

Cette étude a pour but de montrer que si le processus est mélangeant, alors on peut séparer le problème d'estimation de θ sur $[0, \varphi(T)]$, du problème probabiliste sur la "mémoire" du processus entre $[T - k_T, T]$.

Plus précisément, on montrera dans la section 2 la consistance du prédicteur, i.e. la convergence vers 0 de l'erreur statistique de prévision, avant de montrer un exemple d'application inspiré par la décomposition de Wold dans la section 3, pour finir par l'étude de la loi asymptotique du prédicteur statistique dans la section 4.

2.1.3 Discussion sur le modèle

On a dit que le processus est approximativement k_T markovien, ce qui revient à considérer que k_T est imposé par le processus. On peut aussi considérer le modèle ci-dessus comme

un modèle additif non-linéaire généralisé au sens où le processus vérifie

$$X_{T+1} = \sum_{i=0}^{+\infty} r_i(X_{T-i}; \theta) + \varepsilon_{T+1}$$

où l'innovation (ε) est telle que $E[\varepsilon_{T+1}|X_{-\infty}^{T}] = 0$ et où la convergence de la série est à comprendre au sens de la convergence en moyenne quadratique. Une condition pour la convergence de cette série est donnée dans le corollaire 3.1 de Rio [111] (le premier corollaire 3.1 p.51), rappelé dans le lemme ci-dessous.

Lemma 2.2 (Rio) *Soit $(Y_i)_{i\in\mathbb{N}}$ une suite de variables réelles centrées de variance finie. Soit Q_i est la fonction de quantile de $|Y_i|$ i.e. l'inverse généralisé continu à droite de la fonction $H_{Y_i}(t) = P(|Y_i| > t)$, et $\alpha(y) = \alpha[y]$ où $[y]$ désigne la partie entière de y et $\alpha(k)$ le coefficient de mélange fort de Rosenblatt [112] (voir ci-dessous). Alors la série $\sum_{i=1}^{\infty} Y_i$ converge p.s. si la condition suivante est réalisée :*

$$\sum_{i=1}^{\infty} \int_0^1 \alpha^{-1}(u) Q_i^2(u) du < +\infty$$

On notera que ce lemme généralise le théorème des deux séries de Kolmogorov qui traite du cas i.i.d. et requiert la convergence des moments d'ordre 1 et 2.

La convergence de la série $\sum_{i=0}^{+\infty} r_i(X_{T-i}; \theta)$ entraîne à son tour que

$$\sum_{i=k}^{+\infty} r_i(X_{T-i}; \theta) \overset{p.s.}{\to} 0$$

pour $k \to +\infty$. En posant $\eta_{k_T} = \sum_{i=k_T+1}^{+\infty} r_i(X_{T-i}; \theta)$, l'écriture

$$X_{T+1}^{*} := E_\theta \left[X_{T+1} | X_{-\infty}^{T} \right] := \sum_{i=0}^{k_T} r_i(X_{T-i}, \theta) + \eta_{k_T}(\mathbb{X}, \theta)$$

avec l'introduction de $\eta_{k_T}(\mathbb{X}, \theta)$ peut se comprendre comme un choix du statisticien de prendre un $k_T \to +\infty$ de façon à rendre la contribution du passé lointain dans la prévision négligeable, i.e. à avoir $\eta_{k_T}(\mathbb{X}, \theta) \to 0$ pour $T \to +\infty$.

2.2 Consistance du prédicteur statistique

On rappelle la notion de α-mélangeance (cf. Rosenblatt [112]):

Definition 2.3 (Rosenblatt) *Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé et $\mathcal{B}$, $\mathcal{C}$ deux sous-tribus de $\mathcal{A}$. On définit le coefficient de α-mélange entre les deux tribus $\mathcal{B}$, $\mathcal{C}$ par*

$$\alpha(\mathcal{B},\mathcal{C}) = \sup_{\substack{B\in\mathcal{B}\\ C\in\mathcal{C}}} |P(B\cap C) - P(B)P(C)|$$

et le coefficient de α-mélange d'ordre k pour le processus $\mathbb{X} = \{X_t, t \in \mathbb{N}\}$ défini sur l'espace probabilisé $(\Omega, \mathcal{A}, P)$ par

$$\alpha(k) = \sup_{t\in\mathbb{N}} \alpha(\sigma(X_s, s \le t), \sigma(X_s, s \ge t+k))$$

On rappelle en outre l'inégalité de Davydov (cf. Bosq [21], p. 21) : Notons $\sigma(X)$ la σ-algèbre des événements engendrés par la variable X et $\|X\|_q = \left\{\mathbb{E}(X^q)\right\}^{1/q}$ pour $1 \le q \le \infty$.

Lemma 2.4 (Davydov) *Soient $X \in L^q(\mathbb{P})$ et $Y \in L^r(\mathbb{P})$, si $q > 1$, $r > 1$ et $\frac{1}{r}+\frac{1}{q} = 1-\frac{1}{p}$, alors*

$$|Cov(X,Y)| \le 2p\big(2\alpha(\sigma(X),\sigma(Y))\big)^{\frac{1}{p}}\|X\|_q\|Y\|_r.$$

On se place donc dans le cadre suivant :

- Le processus $\mathbb{X}$ est du second ordre, faiblement stationnaire, α mélangeant.
- On suppose que le prédicteur probabiliste s'écrit :

$$X^*_{T+1} := E_\theta\left[X_{T+1}|X^T_{-\infty}\right] = \sum_{i=0}^{k_T} r_i(X_{T-i},\theta) + \eta_{k_T}(\mathbb{X},\theta)$$

On effectue alors les hypothèses suivantes :

Hypothèse $\mathbf{H_0}$ sur le processus $\mathbb{X}$

(i) $\lim_{T\to\infty} E_\theta(\eta^2_{k_T}(\mathbb{X},\theta)) = 0$;

(ii) pour tout $i \in \mathbb{N}$, $\|r_i(X_{T-i}, \theta_1) - r_i(X_{T-i}, \theta_2)\| \leq H_i(X_{T-i}) \|\theta_1 - \theta_2\|$, $\forall \theta_1, \theta_2$;

(iii) il existe $r > 1$ tel que $\sup_{i \in \mathbb{N}} E_\theta H_i^{2r}(X_{T-i}) < \infty$.

Hypothèse $\mathbf{H_1}$ sur l'estimateur $\hat{\theta}_T$ On suppose qu'on dispose d'un estimateur consistant $\hat{\theta}_T$ de θ à la vitesse (paramétrique) T.

(i) $\limsup_{T \to \infty} T.E_\theta(\hat{\theta}_T - \theta)^2 < \infty$;

(ii) il existe $q > 1$ tel que $\limsup_{T \to \infty} T^q E(\hat{\theta}_T - \theta)^{2q} < \infty$.

Hypothèse $\mathbf{H_2}$ sur les coefficients

(i) $\frac{k_T^2}{\varphi(T)} \underset{T \to \infty}{\to} 0$;

(ii) $(T - k_T - \varphi(T)) \underset{T \to \infty}{\to} \infty$.

On est alors en mesure de formuler le théorème suivant :

Theorem 2.5 *Si les hypothèses* $\mathbf{H_0}$,$\mathbf{H_1}$,$\mathbf{H_2}$ *sont vérifiées, alors*

$$\limsup_{T \to \infty} E_\theta(\hat{X}_{T+1} - X^*_{T+1})^2 = 0$$

Proof

$$\begin{aligned}
& E_\theta(X^*_{T+1} - \hat{X}_{T+1})^2 \\
= \; & E_\theta \left(\sum_{i=0}^{k_T} (r_i(X_{T-i}, \theta) - r_i(X_{T-i}, \hat{\theta}_{\varphi(T)})) + \eta_{k_T}(\mathbb{X}, \theta) \right)^2 \\
\leq \; & 2E_\theta(\eta^2_{k_T}(\mathbb{X}, \theta)) + 2E_\theta \left(\sum_{i=0}^{k_T} (r_i(X_{T-i}, \theta) - r_i(X_{T-i}, \hat{\theta}_{\varphi(T)})) \right)^2 \\
\leq \; & 2E_\theta(\eta^2_{k_T}(\mathbb{X}, \theta)) + 2(k_T + 1) \sum_{i=0}^{k_T} E_\theta \left(r_i(X_{T-i}, \theta) - r_i(X_{T-i}, \hat{\theta}_{\varphi(T)}) \right)^2 \\
\leq \; & 2E_\theta(\eta^2_{k_T}(\mathbb{X}, \theta)) + 2(k_T + 1) \sum_{i=0}^{k_T} E_\theta \left(H_i(X_{T-i}) \left\| \hat{\theta}_{\varphi(T)} - \theta \right\| \right)^2
\end{aligned}$$

par application de l'hypothèse $\boldsymbol{H}_0$ *(ii), d'où*

$$\begin{aligned} E_\theta(X^*_{T+1} - \hat{X}_{T+1})^2 &\leq 2E_\theta(\eta^2_{k_T}(\mathbb{X}, \theta)) \\ &+ 2(k_T+1)\sum_{i=0}^{k_T} E_\theta H_i^2(X_{T-i}) E_\theta \left\|\hat{\theta}_{\varphi(T)} - \theta\right\|^2 \\ &+ 2(k_T+1)\sum_{i=0}^{k_T} \delta_{i,T} \\ &\leq I_1 + I_2 + I_3 \end{aligned}$$

où on a appliqué l'inégalité du lemme 2.4 avec $X = H_i^2(X_{T-i})$ *et* $Y = \left\|\hat{\theta}_{\varphi(T)} - \theta\right\|^2$*, et où on a posé*

$$\begin{aligned} I_1 &= 2E_\theta(\eta^2_{k_T}(\mathbb{X}, \theta)) \\ I_2 &= 2(k_T+1)\sum_{i=0}^{k_T} E_\theta H_i^2(X_{T-i}) E_\theta \left\|\hat{\theta}_{\varphi(T)} - \theta\right\|^2 \\ I_3 &= 2(k_T+1)\sum_{i=0}^{k_T} \delta_{i,T} \end{aligned}$$

avec

$$\delta_{i,T} = 2p\left(2\alpha(T-i-\varphi(T))\right)^{1/p}\left(E_\theta H_i^{2r}(X_{T-i})\right)^{1/r}\left(E_\theta\left\|\hat{\theta}_{\varphi(T)} - \theta\right\|^{2q}\right)^{1/q}$$

Par l'hypothèse $\boldsymbol{H}_0$ *(i),* $\lim_{T\to\infty} I_1 = 0$*.*

Par l'hypothèse $\boldsymbol{H}_1$ *(i),* $\limsup_{T\to\infty} \varphi(T).E_\theta(\hat{\theta}_{\varphi(T)} - \theta)^2 < \infty$ *Donc,*

$$\limsup_{T\to\infty} I_2 \leq \limsup_{T\to\infty} \frac{k_T}{\varphi(T)} \cdot \sum_{i=0}^{k_T} E_\theta H_i^2(X_{T-i})$$

Or par $\boldsymbol{H}_1$ *(iii), il existe* $r > 1$ *tel que* $\sup_{i\in\mathbb{N}} E_\theta H_i^{2r}(X_{T-i}) < \infty$*, donc*

$$\limsup_{T\to\infty} I_2 \leq \limsup_{T\to\infty} \frac{k_T^2}{\varphi(T)}.$$

$$I_3 = 4pk_T\left(E_\theta\left\|\hat{\theta}_{\varphi(T)} - \theta\right\|^{2q}\right)^{\frac{1}{q}} \sum_{i=0}^{k_T}\left(2\alpha(T-i-\varphi(T))\right)^{1/p}\left(E_\theta H_i^{2r}(X_{T-i})\right)^{\frac{1}{r}}$$

Par les hypothèses $\boldsymbol{H}_0$ *(iii) et* $\boldsymbol{H}_1$ *(ii), on a*

$$\sup_{i\in\mathbb{N}} \left(E_\theta H_i^{2r}(X_{T-i})\right)^{1/r} < \infty$$
$$\limsup_{T\to\infty} \varphi(T)^q E(\hat{\theta}_{\varphi(T)} - \theta)^{2q} < \infty$$

Donc,

$$\limsup_{T\to\infty} I_3 \leqslant 4p. \limsup_{T\to\infty} \frac{k_T}{\varphi(T)} \sum_{i=0}^{k_T} (2\alpha(T-i-\varphi(T)))^{1/p}$$

Comme $\mathbb{X}$ *est* α*-mélangeant,* $\alpha(k)$ *est décroissant lorsque* k *croît, donc les coefficients de la somme ci dessus sont majorés par* $\alpha(T-k_T-\varphi(T))$*, donc*

$$\limsup_{T\to\infty} I_3 \leq \limsup_{T\to\infty} \frac{k_T^2}{\varphi(T)} \alpha^{1/p}(T-k_T-\varphi(T))$$

Les conditions $\boldsymbol{H}_2$ *(i), et* $\boldsymbol{H}_2$ *(ii) sur les coefficients assurent alors que*

$$\limsup_{T\to\infty} I_3 = \limsup_{T\to\infty} I_2 = 0.$$

Dans le cadre d'une prévision à "mémoire fixe", i.e. où $k_T = k$, les hypothèses se simplifient et on peut formuler le corollaire suivant:

Corollary 2.6 *On suppose que le prédicteur probabiliste s'écrit :*

$$X_{T+1}^* := E_\theta \left[X_{T+1} \left| X_{-\infty}^T\right.\right] = \sum_{i=0}^{k} r_i(X_{T-i},\theta) + \eta_T(\mathbb{X},\theta)$$

On garde les hypothèses $\mathbf{H_0}$,$\mathbf{H_1}$,$\mathbf{H_2}$ *que l'on modifie de la façon suivante:*

- *pour l'hypothèse* $\mathbf{H_0}$*(i) sur le processus* $\mathbb{X}$ *par* $\lim_{T\to\infty} E_\theta(\eta_T^2(\mathbb{X},\theta)) = 0$ *;*
- *pour les hypothèses* $\mathbf{H_2}$ *sur les coefficients par*
  - $\mathbf{H_2}$*(i)* $\varphi(T) \to \infty$*;*
  - $\mathbf{H_2}$*(ii)* $T - \varphi(T) \underset{T\to\infty}{\to} \infty$.

Alors, sous ces nouvelles hypothèses, $\limsup_{T\to\infty} E_\theta(\hat{X}_{T+1} - X_{T+1}^*)^2 = 0.$

2.3 Exemple d'application

Dans cette section on explicite les hypothèses du théorème 2.5 en basant notre discussion sur la décomposition de Wold d'un processus (cf. par exemple Brockwell & Davis [29]): pour un processus linéaire à temps discret, faiblement stationnaire, centré, purement non déterministe et inversible, sa décomposition peut s'écrire sous la forme suivante:

$$X_T = e_T + \sum_{i=1}^{k_T} \varphi_i(\theta) X_{T-i} + \sum_{i>k_T} \varphi_i(\theta) X_{T-i}$$

avec $\sum_{i=1}^{\infty} \varphi_i^2(\theta) < \infty$. On a alors

$$\begin{aligned} X_{T+1}^* &:= E\left[X_{T+1} \left| X_{-\infty}^T \right.\right] \\ &= \sum_{i=1}^{k_T+1} \varphi_i(\theta) X_{T+1-i} + \sum_{i>k_T+1} \varphi_i(\theta) X_{T+1-i} + E\left[e_{T+1} \left| X_{-\infty}^T \right.\right] \\ &:= \sum_{i=0}^{k_T} r_i(X_{T-i}, \theta) + \eta_{k_T}(\mathbb{X}, \theta) \end{aligned}$$

avec

$$r_i(X_{T-i}, \theta) := \varphi_{i+1}(\theta) X_{T-i}$$

$$\eta_{k_T}(\mathbb{X}, \theta) := \sum_{i>k_T} \varphi_{i+1}(\theta) X_{T-i} + E[e_{T+1} | X_{-\infty}^T]$$

Si pour tout i, φ_i est dérivable et $\varphi_i'(.)$ est borné, on peut écrire pour tous θ_1 et θ_2,

$$\begin{aligned} \|r_i(X_{T-i}, \theta_1) - r_i(X_{T-i}, \theta_2)\| &= \|(\varphi_{i+1}(\theta_1) - \varphi_{i+1}(\theta_2)) X_{T-i}\| \\ &\leqslant \left\|\varphi_{i+1}'(.)\right\|_\infty \|\theta_1 - \theta_2\| \, \|X_{T-i}\| \end{aligned}$$

et les hypothèses $\mathbf{H}_0$ (ii), (iii), sont vérifiées avec $H_i = id$. $\mathbf{H}_0$ (iv) est vérifiée si X admet un moment d'ordre $2r$, pour un $r > 1$. Si $E\left[e_{T+1} \left| X_{-\infty}^T \right.\right] = 0$,

$$E\eta_{k_T}^2(\mathbb{X}, \theta) = \sum_{i>k_T} \sum_{j>k_T} \varphi_{i+1}(\theta) \varphi_{j+1}(\theta) \operatorname{cov}(X_{T-i}, X_{T-j})$$

en utilisant l'inégalité du lemme 2.4,

$$E\eta_{k_T}^2(\mathbb{X},\theta) \leqslant 2^{(2p+1)/p} p \left\|X_0\right\|_q \left\|X_0\right\|_r \sum_{i,j>k_T} \varphi_{i+1}(\theta)\varphi_{j+1}(\theta)\alpha^{1/p}(|i-j+1|)$$

et la condition,

$$\sum_{i,j} \varphi_{i+1}(\theta)\varphi_{j+1}(\theta)\alpha^{1/p}\left(|i-j+1|\right) < \infty$$

assure la validité de l'hypothèse $\mathbf{H}_0$ (i).

On a donc montré la proposition suivante :

Proposition 2.7 *Si* $\mathbb{X}$ *vérifie les conditions*

1. $\forall i$, φ_i *est dérivable et* $\left\|\varphi_i'(.)\right\|_\infty < \infty$;
2. *il existe un* $r > 1$ *tel que* (X_t) *admet un moment d'ordre* $2r$*;*
3. $E\left[e_{T+1}\left|X_{-\infty}^T\right.\right] = 0$*;*
4. $\mathbb{X}$ *est* α*-mélangeant vérifie* $\sum_{i,j} \varphi_{i+1}(\theta)\varphi_{j+1}(\theta)\alpha^{1/p}\left(|i-j|\right) < \infty$.

Alors $\mathbb{X}$ *vérifie les conditions du théorème 2.5.*

Dans le cas encore plus particulier d'un processus auto-régressif d'ordre p, le corollaire 2.6 s'applique avec les conditions suivantes:

Corollary 2.8 *Si* $\mathbb{X}$ *vérifie les conditions*

1. $\forall i$, φ_i *est dérivable et* $\left\|\varphi_i'(.)\right\|_\infty < \infty$;
2. *il existe un* $r > 1$ *tel que* (X_t) *admet un moment d'ordre* $2r$*;*
3. $E\left[e_{T+1}\left|X_{-\infty}^T\right.\right] = 0$*;*
4. $\mathbb{X}$ *est* α*-mélangeant.*

Alors $\mathbb{X}$ *vérifie les conditions du corollaire 2.6.*

2.4 Loi limite du prédicteur statistique

2.4.1 Un lemme d'indépendance asymptotique

On note $\overset{d}{\rightsquigarrow}$ la convergence en loi. On établit tout d'abord une conséquence de la mélangeance sur la convergence en loi des vecteurs aléatoires par le lemme suivant :

Lemma 2.9 *Soit (X'_n) et (X''_n) deux suites de variables aléatoires réelles de lois respectives P'_n et P''_n définies sur l'espace probabilisé $(\Omega, \mathcal{A}, P)$. On suppose que (X'_n) et (X''_n) sont asymptotiquement deux à deux mélangeants au sens où il existe une suite de coefficients $\alpha(n)$ avec $\alpha(n) \underset{n\to\infty}{\to} 0$ tels que, pour tous boréliens A et B de $\mathcal{R}$,*

$$|P(X'_n \in A, X''_n \in B) - P(X'_n \in A)P(X''_n \in B)| \leq \alpha(n)$$

Alors, si

1. *$X'_n \overset{d}{\rightsquigarrow} X'$ de loi P';*

2. *$X''_n \overset{d}{\rightsquigarrow} X''$ de loi P'';*

$(X'_n, X''_n) \overset{d}{\rightsquigarrow} (X', X'')$, et la loi de (X', X'') est $P' \otimes P''$.

Proof *D'après Billingsley [17] Théorème 2.1, $P_n \to P \Leftrightarrow \lim_n P_n(C) = P(C)$ pour tout C, ensemble de continuité de P, i.e. pour tout ensemble C tel que $P(\partial C) = 0$ où ∂C représente la frontière de C.*
$X'_n \overset{d}{\rightsquigarrow} X'$ donc $\forall A$ ensemble de continuité de la loi de X', $P(X'_n \in A) \to P(X' \in A)$.
$X''_n \overset{d}{\rightsquigarrow} X''$ donc $\forall B$ ensemble de continuité de la loi de X'', $P(X''_n \in B) \to P(X'' \in B)$.
Donc quel que soit $\varepsilon > 0$, il existe un rang n_0 tel que pour tout $n > n_0$, on ait

$$|P(X'_n \in A)P(X''_n \in B) - P(X' \in A)P(X'' \in B)| \leq \varepsilon$$

Par ailleurs, comme X'_n et X''_n sont deux à deux α-mélangeants, on a

$$|P(X'_n \in A, X''_n \in B) - P(X'_n \in A)P(X''_n \in B)| \leq \alpha(n)$$

Donc, pour tout $\varepsilon > 0$,

$$\begin{aligned} & |P(X'_n \in A, X''_n \in B) - P(X' \in A)P(X'' \in B)| \\ \leq\ & |P(X'_n \in A, X''_n \in B) - P(X'_n \in A)P(X''_n \in B)| \\ +\ & |P(X'_n \in A)P(X''_n \in B) - P(X' \in A)P(X'' \in B)| \\ \leq\ & \alpha(n) + \varepsilon \\ \leq\ & 2\varepsilon \text{ pour } n \text{ assez grand} \end{aligned}$$

D'après Billingsley [17] Théorème 3.1, $(X'_n, X''_n) \overset{d}{\rightsquigarrow} (X', X'')$ *si et seulement si pour tout* A*, ensemble de continuité de la loi de* X'*, pour tout* B *ensemble de continuité de la loi de* X''*, on a* $P(X'_n \in A, X''_n \in B) \to P(X' \in A, X'' \in B)$ *quand* $n \to \infty$.
Donc $(X'_n, X''_n) \overset{d}{\rightsquigarrow} (X', X'')$ *et la loi limite de* (X', X'') *est* $P' \otimes P''$.

2.4.2 Convergence en loi

On notera par la suite ∂_θ l'opération de dérivation par rapport à θ. On reprend le cadre de la section 2, en introduisant des hypothèses supplémentaires. Pour plus de clarté, nous écrivons toutes les hypothèses nécessaires:

Hypothèses sur le processus $\mathbb{X}$

(i) $\mathbb{X}$ est un processus du second ordre, faiblement stationnaire, α mélangeant.

(ii) On suppose que le prédicteur probabiliste s'écrit :

$$X^*_{T+1} := E_\theta \left[X_{T+1} \left| X_{-\infty}^T \right. \right] = \sum_{i=0}^{k_T} r_i(X_{T-i}, \theta) + \eta_{k_T}(\mathbb{X}, \theta)$$

Hypothèse $\mathbf{H}'_0$

(i) $\theta \to r_i(X_{T-i}, \theta)$ est deux fois différentiable en θ;

(ii) la norme infinie de $\partial^2_\theta r_i(X_{T-i}, .)$ est borné en probabilité uniformément en i, i.e.

$\partial^2_\theta r_i(X_{T-i},.)$ est uniformément tendue :

$$\sup_i \left\|\partial^2_\theta r_i(X_{T-i},.)\right\|_\infty = O_P(1);$$

(iii) $\eta_{k_T}(\mathbb{X},\theta) = o_P\left(\sqrt{\frac{1}{\varphi(T)}}\right)$;

(iv) $\sum_{i=0}^{+\infty} \partial_\theta r_i(X_{T-i};\theta)$ existe et converge presque sûrement vers un vecteur V lorsque $T \to +\infty$.

Hypothèse $\mathbf{H}'_1$ sur l'estimateur $\hat{\theta}_T$

(i) $\sqrt{T}(\hat{\theta}_T - \theta) \overset{d}{\rightsquigarrow} N(0,\sigma^2(\theta))$.

Hypothèse $\mathbf{H}'_2$ sur les coefficients

(i) $k_T = o(\sqrt{\varphi(T)})$;

(ii) $(T - k_T - \varphi(T)) \underset{T\to\infty}{\to} \infty$.

On a alors le théorème suivant :

Theorem 2.10 *Si les hypothèses $\mathbf{H}'_0$,$\mathbf{H}'_1$,$\mathbf{H}'_2$ sont vérifiées, alors*

$$\sqrt{\varphi(T)}(\hat{X}_{T+1} - X^*_{T+1}) \overset{d}{\rightsquigarrow} < U, V >$$

où U et V sont deux variables indépendantes, U de loi $\mathcal{N}(0,\sigma^2(\theta))$ et V la loi limite de $\sum_{i=0}^{+\infty} \partial_\theta r_i(X_{T-i};\theta)$.

Proof *Par l'hypothèse $\mathbf{H}'_0$(i), il existe une suite de variables $(\xi_{T,i})$ qui soient $\sigma(X_0^{T-i})$ mesurable et incluses dans un voisinage de θ de sorte que*

$$\begin{aligned} r_i(X_{T-i};\hat{\theta}_{\phi(T)}) - r_i(X_{T-i};\theta) &= (\hat{\theta}_{\phi(T)} - \theta)^t.\partial_\theta r_i(X_{T-i};\theta) \\ &+ (\hat{\theta}_{\phi(T)} - \theta)^t \frac{\partial^2_\theta r_i(X_{T-i};\xi_{T,i})}{2} (\hat{\theta}_{\phi(T)} - \theta) \end{aligned}$$

où $(\hat{\theta}_{\phi(T)} - \theta)^t$ *désigne la transposée du vecteur* $(\hat{\theta}_{\phi(T)} - \theta)$ *et* $\partial_\theta r_i(X_{T-i};\theta)$ *le gradient par rapport à* θ *de* $r_i(X_{T-i};\theta)$. *Donc,*

$$\begin{aligned}
\hat{X}_{T+1} - X^*_{T+1} &= \sum_{i=0}^{k_T} (r_i(X_{T-i}, \hat{\theta}_{\varphi(T)}) - r_i(X_{T-i},\theta)) - \eta_{k_T}(\mathbb{X},\theta) \\
&= \sum_{i=0}^{k_T} (\hat{\theta}_{\phi(T)} - \theta)^t . \partial_\theta r_i(X_{T-i},\theta) \\
&+ \sum_{i=0}^{k_T} (\hat{\theta}_{\phi(T)} - \theta)^t \frac{\partial^2_\theta r_i(X_{T-i};\xi_{T,i})}{2} (\hat{\theta}_{\phi(T)} - \theta) \\
&- \eta_{k_T}(\mathbb{X},\theta)
\end{aligned}$$

Donc,

$$\begin{aligned}
\sqrt{\varphi(T)}(\hat{X}_{T+1} - X^*_{T+1}) &= \sqrt{\varphi(T)}(\hat{\theta}_{\varphi(T)} - \theta)^t \sum_{i=0}^{k_T} (\partial_\theta r_i(X_{T-i},\theta)) \\
&+ \sqrt{\varphi(T)} \frac{(\hat{\theta}_{\varphi(T)} - \theta)^t}{2} \sum_{i=0}^{k_T} \partial^2_\theta r_i(X_{T-i};\xi_{T,i})(\hat{\theta}_{\varphi(T)} - \theta) \\
&- \sqrt{\varphi(T)} \eta_{k_T}(\mathbb{X},\theta) \\
&:= I_1 + I_2 - I_3
\end{aligned}$$

Etude de I_2.

Par l'hypothèse $\mathbf{H}'_1(i)$, $\sqrt{\varphi(T)}(\hat{\theta}_{\varphi(T)} - \theta) \overset{d}{\rightsquigarrow} N(0,\sigma^2(\theta))$. *Donc* $||\hat{\theta}_{\varphi(T)} - \theta||^2 = O_P(1/\varphi(T))$. *En rajoutant la condition* $\mathbf{H}'_0(ii)$, $\sup_i \partial^2_\theta r_i(X_{T-i},\xi_{T,i}) = O_P(1)$, *on obtient*

$$I_2 = \sqrt{\varphi(T)} O_P\left(\frac{1}{\varphi(T)}\right) \sum_{i=0}^{k_T} O_P(1) = O_P\left(\frac{k_T}{\sqrt{\varphi(T)}}\right)$$

Etude de I_3.

$I_3 = \sqrt{\varphi(T)}\eta_{k_T}(\mathbb{X},\theta)$. *Si on suppose, par l'hypothèse* $\mathbf{H}'_0(iii)$, *que*

$$\eta_{k_T}(\mathbb{X},\theta) = o_P\left(\sqrt{\frac{1}{\varphi(T)}}\right)$$

alors $I_3 = o_P(1)$.

Etude de I_1.

On pose $I_1 =< U_T, V_T >$ *avec*

$$U_T = \sqrt{\varphi(T)}(\hat{\theta}_{\varphi(T)} - \theta)$$

$$V_T = \sum_{i=0}^{k_T} \partial_\theta r_i(X_{T-i}, \theta)$$

U_T *est* $\sigma(X_0^{\varphi(T)})$ *mesurable et* V_T *est* $\sigma(X_{T-k_T}^T)$ *mesurable donc,* U_T *et* V_T *ont pour coefficient de* α*-mélange,* $\alpha(T - k_T - \varphi(T))$*, et sont donc deux à deux mélangeants pour* $T - k_T - \varphi(T) \to \infty$ *(hypothèse* $\mathbf{H}_2'$ *(ii). Par le lemme 2.9, il suffit donc d'étudier la convergence en loi de* U_T *et* V_T *séparément pour en déduire la convergence de* (U_T, V_T) *et par continuité celle de* $U_T.V_T$*.*

Or, par hypothèse, $U_T = \sqrt{\varphi(T)}(\hat{\theta}_{\varphi(T)} - \theta) \overset{d}{\rightsquigarrow} \mathcal{N}(0, \sigma^2(\theta)) := U$ *Pour* V_T*, la condition* $\mathbf{H}_0'$*(iv), assure l'existence d'un vecteur*

$$V := \sum_{i=0}^{+\infty} \partial_\theta r_i(X_{T-i}; \theta)$$

tel que

$$\sum_{i=0}^{T} \partial_\theta r_i(X_{T-i}; \theta) \overset{\mathcal{L}^2}{\to} V$$

Donc pour la sous suite k_T*,* $V_T \to V$ *en moyenne quadratique . Donc* $V_T \overset{d}{\rightsquigarrow} V$ *en loi.*

En utilisant le lemme 2.9 précité, on obtient alors

$$< U_T, V_T > \overset{d}{\rightsquigarrow} < U, V >$$

On a donc $I_1 + I_2 - I_3 =< U_T, V_T > + O_P\left(\frac{k_T}{\sqrt{\varphi(T)}}\right) + o_P(1)$*. Si* $k_T = o(\sqrt{\varphi}(T))$ *(hypothèse* $\mathbf{H}_2'$ *(i),* $I_1 + I_2 - I_3 =< U_T, V_T > + o_P(1)$*. Par le lemme de Slutsky, on en déduit que*

$$\sqrt{\varphi(T)}(\hat{X}_{T+1} - X_{T+1}^*) \overset{d}{\rightsquigarrow} < U, V >$$

avec U *de loi* $\mathcal{N}(0, \sigma^2(\theta))$ *.*

2.4.3 Discussion

Les conditions sur les coefficients k_T et $\varphi(T)$ donnent une indication du compromis à effectuer entre l'approximation probabiliste et l'estimation statistique. En effet, pour

$$\eta_{k_T} = \sum_{i=k_T+1}^{+\infty} \partial_\theta r_i(X_{T-i};\theta)$$

On a par Minkowski,

$$\|\eta_{k_T}\|_2 \leqslant \sum_{i=k_T}^{+\infty} \|\partial_\theta r_i(X_{T-i};\theta)\|_2$$

Si par exemple

$$\|\partial_\theta r_i(X_{T-i};\theta)\|_2 \sim \frac{1}{i^a}$$

pour un $a > 1$, alors

$$\|\eta_{k_T}\|_2 \sim \sum_{i=k_T}^{+\infty} \frac{1}{i^a} \sim \frac{k_T^{1-a}}{1-a}$$

et la condition $\mathbf{H}_0'$(iii) est vérifiée pour

$$\frac{\sqrt{\varphi(T)}}{k_T^{a-1}} = o(1)$$

On doit effectuer un compromis pour rendre compatibles cette dernière condition et la condition $\mathbf{H}_2'$(i) : $k_T = o(\sqrt{\varphi(T)})$.

Remarque : on aurait pu généraliser ce résultat à d'autres vitesses d'estimation que $\sqrt{T}$ mais on a préféré ne pas présenter un cas plus général pour des raisons de lisibilité.

2.4.4 Cas de la mémoire fixe

Lorsque $k_T = k$ est fixe, le théorème précédent donne le corollaire suivant :

Corollary 2.11 *On suppose que le prédicteur probabiliste s'écrit :*

$$X_{T+1}^* := E_\theta\left[X_{T+1}\left|X_{-\infty}^T\right.\right] = \sum_{i=0}^{k} r_i(X_{T-i},\theta) + \eta_T(\mathbb{X},\theta)$$

sous les hypothèses modifiées suivantes

Hypothèse $\mathbf{H}_0'$

(i) $\theta \to r_i(X_{T-i}, \theta)$ *est deux fois différentiable en* θ*;*

(ii) la norme infinie de $\partial_\theta^2 r_i(X_{T-i}, .)$ *est borné en probabilité uniformément en* i*, i.e.* $\partial_\theta^2 r_i(X_{T-i}, .)$ *est uniformément tendue :*

$$\sup_i \left\| \partial_\theta^2 r_i(X_{T-i}, .) \right\|_\infty = O_P(1);$$

(iii) $\eta_T(\mathbb{X}, \theta) = o_P\left(\sqrt{\frac{1}{\varphi(T)}}\right)$*;*

(iv) $\sum_{i=0}^{k} \partial_\theta r_i(X_{T-i}; \theta)$ *converge en loi vers un vecteur* V *lorsque* $T \to +\infty$*.*

Hypothèse $\mathbf{H}_1'$ sur l'estimateur $\hat{\theta}_T$

(i) $\sqrt{T}(\hat{\theta}_T - \theta) \overset{d}{\rightsquigarrow} N(0, \sigma^2(\theta))$*.*

Hypothèse $\mathbf{H}_2'$ sur les coefficients

(i) $(T - \varphi(T)) \underset{T\to\infty}{\to} \infty$*.*

$$\sqrt{\varphi(T)}(\hat{X}_{T+1} - X_{T+1}^*) \overset{d}{\rightsquigarrow} < U, V >$$

où U *et* V *sont deux variables indépendantes.*

2.4.5 Application à la prévision d'un AR(l)

Un exemple important dans les applications est celui d'un processus linéaire autorégressif d'ordre l,

$$X_t = \sum_{i=1}^{l} \theta_i X_{t-i} + \varepsilon_t$$

où (ε_t) est une suite i.i.d centrée. On pose $P(z) = z^l - \sum_{i=1}^{l} \theta_i z^{l-i}$. D'après Doukhan[46], théorème 6 et corollaire 3 de la section 2.4.1.2, si le polynôme P a toutes ses racines

à l'intérieur du disque unité ouvert et si la distribution marginale commune des (ε_t) est dominée par la mesure de Lebesgue, le processus (X_t) est stationnaire et géométriquement ergodique donc géométriquement mélangeant. Il existe donc $a > 0$ et $b > 0$ tel que le coefficient de α-mélange vérifie $\alpha(u) \leq a \exp(-bu)$, pour $u \geq 0$.

D'après Box et Jenkins [28], la stationnarité du processus entraîne que les coefficients θ satisfont les équations de Yule-Walker, à savoir $\gamma(k) = \theta_1\gamma(k-1) + \ldots + \theta_l\gamma(k-l)$ pour $k > 0$, où on a noté $\gamma(k) = Cov(X_0, X_k)$ les autocovariances. On considère alors l'estimateur sans biais des autocovariances $\hat{\gamma}_T = (\hat{\gamma}_T(1), \ldots, \hat{\gamma}_T(l))$ défini par

$$\hat{\gamma}_T(i) = \frac{1}{T-i} \sum_{t=1}^{T-i} X_t X_{t+i}$$

pour $i = 1, \ldots, l$ et $T > i$. En inversant l'équation de Yule-Walker, on obtient un estimateur de θ:

$$\hat{\theta}_T = \hat{\Gamma}_{l-1}^{-1} \hat{\gamma}_T$$

où $\hat{\gamma}_T$ est le vecteur $(\hat{\gamma}_T^1, \ldots, \hat{\gamma}_T^l)$ et $\hat{\Gamma}_{l-1}$ la matrice

$$\begin{bmatrix} 1 & \hat{\gamma}_T(1) & \cdots & \hat{\gamma}_T(l-1) \\ \hat{\gamma}_T(1) & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \hat{\gamma}_T(1) \\ \hat{\gamma}_T(l-1) & \ldots & \hat{\gamma}_T(1) & 1 \end{bmatrix}$$

On a alors, cf. Gouriéroux et Monfort [63] propriété 9.7 et 9.8, Si $\mathbb{X}$ est un processus gaussien, stationnaire admettant une représentation moyenne mobile infinie, tel que $\sum_{k=-\infty}^{+\infty} |k\gamma(k)| < +\infty$, alors

$$\sqrt{T}(\hat{\gamma}_T - \gamma) \overset{d}{\rightsquigarrow} N(0, \Sigma)$$

où N est un vecteur Gaussien l dimensionnel de matrice de covariance $\Sigma = [\Sigma_{kl}] = [\sum_{j=-\infty}^{+\infty} \gamma(j)(\gamma(j+k+l) + \gamma(j+k-l))]$.

On peut aussi estimer θ par la méthode du maximum de vraisemblance, où la variance σ^2 de ε_t est supposée inconnue. On a alors, sous les conditions de la propriété 9.51 de

Gouriéroux et Monfort [62],

$$\sqrt{T}\begin{pmatrix} \hat{\theta}_T - \theta \\ \hat{\sigma}_T^2 - \sigma^2 \end{pmatrix} \overset{d}{\rightsquigarrow} N(0, J(\theta, \sigma^2)^{-1})$$

où $J(\theta, \sigma^2)$ est la matrice d'information de Fisher. La matrice de variance covariance asymptotique de $\sqrt{T}(\hat{\theta}_T - \theta)$ vaut $\sigma^2/\gamma(0)\Gamma_{l-1}^{-1}$, où Γ_{l-1} est l'analogue sans estimateurs de la matrice $\hat{\Gamma}_{l-1}$ décrite précédemment.

Le prédicteur probabiliste s'écrit alors

$$X_{T+1}^* := E_\theta[X_{T+1}|X_{-\infty}^T] = \sum_{i=0}^{l-1} \theta_{i+1} X_{T-i}$$

qui est de la forme voulue avec $r_i(X_{T-i}, \theta) := \theta_{i+1} X_{T-i}$ et $\eta_T(\mathbb{X}, \theta) = 0$. On utilise l'une ou l'autre des méthodes d'estimation de θ pour construire le prédicteur statistique

$$\hat{X}_{T+1} = \sum_{i=0}^{l-1} \hat{\theta}_{\varphi(T)}(i+1) X_{T-i}$$

Les conditions $\mathbf{H}_0'$ (i)-(iii) d'application du corollaire 4.3 avec $k = l-1$ sont trivialement satisfaites. $\partial_{\theta_j} r_i(X_{t-i}, \theta) = X_{T-i}$ pour $j = i+1$, et 0 sinon. On a donc

$$\sum_{i=0}^{l-1} \partial_\theta r_i(X_{T-i}; \theta) = \sum_{i=0}^{l-1} \partial_\theta(\theta_{i+1} X_{T-i}) = \begin{bmatrix} X_T \\ X_{T-1} \\ \vdots \\ X_{T-l+1} \end{bmatrix} := V_T$$

Par ergodicité du processus, il existe un vecteur V tel que $V_T \overset{d}{\rightsquigarrow} V$, et la condition $\mathbf{H}_0'$ (iv) est vérifiée. En choisissant par exemple $\varphi(T) = T - \log T$, le corollaire 4.3 entraîne que

$$\sqrt{\varphi(T)}(\hat{X}_{T+1} - X_{T+1}^*) \overset{d}{\rightsquigarrow} < U, V >$$

où U est la loi limite de $\sqrt{T}(\hat{\theta}_T - \theta)$. □

Chapter 3

A quantile copula approach to conditional density estimation

Abstract : In this chapter, we investigate the problem of estimating the conditional density of a real-valued random variable Y given the value of an explanatory variable X. To that end, we present a new non-parametric estimator of the conditional density of the kernel type. It is based on a transformation of the data by quantile transform. By use of the copula representation, it turns out to have a remarkable product form. This chapter is organised as follows : in section 3.1, we present the two main approaches to tackle the problem and briefly review the literature. In section 3.2 we introduce the quantile transform and the copula representation which leads to the definition of our estimator. In section 3.3, the main asymptotic results are established. Proofs are mainly based on a series of auxiliary lemmas which are given in section 3.4. At last, the pointwise asymptotic results of section 3.3 are extended to uniformity on compact sets in section 3.5. Parts of this chapter are drawn from a paper, "A quantile copula approach to conditional density estimation" , submitted for publication and in revision. Discussions concerning its numerical implementation, comparison with competitors and possible extensions are

postponed to chapter 4. Applications to prediction are deferred until chapter 5. Note the followed approach could be also used to predict the conditional distribution, in the spirit of Bosq and Blanke [24].

3.1 Introduction

3.1.1 Motivation

Let $((X_i, Y_i); i = 1, \ldots, n)$ be an independent identically distributed sample from real-valued random variables (X, Y) sitting on a given probability space. For predicting the response Y of the input variable X at a given location x, it is of great interest to estimate not only the conditional mean or *regression function* $E(Y|X = x)$, but the full *conditional density* $f(y|x)$. Indeed, estimating the conditional density is much more informative, since it allows not only to recalculate from the density the conditional expected value $E(Y|X)$, but also many other characteristics of the distribution such as the conditional variance. In particular, having knowledge of the general shape of the conditional density, is especially important for multi-modal or skewed densities, which often arise from nonlinear or non-Gaussian phenomenas, where the expected value might be nowhere near a mode, i.e. the most likely value to appear. Moreover, for situations in which confidence intervals are preferred to point estimates, the estimated conditional density is an object of obvious interest.

3.1.2 Estimation by kernel density smoothing

A natural approach to estimate the conditional density $f(y|x)$ of Y given $X = x$ would be to exploit the identity

$$f(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}, \quad f_X(x) \neq 0 \tag{3.1}$$

where f_{XY} and f_X denote the joint density of (X, Y) and X, respectively. By introducing Parzen-Rosenblatt [102, 113] kernel estimators of these densities, namely,

$$\hat{f}_{n,XY}(x,y) := \frac{1}{n}\sum_{i=1}^{n} K'_{h'}(X_i - x)K_h(Y_i - y)$$

$$\hat{f}_{n,X}(x) := \frac{1}{n}\sum_{i=1}^{n} K'_{h'}(X_i - x)$$

where $K_h(.) = 1/hK(./h)$ and $K'_{h'}(.) = 1/h'K'(./h')$ are (rescaled) kernels with their associated sequence of bandwidth $h = h_n$ and $h' = h'_n$ going to zero as $n \to \infty$, one can construct the quotient

$$\hat{f}^R_n(y|x) := \frac{\hat{f}_{n,XY}(x,y)}{\hat{f}_{n,X}(x)}$$

and obtain an estimator of the conditional density. Such an estimator was first proposed by Rosenblatt [114]. See also Bosq [20]. Since then, several authors have studied it, among whom we may cite Youndjé, Sarda and Vieu [146, 147, 148], and Hyndman et al. [81], who slightly improved on Rosenblatt's kernel based estimator. In a dependent context, one can also cite Roussas [115], Collomb et al. [37], Lecoutre and Ould-Said [91].

3.1.3 Estimation by regression techniques

As pointed out by numerous authors, see e.g. Fan and Yao [51] chapter 6, this approach is equivalent to the one arising from considering this conditional density estimation problem in a regression framework. Indeed, let $F(y|x)$ be the cumulative conditional distribution function of Y given $X = x$. It stems from the fact that

$$E\left(\mathbb{1}_{|Y-y|\leq h}|X = x\right) = F(y + h|x) - F(y - h|x) \approx 2h.f(y|x)$$

as $h \to 0$, that, if one replace the expectation in the above expression by its empirical counterpart, one can apply the usual local averaging methods and perform a regression estimation on the synthetic data $((1/2h)\mathbb{1}_{|Y_i-y|\leq h};\ i = 1, \ldots, n)$. By a Bochner type

theorem, one can even replace the transformed data by its smoothed version

$$Y_i' := K_h(Y_i - y) := \frac{1}{h} K\left(\frac{Y_i - y}{h}\right).$$

In particular, the popular Nadaraya-Watson regression estimator [99, 144] can be applied in this setting,

$$\hat{f}_n^{NW}(y|x) := \frac{\sum_{i=1}^n Y_i' K_{h'}'(X_i - x)}{\sum_{i=1}^n K_{h'}'(X_i - x)}$$

The obtained estimator reduces itself to the same estimator of the conditional density of the double kernel type as before

$$\hat{f}_n^{NW}(y|x) := \frac{\sum_{i=1}^n K_h(Y_i - y) K_{h'}'(X_i - x)}{\sum_{i=1}^n K_{h'}'(X_i - x)} = \hat{f}_n^R(y|x).$$

Taking advantage of this regression formulation, Fan, Yao and Tong [52] proposed a conditional density estimator which generalizes the kernel one by use of the local polynomial techniques. In particular, it allows to tackle the bias issues of the kernel smoothing. However, and unlike the former, it is no longer guaranteed to have positive value nor to integrate to 1 with respect to y. With these issues in mind, Hyndman and Yao [82] built on local polynomial techniques and suggested two improved methods, the first one based on locally fitting a log-linear model and the second one on constrained local polynomial modeling. An overview can be found in Fan and Yao [51] (chapter 6 and 10). Very recently, Györfi and Kohler [65] studied a partitioning type estimate and studied its properties in total variation norm and Lacour [87] a projection-type estimate for Markov chains. Extension to functional data is studied in Laksaci [88, 89] and also in the book by Ferraty and Vieu [57]. Minimax rates of convergences are obtained in Efromovich [49]. A Bayesian approach is to be found in Tang and Ghosal [132] and the k nearest neighbour estimate in Yu [149]. De Gooijer et al. [39] study a linear combination of the above kernel and local polynomial estimates to combine the advantages of the two classical approaches. See remark 3.4 below and subsection 4.3 for a more detailed accounts of some of these estimators.

3.1.4 A product shaped estimator

However, these two equivalent approaches suffer from several drawbacks: first, by its form as a quotient of two estimators, the probabilistic behaviour of the Nadaraya-Watson estimator (or its local polynomial counterpart) is tricky to study. It is usually dealt with by a centering at expectation for both numerator and denominator and a linearizing of the inverse, see e.g. [51], or [21] for details. Second, at a conceptual level, one could argue that implementing regression estimation techniques in this setting is, in a sense, unnatural: estimating a density, even if it is a conditional one, should resort to density estimation techniques only. Finally, practical implementations of these estimators can lead to numerical instability when the denominator is close to zero.

To remedy these problems, we propose an estimator which builds on the idea of using synthetic data, i.e. a representation of the data more adapted to the problem than the original one. By transforming the data by quantile transforms and making use of the copula function, the estimator turns out to have a remarkable *product* form

$$\hat{f}_n(y|x) = \hat{f}_Y(y)\hat{c}_n(F_n(x), G_n(y))$$

where $\hat{f}_Y$, $\hat{c}_n$, $F_n(x)$, $G_n(y)$ are estimators of the density f_Y of Y, the copula density c, the c.d.f. F of X and G of Y respectively (see next section below for definitions). Its study then reveals to be particularly simple: it reduces to the ones already done on nonparametric density estimation.

3.2 Presentation of the estimator

3.2.1 The quantile transform

The idea of transforming the data is not new. It has been used to improve the range of applicability and performance of classical estimation techniques, e.g. to deal with

skewed data, heavy tails, or restrictions on the support (see e.g. Carroll and Ruppert [31], Devroye and Lugosi [45] chapter 14 and the references therein, and also Van der Vaart [136] chapter 3.2 for the related topic of variance stabilizing transformations in a parametric context). In order to make inference on Y from X, a natural question which then arises is, what is the "best" transformation, if this question has a sense. As one can note from the above references, the "best" transformation is intimately linked to the distribution of the underlying data. We will see below that, for our problem, the natural candidate is the quantile transform.

The quantile transform is a well-known probabilistic device which is used to reduce proofs, e.g. in empirical process theory, for arbitrary real valued random variables X to ones for random variables U uniformly distributed on the interval $[0, 1]$. First of all, let us recall the definition of the generalised inverse F^{-1} of a non-decreasing right-continuous distribution function, i.e. the (left continuous version of the) quantile function Q:

Definition 3.1 *The generalised inverse F^{-1} of a non-decreasing right-continuous distribution function F is*

$$F^{-1}(u) := Q(u) = \inf\{x \in \mathbb{R} | F(x) \geqslant u\}, \quad u \in (0, 1)$$

The quantile transform is based on the following well-known lemma (See e.g. Shorack and Wellner [120], chapter 1).

Lemma 3.2 *For any real-valued random variable X, with distribution function F, the following property holds.*

i) Whenever F is continuous, the random variable $U = F(X)$ is uniformly distributed on $(0, 1)$;

ii) Conversely, when F is arbitrary, if U is a uniformly distributed random variable on $(0, 1)$, X is equal in law to $F^{-1}(U)$.

As a consequence, given a sample $(X_1, \ldots, X_n)$ of random variables with common continuous c.d.f. F sitting on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, one can always enlarge this probability space to carry a sequence $(U_1, \ldots, U_n)$ of uniform $(0, 1)$ random variables such that $U_i = F(X_i)$, that is to say to construct a pseudo-sample with a *prescribed uniform* marginal distribution.

Thus, the quantile transform can be viewed as a symmetrization or invariance device. When applied to a multivariate distribution, it leads to the definition of the copula function, as presented below.

3.2.2 The copula representation

Formally, a copula is a bi-(or multi)variate distribution function whose marginal distribution functions are uniform on the interval $[0, 1]$. Indeed, Sklar [122] proved the following fundamental result:

Theorem 3.3 (Sklar) *For any bivariate cumulative distribution function $F_{X,Y}$ on $\mathbb{R}^2$, with marginal cumulative distribution functions F of X and G of Y, there exists some function $C : [0, 1]^2 \to [0, 1]$, called the dependence or copula function, such as*

$$F_{X,Y}(x, y) = C(F(x), G(y)) \ , \ -\infty \leq x, y \leq +\infty. \tag{3.2}$$

If F and G are continuous, this representation is unique with respect to (F, G). The copula function C is itself a cumulative distribution function on $[0, 1]^2$ with uniform marginals.

This theorem gives a representation of the bivariate c.d.f. as a function of each univariate c.d.f. In other words, the copula function captures the dependence structure among the components X and Y of the vector (X, Y), irrespectively of the marginal distribution F and G. Simply put, it allows to deal with the randomness of the dependence structure and the randomness of the marginals *separately*.

Copulas appear to be naturally linked with the quantile transform : in the case F and G are continuous, formula (3.2) is simply obtained by defining the copula function as $C(u,v) = F_{X,Y}(F^{-1}(u), G^{-1}(v))$, $0 \leq u \leq 1$, $0 \leq v \leq 1$. For more details regarding copulas and their properties, one can consult for example the book of Joe [83]. Copulas have witnessed a renewed interest in statistics, especially in finance, since the pioneering work of Rüschendorf [117] and Deheuvels [42], who introduced the empirical copula process. Weak convergence of the empirical copula process was investigated by Deheuvels [43], Van der Vaart and Wellner [137], Fermanian, Radulovic and Wegkamp [56]. For the estimation of the copula density, refer to Gijbels and Mielniczuk [62], Fermanian [54] and Fermanian and Scaillet [55].

From now on, we assume that the copula function $C(u,v)$ has a density $c(u,v)$ with respect to the Lebesgue measure on $[0,1]^2$ and that F and G are strictly increasing and differentiable with densities f and g. $C(u,v)$ and $c(u,v)$ are then the cumulative distribution function (c.d.f.) and density respectively of the transformed variables $(U,V) = (F(X), G(Y))$. By differentiating formula (3.2), we get for the joint density,

$$f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y} = f(x)g(y)c(F(x), G(y))$$

where $c(u,v) := \frac{\partial^2 C(u,v)}{\partial u \partial v}$ is the above mentioned copula density. Eventually, we can obtain the following explicit formula of the conditional density

$$f(y|x) = \frac{f_{XY}(x,y)}{f(x)} = g(y)c(F(x), G(y)) \tag{3.3}$$

provided $f(x) \neq 0$.

3.2.3 Construction of the estimator

Starting from the previously stated product type formula (3.3), a natural plug-in approach to build an estimator of the conditional density is to use

- a Parzen-Rosenblatt kernel type non parametric estimator of the marginal density

g of Y,

$$\hat{g}_n(y) := \frac{1}{nh_n} \sum_{i=1}^{n} K_0 \left(\frac{y - Y_i}{h_n} \right)$$

- the empirical distribution functions $F_n(x)$ and $G_n(y)$ for $F(x)$ and $G(y)$ respectively,

$$F_n(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{X_j \leq x} \text{ and } G_n(y) := \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{Y_j \leq y}.$$

Concerning the copula density $c(u, v)$, we noted that $c(u, v)$ is the joint density of the transformed variables $(U, V) := (F(X), G(Y))$. Therefore, $c(u, v)$ can be estimated by the bivariate Parzen-Rosenblatt kernel type non parametric density (pseudo) estimator,

$$c_n(u, v) := \frac{1}{na_n b_n} \sum_{i=1}^{n} K \left(\frac{u - U_i}{a_n}, \frac{v - V_i}{b_n} \right) \tag{3.4}$$

where K is a bivariate kernel and a_n, b_n its associated bandwidth. For simplicity, we restrict ourselves to product kernels, i.e. $K(u, v) = K_1(u)K_2(v)$ with the same bandwidths $a_n = b_n$.

Nonetheless, since F and G are unknown, the random variables (U_i, V_i) are not observable, i.e. c_n is not a true statistic. Therefore, we approximate the pseudo-sample $(U_i, V_i), i = 1, \ldots, n$ by its empirical counterpart $(F_n(X_i), G_n(Y_i)), i = 1, \ldots, n$. We therefore obtain a genuine estimator of $c(u, v)$

$$\hat{c}_n(u, v) := \frac{1}{na_n^2} \sum_{i=1}^{n} K_1 \left(\frac{u - F_n(X_i)}{a_n} \right) K_2 \left(\frac{v - G_n(Y_i)}{a_n} \right). \tag{3.5}$$

Eventually, the conditional density estimator is written as

$$\hat{f}_n(y|x) := \left[\frac{1}{nh_n} \sum_{i=1}^{n} K_0 \left(\frac{y - Y_i}{h_n} \right) \right] \cdot \left[\frac{1}{na_n^2} \sum_{i=1}^{n} K_1 \left(\frac{F_n(x) - F_n(X_i)}{a_n} \right) K_2 \left(\frac{G_n(y) - G_n(Y_i)}{a_n} \right) \right]$$

or, under a more compact form,

$$\hat{f}_n(y|x) := \hat{g}_n(y)\hat{c}_n(F_n(x), G_n(y)). \tag{3.6}$$

Remark 3.4 *To our knowledge, the estimator studied in this paper has never been proposed in the literature. However, and after we formulated our estimator, we found that*

some connections can be made with the smooth nearest neighbor one proposed by Yang [145] and Stute [128] and the Gasser and Müller [58] and Priestley and Chao [107] one in the context of regression estimation. See also Stute [130] and [131] for the application of his estimator to the estimation of the conditional cumulative distribution function and conditional empirical process. Indeed, the Gasser and Müller estimators tackle the issue of having a random denominator by first transforming the design $X_1, \ldots, X_n$ to a uniform (random) one. This result in assigning the surfaces under the kernel function instead of its heights as weights. Contrary to our estimator, they do not make transformations of the data in both directions X and Y. The Gasser and Müller-type estimators, which, has a convolution shape, is presented below, for convenience:

$$m_n^{GM(1)}(x) := \frac{1}{h_n} \sum_{i=1}^{n-1} \left\{ \int_{X_{i,n}}^{X_{i+1,n}} K\left(\frac{x-u}{h_n}\right) du \right\} Y_{[i]}$$

$$m_n^{GM(2)}(x) := \frac{1}{h_n} \sum_{i=1}^{n} (X_{i+1,n} - X_{i,n}) K\left(\frac{x-X_{i,n}}{h_n}\right) Y_{[i]}$$

where $X_{i,n}$ denotes the ith order statistic of the sample $(X_1, \ldots, X_n)$ and $Y_{[i]}$ its corresponding Y value (i.e its concomitant). Regarding Yang's [145] and Stute's [128] estimators of the regression function, they are based on the reduction of the design data to a uniform one by use of the quantile transform,

$$m_n^{YS(1)}(x) := (na_n)^{-1} \sum_{i=1}^{n} Y_i K\left(\frac{F_n(x) - F_n(X_i)}{a_n}\right)$$

$$m_n^{YS(2)}(x) := (na_n)^{-1} \frac{\sum_{i=1}^{n} Y_i K\left(\frac{F_n(x)-F_n(X_i)}{a_n}\right)}{\sum_{i=1}^{n} K\left(\frac{F_n(x)-F_n(X_i)}{a_n}\right)}$$

Note that the first version $m_n^{YS(1)}$ of the estimator of the regression function is also product shaped. However, these estimators are not based on a transformation of the data in both direction.

3.3 Pointwise Asymptotic results

3.3.1 Notations and assumptions

We note the ith moment of a generic kernel (possibly multivariate) K as

$$m_i(K) := \int u^i K(u)du$$

and the $\mathbb{L}_p$ norm of a function h by $||h||_p := \int h^p$. We use the sign $\simeq$ to denote the order of the bandwidths, i.e. $h_n \simeq u_n$ means that $h_n = c_n u_n$ with $c_n \to c > 0$. The support of the densities function f and g are noted by $\text{supp}(f) = \overline{\{x \in \mathbb{R}; f(x) > 0\}}$ and $\text{supp}(g) = \overline{\{y \in \mathbb{R}; g(y) > 0\}}$, where $\overline{A}$ stands for the closure of a set A. Finally, $o_P(.)$ and $O_p(.)$ (respectively $o_{a.s.}(.)$ and $O_{a.s}(.)$)will stands for convergence and boundedness in probability (respectively almost surely) as in [136].

To state our results, we will have to make some regularity assumptions on the kernels and the densities which, although far from being minimal, are somehow customary in kernel density estimation (see subsections 3.4.2 and 3.5.2 for discussions and details). Set x and y two fixed points in the interior of $\text{supp}(f)$ and $\text{supp}(g)$ respectively.

Assumption A

(i) the c.d.f F of X and G of Y are strictly increasing and differentiable;

(ii) the densities g and c are twice continuously differentiable with bounded second derivatives on their support.

(iii) the densities g anc c are uniformly continuous and non-vanishing almost everywhere on a compact set $J := [a, b]$ and $D \subset (0, 1) \times (0, 1)$ included in the interior of $\text{supp}(g)$ and $\text{supp}(c)$, respectively.

Moreover, we assume that the kernels K_0 and K satisfy the following:

Assumption B

(i) K and K_0 are of bounded support and of bounded variation;

(ii) $0 \leq K \leq C$ and $0 \leq K_0 \leq C$ for some constant C;

(iii) K and K_0 are second order kernels: $m_0(K) = \int K = 1$, $m_1(K) = \int xK(x)dx = 0$ and $m_2(K) = \int x^2K(x)dx < +\infty$, and the same for K_0;

(iv) K it is twice differentiable with bounded second partial derivatives.

3.3.2 Heuristic

Recall that $c_n(u, v)$ is the kernel copula (pseudo) density estimator from the unobservable, but fixed with respect to n, pseudo data $(F(X_i), G(X_i))$, and that $\hat{c}_n(u, v)$ is its analogue made from the approximate data $(F_n(X_i), G_n(X_i)$. The heuristic of the reason why our estimator works is that the $n^{-1/2}$ in probability rate of convergence in uniform norm of F_n and G_n to F and G is faster than the $1/\sqrt{na_n^2}$ rate of the non parametric kernel estimator c_n of the copula density c. Therefore, the approximation step of the unknown transformations F and G by their empirical counterparts F_n and G_n does not have any impact asymptotically on the estimation step of c by c_n. Put in another way, one can approximate $\hat{c}_n(F_n(x), G_n(y))$ by $c_n(F(x), G(y))$ at a faster rate than the convergence rate of $c_n(F(x), G(y))$ to $c(F(x), G(y))$. This is what is proved in the two approximation propositions of section 3.4, which imposes some conditions on the bandwidth a_n for the approximation to hold, among which $na_n^4 \to \infty$ for the in probability rate. The convergence properties of our estimator will then result from the well-known convergence properties of the kernel density estimators, which are also recalled in section 3.4.

3.3.3 Weak and strong consistency of the estimator

We have the following pointwise weak consistency theorem:

Theorem 3.5 *Let the regularity assumptions* ***A*** *(i)-(ii) and* ***B*** *(i)-(iv) on the densities*

and kernels be satisfied, if h_n and a_n tends to zero as $n \to \infty$ in such a way that

$$nh_n \to \infty, \quad na_n^4 \to \infty, \quad \frac{\sqrt{\ln \ln n}}{na_n^3} \to 0,$$

then,

$$\hat{f}_n(y|x) = f(y|x) + O_P\left(\frac{1}{\sqrt{nh_n}} + h_n^2 + a_n^2 + \frac{1}{\sqrt{na_n^2}} + \frac{1}{na_n^4} + \frac{\sqrt{\ln \ln n}}{na_n^3}\right).$$

Proof *Recall from 3.4 and 3.5 that c_n and $\hat{c}_n$ are estimators of the copula density c based respectively on unobservable pseudo-data $(F(X_i), G(Y_i)$, and their approximations $(F_n(X_i), G_n(Y_i))$. The main ingredient of the proof follows from the decomposition:*

$$\begin{aligned}
\hat{f}_n(y|x) - f(y|x) &= \hat{g}_n(y)\hat{c}_n(F_n(x), G_n(y)) - g(y)c(F(x), G(y)) \\
&= [\hat{g}_n(y) - g(y)]\, \hat{c}_n(F_n(x), G_n(y)) \\
&\quad + g(y)\, [\hat{c}_n(F_n(x), G_n(y)) - c(F(x), G(y))] \\
&:= D_1 + D_2
\end{aligned}$$

We proceed one step further in the decomposition of each terms, by first centering at fixed locations,

$$\begin{aligned}
D_1 &= [\hat{g}_n(y) - g(y)]\, [\hat{c}_n(F_n(x), G_n(y)) - \hat{c}_n(F(x), G(y))] \\
&\quad + [\hat{g}_n(y) - g(y)]\, [\hat{c}_n(F(x), G(y)) - c_n(F(x), G(y))] \\
&\quad + [\hat{g}_n(y) - g(y)]\, [c_n(F(x), G(y)) - c(F(x), G(y))] \\
&\quad + [\hat{g}_n(y) - g(y)]\, [c(F(x), G(y))]
\end{aligned} \tag{3.7}$$

$$\begin{aligned}
D_2 &= g(y)\, [\hat{c}_n(F_n(x), G_n(y)) - \hat{c}_n(F(x), G(y))] \\
&\quad + g(y)\, [\hat{c}_n(F(x), G(y)) - c_n(F(x), G(y))] \\
&\quad + g(y)\, [c_n(F(x), G(y)) - c(F(x), G(y))]
\end{aligned} \tag{3.8}$$

On the one hand, convergence results for the kernel density estimators of section 3.4.2

entail that,

$$\hat{g}_n(y) - g(y) = O_p(h_n^2 + 1/\sqrt{nh_n})$$
$$c_n(F(x), G(y)) - c(F(x), G(y)) = O_p(a_n^2 + 1/\sqrt{na_n^2})$$

by lemma 3.14 and 3.15 respectively. On the other hand, as it is assumed that $a_n \to 0$ *and* $na_n^4 \to \infty$*, we have that* $na_n^3 \to \infty$ *and approximation propositions 3.16 and 3.17 of sections 3.4.4 and 3.4.5 apply and entail that*

$$\hat{c}_n(F(x), G(y)) - c_n(F(x), G(y)) = O_P\left(n^{-1/2} + \frac{\sqrt{\ln\ln n}}{na_n^3} + \frac{1}{na_n^4}\right)$$
$$\hat{c}_n(F_n(x), G_n(y)) - \hat{c}_n(F(x), G(y)) = O_P\left(\frac{1}{\sqrt{n}} + \frac{1}{na_n^4}\right).$$

and the conditions

$$\frac{\sqrt{\ln\ln n}}{na_n^3} \to 0, \quad \frac{1}{na_n^4} \to 0.$$

entail these latter terms be negligible in probability.

We therefore obtain, by neglecting the $n^{-1/2}$ *terms, that*

$$\begin{aligned} D_1 &= [\hat{g}_n(y) - g(y)]\,[c(F(x), G(y)) + o_P(1)] \\ &= O_P\left(h_n^2 + 1/\sqrt{nh_n}\right) \\ D_2 &= g(y) O_P\left(a_n^2 + \frac{1}{\sqrt{na_n^2}} + \frac{1}{na_n^4} + \frac{\sqrt{\ln\ln n}}{na_n^3}\right) \end{aligned}$$

Thus the claimed result. □

Remark 3.6 *As a corollary, we get the rate of convergence, by choosing the bandwidths which balance the bias and variance trade-off: for an optimal choice of* $h_n \simeq n^{-1/5}$ *and* $a_n \simeq n^{-1/6}$*, we get*

$$\hat{f}(y|x) = f(y|x) + O_P(n^{-1/3}).$$

Therefore, our estimator is rate optimal in the sense that it reaches the minimax rate $n^{-1/3}$ *of convergence, according to Efromovich [49].*

Almost sure results can be proved in the same way: we have the following strong consistency result,

Theorem 3.7 *Let the regularity conditions* ***A*** *(i)-(ii) and* ***B*** *(i)-(iv) on the densities and kernels be satisfied. If the bandwidths h_n and a_n tends to zero as $n \to \infty$ in such a way that*

$$\frac{nh_n}{\ln\ln n} \to \infty, \quad \frac{(\ln n)^{1/2}(\ln\ln n)^{1/2}}{na_n^3} \to 0, \quad \frac{\ln\ln n}{na_n^4} \to 0,$$

then,

$$\hat{f}_n(y|x) = f(y|x) + O_{a.s.}\left(h_n^2 + \sqrt{\frac{\ln\ln n}{nh_n}} + a_n^2 + \sqrt{\frac{\ln\ln n}{na_n^2}} + \frac{\ln\ln n}{na_n^4} + \frac{\sqrt{\ln n \ln\ln n}}{na_n^3}\right).$$

Proof *It follows the same lines as the preceding theorem, but uses the a.s. results of the consistency of the kernel density estimators of lemmas 3.14 and 3.15 and of the approximation propositions 3.16 and 3.17. It is therefore similar and omitted.*

Remark 3.8 *For $h_n \simeq (\ln\ln n/n)^{1/5}$ and $a_n \simeq (\ln\ln n/n)^{1/6}$ which is the optimal trade-off between the bias and the stochastic term, one gets the optimal rate $(\ln\ln n/n)^{1/3}$.*

3.3.4 Convergence in distribution

Theorem 3.9 *Let the regularity conditions* ***A*** *(i)-(ii) and* ***B*** *(i)-(iv) on the densities and kernels be satisfied. If (x, y) are such that $g(y) > 0$ and $c(F(x), G(y)) > 0$, and the bandwidths h_n and a_n tend to zero in such a way that*

$$nh_n \to \infty, \quad \frac{\sqrt{\ln\ln n}}{na_n^3} \to 0, \quad na_n^4 \to \infty, \quad na_n^6 \to 0,$$

then,

$$\sqrt{na_n^2}\left(\hat{f}_n(y|x) - f(y|x)\right) \overset{d}{\rightsquigarrow} \mathcal{N}\left(0, g(y)f(y|x)||K||_2^2\right).$$

Proof *With the conditions on the bandwidths, the previous decomposition 3.7 and 3.8*

writes

$$\sqrt{na_n^2}\left(\hat{f}_n(y|x) - f(y|x)\right) = \sqrt{na_n^2}[c_n(F(x), G(y)) - c(F(x), G(y))]\,[g(y) + o_P(1)] + o_P(1)$$

Now, $a_n \to 0$, $na_n^2 \to \infty$ *and* $na_n^6 \to 0$ *entails, via lemma 3.15 of section 3.4, that* $c_n(F(x), G(y)) - c(F(x), G(y))$ *is asymptotically normal,*

$$\sqrt{na_n^2}g(y)\,[c_n(F(x), G(y)) - c(F(x), G(y))] \overset{d}{\rightsquigarrow} \mathcal{N}\left(0, g^2(y)c(F(x), G(y))\,\|K\|_2^2\right).$$

An application of Slutsky's lemma yields the desired result.

For a vector $(y_1, \ldots, y_d)$, one can get a multidimensional version of the convergence in distribution (fidi convergence):

Corollary 3.10 *With the same assumptions, for* $(y_1, \ldots, y_d)$ *in the interior of* $supp(g)$ *such that* $g(y_i)f(y_i|x) \neq 0$,

$$\sqrt{na_n^2}\left(\left(\frac{\hat{f}_n(y_i|x) - f(y_i|x)}{\sqrt{g(y_i)f(y_i|x)}\,\|K\|_2}\right), i = 1, ..., m\right) \overset{d}{\rightsquigarrow} N^{(m)}$$

where $N^{(m)}$ *is the standard* m*-variate centered normal distribution with identity variance matrix.*

Proof *It simply follows from the use of the Cramér-Wold device and is therefore omitted. For details, see e.g. [21], theorem 2.3.*

3.3.5 Asymptotic Bias, Variance and Mean square error

With the rates involved in the previous theorems , the estimator can be written in the following form,

$$\hat{f}_n(y|x) = f(y|x) + g(y)N_n(x, y) + g(y)B_n(x, y) + R_n$$

with

$$\begin{aligned} R_n &= o_{a.s}(a_n^2 + (na_n^2)^{-1/2}) \\ B_n(x,y) &= Ec_n(F(x),G(y)) - c(F(x),G(y)) \\ &= B_K(c,x,y)\frac{a_n^2}{2} + o(a_n^2) \\ N_n(x,y) &= c_n(F(x),G(y)) - Ec_n(F(x),G(y)) \end{aligned}$$

B_n is the deterministic bias of the bivariate kernel pseudo density estimator with constant

$$B_K(c,x,y) := m_2(K_1)\frac{\partial^2 c(F(x),G(y))}{\partial u^2} + m_2(K_2)\frac{\partial^2 c(F(x),G(y))}{\partial v^2},$$

and N_n is asymptotically Normal, i.e. such as

$$\sqrt{na_n^2}\left[\frac{N_n(x,y)}{c^{1/2}(F(x),G(y))||K||_2}\right] \overset{d}{\rightsquigarrow} \mathcal{N}(0,1).$$

The asymptotic bias and variance of the statistics which are asymptotically equivalent to the estimator are thus given by

$$g(y)B_n(x,y) = g(y)B_K(c,x,y)\frac{a_n^2}{2} + o(a_n^2)$$

and

$$1/(na_n^2)g(y)f(y|x)||K||_2^2 + o(1/(na_n^2)),$$

respectively.

We thus have the following claim on the bias of the proposed estimator:

Claim 3.11 *With the assumptions of the previous theorems, we have*

$$B_0 := E(\hat{f}_n(y|x)) - f(y|x) = g(y)B_K(c,x,y)\frac{a_n^2}{2} + o(a_n^2)$$

with $B_K(c,x,y) := m_2(K_1)\frac{\partial^2 c(F(x),G(y))}{\partial u^2} + m_2(K_2)\frac{\partial^2 c(F(x),G(y))}{\partial v^2}$.

A similar statement about the asymptotic variance of the estimator can be made, which leads to the following claim on the asymptotic mean squared error:

Claim 3.12 *The Asymptotic Mean Squared Error (AMSE) E_0 at (x, y) is*

$$\begin{aligned} E_0 &:= B_0^2 + V_0 \\ &= \frac{a_n^4 g^2(y) \left(B_k(c,x,y)\right)^2}{4} + \frac{g(y) f(y|x) ||K||_2^2}{n a_n^2} + o\left(a_n^4 + \frac{1}{n a_n^2}\right) \end{aligned}$$

which gives, for the choice of the bandwidth $a_n \simeq n^{-1/6}$ mentioned above,

$$E_0 = n^{-2/3} g^2(y) \left(\frac{B_K^2(c,x,y)}{4} + c(F(x), G(y)) ||K||_2^2 \right) + o(n^{-2/3}).$$

3.4 Intermediate and auxiliary results

In this section, we gather some preliminary results which we will need as basic tools for the demonstrations of section 3.3. In subsection 3.4.1, we recall classical results about the convergence of the Kolmogorov-Smirnov statistic. Next, we make a brief overview of kernel density estimation and apply these results to the estimators $\hat{g}_n$ (section 3.4.2) and c_n (section 3.4.3). Eventually, we need two approximation propositions of $\hat{c}_n$ by c_n in sections 3.4.4 and 3.4.5.

3.4.1 Approximation of the pseudo-variables $F(X_i)$ by their estimates $F_n(X_i)$

For $(X_i, i = 1, \ldots, n)$ an i.i.d. sample of a real random variable X with common c.d.f. F, the Kolmogorov-Smirnov statistic is defined as $D_n := \|F_n - F\|_\infty$. Glivenko-Cantelli, Kolmogorov and Smirnov, Chung, Donsker among others have studied its convergence properties in increasing generality (See e.g. [120] and [137] for recent accounts). For our purpose, we only need to formulate these results in the following rough form:

Lemma 3.13 *For an i.i.d. sample from a continuous c.d.f. F,*

$$\|F_n - F\|_\infty = O_{a.s.}\left(\sqrt{\frac{\ln\ln n}{n}}\right) \tag{3.9}$$

$$\|F_n - F\|_\infty = O_P\left(\frac{1}{\sqrt{n}}\right). \tag{3.10}$$

Since F is unknown, the random variables $U_i = F(X_i)$ are not observed. As a consequence of the preceding lemma 3.13, one can naturally approximate these variables by the statistics $F_n(X_i)$. Indeed,

$$|F(X_i) - F_n(X_i)| \leq \sup_{x\in R} |F(x) - F_n(x)| = \|F_n - F\|_\infty \quad \text{a.s.}$$

Thus, $|F(X_i) - F_n(X_i)|$ is no more than an $O_P((\ln\ln n/n)^{1/2})$ or an $O_{a.s.}(n^{-1/2})$. These rates of approximation appears to be faster than those of statistical estimation of densities, as is shown in the next subsection.

3.4.2 Convergence of the kernel density estimator $\hat{g}_n$

We recall below some classical results about the convergence of the Parzen-Rosenblatt kernel non-parametric estimator $\hat{f}_n$ of a d-variate density f. Since its inception by Rosenblatt [113] and Parzen [102], it has been studied by a great deal of authors. See, among others, Prakasa Rao [105], Bosq and Lecoutre [26], Nadaraya [100], Scott [119], Wand and Jones [143], for details. See also Bosq [21] chapter 2.

It is well known that the bias of the kernel density estimator depends on the degree of smoothness of the underlying density, measured by its number of derivatives or its Lipschitz order. In order to get the convergence of the bias to zero, it suffices to assume that the density is continuous (See [102]). To get further information on the rate of convergence of the estimator, it is necessary to make further assumptions. Moreover, for kernel functions with unbounded support, the rate of convergence also depends on the tail behaviour of the kernel (See Stute [126]). Therefore, for clarity of exposition and simplicity of notations, we will make the customary assumptions that the density is twice

differentiable and that the kernel is of bounded support. We then have the following results:

- Bias: With the previous assumptions, for a x in the interior of $\text{supp}(f)$, $h_n \to 0$ and $nh_n^d \to \infty$ entail that

$$E\hat{f}_n(x) = f(x) + \frac{h_n^2}{2} \int_{\mathbb{R}^d} \sum_{1 \leq i,j \leq d} \frac{\partial^2 f(x)}{\partial x_i \partial x_j} z_i z_j K(z) dz + o(h_n^2).$$

  With the multivariate kernel K as a product of d order two kernels K_i, the above sum reduces to the diagonal terms.

$$E\hat{f}_n(x) = f(x) + \frac{h_n^2}{2} \sum_{1 \leq i \leq d} m_2(K_i) \frac{\partial^2 f(x)}{\partial x_i^2} + o(h_n^2).$$

- Variance: with the same assumptions,

$$Var\left[\hat{f}_n(x)\right] = \frac{f(x)}{nh_n^d} \|K\|_2^2 + o\left(\frac{1}{nh_n^d}\right).$$

- Convergence in quadratic mean: from the previous results, one can show that a necessary and sufficient condition for (pointwise) convergence in quadratic mean is $h_n \to 0$ and $nh_n^d \to \infty$. For a choice of the bandwidth as $h_n \simeq n^{-1/(d+4)}$, which realizes the optimal trade-off between the bias and variance, one gets the rate $n^{-2/(d+4)}$, which is the optimal speed of convergence in the minimax sense in the class of density functions with bounded second derivatives, according to Stone [125].

- Pointwise asymptotic normality: in addition to the previous conditions, if f is strictly positive in x, $h_n \to 0$ and $nh_n^d \to \infty$, then

$$\sqrt{nh_n^d}\left(\hat{f}_n(x) - E\hat{f}_n(x)\right) \overset{d}{\rightsquigarrow} \mathcal{N}(0, f(x)\|K\|_2^2).$$

  If the bandwidth is small enough to make the bias negligible, one gets the following corollary: if f is strictly positive in x, $h_n \to 0$, $nh_n^d \to \infty$ and $nh_n^{d+4} \to 0$, then

$$\sqrt{nh_n^d}\left(\frac{\hat{f}_n(x) - f(x)}{(\hat{f}_n(x)\|K\|_2^2)^{1/2}}\right) \overset{d}{\rightsquigarrow} \mathcal{N}(0,1).$$

- Pointwise almost sure convergence: if moreover $nh_n^d/(\ln\ln n) \to \infty$ (see [41]), we have that

$$\hat{f}_n(x) - E\hat{f}_n(x) = O_{a.s.}\left(\sqrt{\frac{\ln\ln n}{nh_n^d}}\right).$$

  For a choice of the bandwidth as $h_n \simeq ((\ln\ln n)/n)^{1/(d+4)}$, we get the rate of convergence $((\ln\ln n)/n)^{2/(d+4)}$:

$$\hat{f}_n(x) - f(x) = O_{a.s.}\left(\left(\frac{\ln\ln n}{n}\right)^{2/(d+4)}\right).$$

Applied to our case ($d = 1$), we can summarize these results for further reference in the following lemma for the estimator $\hat{g}_n$ of the density g of Y:

Lemma 3.14 *With the previous assumptions, for a point y in the interior of the support of g, we have,*

- *for a bandwidth chosen such as $h_n \simeq n^{-1/5}$,*

$$|\hat{g}_n(y) - g(y)| = O_p(n^{-2/5}),$$

- *for a point y where $g(y) > 0$, and $h_n = o(n^{-1/5})$,*

$$\sqrt{nh_n}\,[\hat{g}_n(y) - g(y)] \overset{d}{\rightsquigarrow} \mathcal{N}\left(0, g(y)\,\|K_0\|_2^2\right),$$

- *for a bandwidth choice of $h_n \simeq (\ln\ln n/n)^{1/5}$,*

$$\hat{g}_n(y) - g(y) = O_{a.s.}\left(\left(\frac{\ln\ln n}{n}\right)^{2/5}\right).$$

3.4.3 Convergence of $c_n(u, v)$

As mentioned before, the assumptions that F and G be differentiable and strictly increasing entail that c is the density of the transformed variables $(U, V) := (F(X), G(Y))$. Therefore, once one convinces oneself that $c_n(u, v)$ is simply the kernel density estimator of the bivariate density $c(u, v)$ of the pseudo-variables (U, V), one directly draws its convergence properties by applying the results of the preceding subsection with $d = 2$:

Lemma 3.15 *With the previous assumptions, for $(u,v) \in (0,1)^2$, we have,*

- *for a bandwidth chosen such as $a_n \simeq n^{-1/6}$,*

$$|c_n(u,v) - c(u,v)| = O_p(n^{-1/3}),$$

- *for a point (u,v) where $c(u,v) > 0$, and $a_n = o(n^{-1/6})$,*

$$\sqrt{na_n^2}\left(\frac{c_n(u,v) - c(u,v)}{(c_n(u,v)\,\|K\|_2^2)^{1/2}}\right) \overset{d}{\rightsquigarrow} \mathcal{N}(0,1).$$

- *for a bandwidth choice of $a_n \simeq (\ln\ln n/n)^{1/6}$,*

$$c_n(u,v) - c(u,v) = O_{a.s.}\left(\left(\frac{\ln\ln n}{n}\right)^{1/3}\right).$$

3.4.4 An approximation proposition of $\hat{c}_n(u,v)$ by $c_n(u,v)$

The proposition of this section gives the rate of approximation of the kernel copula density estimator $\hat{c}_n(u,v)$ computed on the real data $(F_n(X_i), G_n(Y_i))$ by its analogue $c_n(u,v)$ computed on the pseudo-data $(U_i,V_i) := (F(X_i), G(Y_i))$. A similar result, but with a different proof, has been obtained in Fermanian [54] theorem 1.

Proposition 3.16 *Let $(u,v) \in (0,1)^2$. If the kernel $K(u,v) = K_1(u)K_2(v)$ is twice differentiable with bounded second derivatives, then*

$$|\hat{c}_n(u,v) - c_n(u,v)| = O_P\left(n^{-1/2} + \frac{\sqrt{\ln\ln n}}{na_n^3} + \frac{1}{na_n^4}\right)$$

$$|\hat{c}_n(u,v) - c_n(u,v)| = O_{a.s.}\left(\left(\frac{\ln\ln n}{n}\right)^{1/2} + \frac{(\ln n)^{1/2}(\ln\ln n)^{1/2}}{na_n^3} + \frac{\ln\ln n}{na_n^4}\right)$$

Proof *We note $||.||$ a norm for vectors and ^{T} for transpose. Set*

$$\Delta := \hat{c}_n(u,v) - c_n(u,v) = \frac{1}{na_n^2}\sum_{i=1}^{n}\Delta_{i,n}(u,v)$$

with

$$\Delta_{i,n}(u,v) := K\left(\frac{u - F_n(X_i)}{a_n}, \frac{v - G_n(Y_i)}{a_n}\right) - K\left(\frac{u - F(X_i)}{a_n}, \frac{v - G(Y_i)}{a_n}\right)$$

and define

$$Z_{i,n} := \begin{pmatrix} F(X_i) - F_n(X_i) \\ G(Y_i) - G_n(Y_i) \end{pmatrix}.$$

As mentioned in section 3.4.1, $|F_n(X_i) - F(X_i)| \leq ||F_n - F||_\infty$ *and* $|G_n(Y_i) - G(Y_i)| \leq ||G_n - G||_\infty$ *a.s. for every* $i = 1, \ldots, n$*. Lemma 3.13 thus entails that the norm of* $Z_{i,n}$ *is independent of* i *and such that*

$$||Z_{i,n}|| = O_P(1/\sqrt{n}) \ , \ i = 1, \ldots, n \tag{3.11}$$

$$||Z_{i,n}|| = O_{a.s.}(\sqrt{\ln \ln n / n}) \ , \ i = 1, \ldots, n \tag{3.12}$$

Now, for every fixed $(u, v) \in [0, 1]^2$*, since the kernel* K *is twice differentiable, there exists, by Taylor expansion, random variables* $\tilde{U}_{i,n}$ *and* $\tilde{V}_{i,n}$ *such that, almost surely,*

$$\begin{aligned} \Delta &= \frac{1}{na_n^3} \sum_{i=1}^{n} Z_{i,n}^T \nabla K \left(\frac{u - F(X_i)}{a_n}, \frac{v - G(Y_i)}{a_n} \right) \\ &+ \frac{1}{2na_n^4} \sum_{i=1}^{n} Z_{i,n}^T \nabla^2 K \left(\frac{u - \tilde{U}_{i,n}}{a_n}, \frac{v - \tilde{V}_{i,n}}{a_n} \right) Z_{i,n} \\ &:= \Delta_1 + \Delta_2 \end{aligned}$$

where $Z_{i,n}^T$ *denotes the transpose of the vector* $Z_{i,n}$ *and* ∇K *and* $\nabla^2 K$ *the gradient and the Hessian respectively of the multivariate kernel function* K

$$\nabla K = \begin{pmatrix} \frac{\partial K}{\partial u} \\ \frac{\partial K}{\partial v} \end{pmatrix} \ , \ \nabla^2 K = \begin{pmatrix} \frac{\partial^2 K}{\partial u^2} & \frac{\partial^2 K}{\partial u \partial v} \\ \frac{\partial^2 K}{\partial u \partial v} & \frac{\partial^2 K}{\partial v^2} \end{pmatrix}$$

By centering at expectations, decompose further the first term Δ_1 *as,*

$$\begin{aligned} \Delta_1 &= \frac{1}{na_n^3} \sum_{i=1}^{n} Z_{i,n}^T \left(\nabla K \left(\frac{u - F(X_i)}{a_n}, \ldots \right) - E \nabla K \left(\frac{u - F(X_i)}{a_n}, \ldots \right) \right) \\ &+ \frac{1}{na_n^3} \sum_{i=1}^{n} Z_{i,n}^T E \nabla K \left(\frac{u - F(X_i)}{a_n}, \frac{v - G(Y_i)}{a_n} \right) \\ &:= \Delta_{11} + \Delta_{12} \end{aligned}$$

We again decompose one step further Δ_{11}: set

$$A_i = \nabla K\left(\frac{u - F(X_i)}{a_n}, \dots\right) - E\nabla K\left(\frac{u - F(X_i)}{a_n}, \dots\right)$$

Then,

$$\begin{aligned}|\Delta_{11}| &\leq \frac{||Z_n||}{na_n^3}\sum_{i=1}^{n}\left(||A_i|| - E||A_i||\right) + \frac{||Z_n||}{na_n^3}\sum_{i=1}^{n}E||A_i|| \\ &:= \Delta_{111} + \Delta_{112}\end{aligned}$$

We now proceed to the study of the order of each terms in the previous decompositions.

- *Negligibility of Δ_2:*

  By the boundedness assumption on the second-order derivatives of the kernel, and equations 3.11 and 3.12,

  $$\Delta_2 = O_P\left(\frac{1}{na_n^4}\right) \text{ and } \Delta_2 = O_{a.s.}\left(\frac{\ln\ln n}{na_n^4}\right).$$

- *Negligibility of Δ_{12}:*

  Bias results on the bivariate gradient kernel estimator (See Scott [119] chapter 6) entail that

  $$E\nabla K\left(\frac{u - F(X_i)}{a_n}, \frac{v - G(Y_i)}{a_n}\right) = a_n^3\nabla c\,(u, v) + O(a_n^5)$$

  Cauchy-Schwarz inequality yields that

  $$|\Delta_{12}| \leq \frac{n||Z_{i,n}||}{na_n^3}\left\|E\nabla K\left(\frac{u - F(X_i)}{a_n}, \frac{v - G(Y_i)}{a_n}\right)\right\|$$

  In turn, with equations 3.11 and 3.12,

  $$\Delta_{12} = O_P(1/\sqrt{n}) \text{ and } \Delta_{12} = O_{a.s}(\sqrt{\ln\ln n/n}).$$

- *Negligibility of Δ_{11}:*

  - *Negligibility of Δ_{111}*

    Boundedness assumption on the derivative of the kernel imply that $||A_i|| \leq 2C$

a.s. We apply Hoeffding inequality for independent, centered, bounded by M, but non identically distributed random variables (η_j) (e.g. see [21]),

$$P(\sum_{j=1}^{n} \eta_j > t) \leq \exp\left(-\frac{t^2}{2nM^2}\right). \tag{3.13}$$

Here, for every $\epsilon > 0$, with $M = 2C$, $\eta_i = ||A_i|| - E||A_i||$, $t = \epsilon n^{1/2}(\ln \ln n)^{1/2}$, we get that

$$P\left(\sum_{i=1}^{n} (||A_i|| - E||A_i||) > \epsilon\sqrt{n \ln \ln n}\right) \leqslant \exp\left(-\frac{\epsilon^2 \ln \ln n}{4M^2}\right) = \frac{1}{(\ln n)^{\delta}}$$

with a $\delta > 0$ and where the r.h.s. goes to zero as $n \to \infty$. Therefore, $\sum_{i=1}^{n} (||A_i|| - E||A_i||) = O_P(\sqrt{n \ln \ln n})$.

For the almost sure negligibility, we get similarly by inequality 3.13 that, for every $\epsilon > 0$, with $t = \epsilon\sqrt{n \ln n}$,

$$P\left(\sum_{i=1}^{n} (||A_i|| - E||A_i||) > \epsilon\sqrt{n \ln n}\right) \leq n^{-\epsilon^2/4M^2}$$

and the series on the r.h.s is convergent for a $\epsilon > 2M$. Consequently, there exists an $\epsilon > 0$ such that

$$\sum_{n \in \mathbb{N}} P\left(\sum_{i=1}^{n} (||A_i|| - E||A_i||) > \epsilon\sqrt{n \ln n}\right) < \infty$$

which is the definition of almost complete convergence (a. co.), see e.g. [57] definition A.3. p. 230. In turn, it means that

$$\sum_{i=1}^{n} (||A_i|| - E||A_i||) = O_{a.co.}(\sqrt{n \ln n})$$

and by the Borell-Cantelli lemma,

$$\sum\nolimits_{i=1}^{n} (||A_i|| - E||A_i||) = O_{a.s.}(\sqrt{n \ln n}).$$

Therefore, using equations 3.11 and 3.12, we have that

$$\Delta_{111} = O_P\left(\frac{\sqrt{\ln \ln n}}{na_n^3}\right) \text{ and } \Delta_{111} = O_{a.s.}\left(\sqrt{\ln n}\frac{\sqrt{\ln \ln n}}{na_n^3}\right)$$

– *Negligibility of Δ_{112}*

It remains to evaluate $E||A_i||$. First, we have that

$$E||A_i|| \leq 2E||\nabla K((u - F(X_i))/a_n, \ldots)||$$

Second, since K is differentiable and of product form $K(u, v) = K_1(u)K_2(v)$, each sub-kernel is of bounded variations and can be written as a difference of two monotone increasing functions. For example, set $K_1 = K_1^a - K_1^b$ and define $K^* := (K_1^a + K_1^b)K_2$. We have,

$$\left|\frac{\partial K}{\partial u}\right| \leqslant \left(|(K_1^a)'| + |(K_1^b)'|\right) K_2 = ((K_1^a)' + (K_1^b)')K_2 := \frac{\partial K^*}{\partial u}$$

where the equality proceeds from the positivity of the derivatives. As a consequence,

$$E\left|\frac{\partial K}{\partial u}((u - F(X_i))/a_n, \ldots)\right| \leq E\frac{\partial K^*}{\partial u}((u - F(X_i))/a_n, \ldots)$$

and similarly for the other partial derivative. The r.h.s. of the previous inequality is, after an integration by parts, of order a_n^3 by the results on the kernel estimator of the gradient of the density (See Scott [119] chapter 6). Therefore, $\sum_{i=1}^n E||A_i|| = O(na_n^3)$, and

$$\Delta_{112} = \frac{||Z_n||}{na_n^3} \sum_{i=1}^n E||A_i|| = O_P(n^{-1/2}) \text{ or } O_{a.s.}((\ln\ln n/n)^{1/2})$$

by equations 3.11 and 3.12.

Recollecting all elements, we eventually obtain that

$$\begin{aligned}
\Delta &= \Delta_{111} + \Delta_{112} + \Delta_{12} + \Delta_2 \\
&= O_P\left(n^{-1/2}\right) + O_P\left(\frac{\sqrt{\ln\ln n}}{na_n^3}\right) + O_P\left(\frac{1}{na_n^4}\right) \\
\text{or } &= O_{a.s.}\left(\left(\frac{\ln\ln n}{n}\right)^{1/2}\right) + O_{a.s.}\left(\frac{(\ln n)^{1/2}(\ln\ln n)^{1/2}}{na_n^3}\right) + O_{a.s.}\left(\frac{\ln\ln n}{na_n^4}\right)
\end{aligned}$$

□

3.4.5 An approximation of $\hat{c}_n(F_n(x), G_n(y))$ by $\hat{c}_n(F(x), G(y))$

The proposition of this subsection gives the rate of deviation of the kernel copula density estimator $\hat{c}_n$ from a varying location $(F_n(x), G_n(y))$ to a fixed location $(F(x), G(y))$.

Proposition 3.17 *With the same assumptions as in the preceding proposition, we have*

- *if* $a_n \to 0$, $na_n^3 \to \infty$,

$$\hat{c}_n(F_n(x), G_n(y)) - \hat{c}_n(F(x), G(y)) = O_P\left(\frac{1}{\sqrt{n}} + \frac{1}{na_n^4}\right),$$

- *if* $a_n \to 0$, $na_n^3 / \ln\ln n \to \infty$,

$$\hat{c}_n(F_n(x), G_n(y)) - \hat{c}_n(F(x), G(y)) = O_{a.s}\left(\sqrt{\frac{\ln\ln n}{n}} + \frac{1}{na_n^4}\right).$$

Proof *We proceed similarly as in the preceding proposition. Set*

$$\Delta'(x,y) := \hat{c}_n(F_n(x), G_n(y)) - \hat{c}_n(F(x), G(y)) = \frac{1}{na_n^2}\sum_{i=1}^{n} \Delta'_{i,n}(x,y) \tag{3.14}$$

with

$$\begin{aligned}\Delta'_{i,n}(x,y) := & K\left(\frac{F_n(x) - F_n(X_i)}{a_n}, \frac{G_n(y) - G_n(Y_i)}{a_n}\right) \\ & - K\left(\frac{F(x) - F_n(X_i)}{a_n}, \frac{G(y) - G_n(Y_i)}{a_n}\right)\end{aligned}$$

and define

$$Z_n(x,y) := \begin{pmatrix} F_n(x) - F(x) \\ G_n(y) - G(y) \end{pmatrix}.$$

We first expand $\Delta'_{i,n}(x,y)$ to a fixed location $(F(x), G(y))$ by a Taylor expansion and bounding uniformly the second order terms,

$$\Delta'_{i,n}(x,y) = Z_n^T(x,y)\frac{\nabla K}{a_n}\left(\frac{F(x) - F_n(X_i)}{a_n}, \frac{G(y) - G_n(Y_i)}{a_n}\right) + \frac{||Z_n||_\infty^2}{a_n^2} R_3 \tag{3.15}$$

where R_3 is uniformly bounded almost surely, by assumptions on the second order derivatives of the kernel: $R_3 = O_{a.s.}(1)$.

We then expand the gradient from the data $(F_n(X_i), G_n(Y_i))$ to the pseudo but fixed w.r.t. n data $(F(X_i), G(Y_i))$: by a second Taylor expansion, there exists random variables $\tilde{U}_{i,n}$,

$\tilde{V}_{i,n}$ such that

$$\begin{aligned}
&\nabla K\left(\frac{F(x)-F_n(X_i)}{a_n}, \frac{G(y)-G_n(Y_i)}{a_n}\right) \\
&= \nabla K\left(\frac{F(x)-F(X_i)}{a_n}, \frac{G(y)-G(Y_i)}{a_n}\right) \\
&+ Z_{i,n}^T \frac{\nabla^2 K}{a_n}\left(\frac{F(x)-\tilde{U}_{i,n}}{a_n}, \frac{G(y)-\tilde{V}_{i,n}}{a_n}\right). \qquad (3.16)
\end{aligned}$$

Plugging (3.15) and (3.16) in (3.14), we get

$$\begin{aligned}
\Delta'(x,y) &= \frac{Z_n^T(x,y)}{na_n^3}\sum_{i=1}^n \nabla K\left(\frac{F(x)-F(X_i)}{a_n}, \frac{G(y)-G(Y_i)}{a_n}\right) \\
&+ \frac{Z_n^T(x,y)}{na_n^4}\sum_{i=1}^n Z_{i,n}^T \nabla^2 K\left(\frac{F(x)-\tilde{U}_{i,n}}{a_n}, \frac{G(y)-\tilde{V}_{i,n}}{a_n}\right) \\
&+ \frac{||Z_n||_\infty^2}{a_n^4} R_3 \\
&:= \Delta'_1 + \Delta'_2 + \Delta'_3
\end{aligned}$$

As in equation (3.15), bounding uniformly the Hessian, we get that

$$|\Delta'_2| \leq \frac{||Z_n||^2}{a_n^4} R_2$$

where $R_2 = O_{a.s.}(1)$ uniformly. Similarly to proposition 3.16, the consistency properties of the kernel estimator of the derivative of the density (See Scott [119] chapter 6) entails with $a_n \to 0$ and $na_n^3 \to \infty$ that

$$R_1 := \frac{1}{na_n^3}\sum_{i=1}^n \nabla K\left(\frac{F(x)-F(X_i)}{a_n}, \frac{G(y)-G(Y_i)}{a_n}\right) = O_P(1),$$

and with $a_n \to 0$, $na_n^3/\ln\ln n \to \infty$ that

$$R_1 := \frac{1}{na_n^3}\sum_{i=1}^n \nabla K\left(\frac{F(x)-F(X_i)}{a_n}, \frac{G(y)-G(Y_i)}{a_n}\right) = O_{a.s.}(1).$$

Therefore, equation (3.14) becomes

$$|\Delta'(x,y)| \leq ||Z_n||_\infty R_1 + \frac{||Z_n||_\infty^2}{a_n^4}(R_2 + R_3)$$

Combined with equations (3.11) and (3.12), we thus obtain that

$$\Delta'(x,y) = O_P\left(\frac{1}{\sqrt{n}} + \frac{1}{na_n^4}\right)$$
$$or \ = O_{a.s}\left(\sqrt{\frac{\ln\ln n}{n}} + \frac{1}{na_n^4}\right).$$

□

3.5 Uniform consistency results

The weak and strong consistency results of section 3.3 can be extended to hold uniformly on a given compact set, as shown in the next subsection. To that end, classical results on the uniform convergence of the kernel density estimators and uniform approximation propositions similar to those of section 3.4 are required. Those results are presented and established afterwards in subsection 3.5.2 and 3.5.3 respectively.

3.5.1 Uniform consistency of the conditional density estimator

Theorem 3.18 *Let the regularity conditions* ***A*** *(i)-(iii) and* ***B*** *(i)-(iv) be satisfied. If* $h_n \simeq (\ln n/n)^{1/5}$ *and* $a_n \simeq (\ln n/n)^{1/6}$*, then, for* x *in the interior of supp*(f) *and* $[a,b]$ *included in the interior of supp*(g)*,*

$$\sup_{y\in[a,b]} |\hat{f}_n(y|x) - f(y|x)| = O_p\left(\left(\frac{\ln n}{n}\right)^{1/3}\right),$$

and

$$\sup_{y\in[a,b]} \left|\hat{f}_n(y|x) - f(y|x)\right| = O_{a.s.}\left(\left(\frac{\ln n}{n}\right)^{1/3}\right).$$

Proof *The proof is identical to the ones of theorems 3.5 and 3.7, but uses propositions 3.24 and 3.23 below instead of propositions 3.16 and 3.17, and uniform consistency results of the kernel density estimators of theorems 3.20 and 3.21 below applied to* g_n *and* c_n*.*

3.5.2 Uniform consistency of the kernel density estimators

Similarly to section 3.4.2, we recall below some classical results of convergence of the kernel density estimators uniformly on sets. In the following f denotes a generic density on $\mathbb{R}^d$.

Bias

If f is supposed to be twice differentiable with second partial derivatives uniformly bounded on J, the bias is also uniformly bounded on J: indeed,

$$\sup_{t\in J} |Ef_n(t) - f(t)| = h_n^2/2 \int K(y) y^T \{f''(y)\} y dy + o(h_n^2)$$

where

$$f''(t) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\bigg|_{x=t} \right)$$

is a shorthand for the Hessian of f, and where the $o(.)$ is independent of t. This comes from a so-called uniform Bochner type theorem:

Lemma 3.19 (Bochner) *Let $K : \mathbb{R}^d \mapsto \mathbb{R}$ be a convolution kernel, i.e. $K \in \mathbb{L}_1$ w.r.t Lebesgue measure λ, with $\int K d\lambda = 1$, and set $K_h(x) = h^{-1}K(x/h)$, $x \in \mathbb{R}^d$. If a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is uniformly continuous and in $\mathbb{L}_1$, then*

$$\lim_{h\to 0} \sup_{x\in\mathbb{R}^d} |f * K_h(x) - f(x)| = 0$$

Proof *See e.g. Stein [123] or Bosq and Lecoutre [26].*

Uniform convergence in probability

One can refine the pointwise consistency results of the kernel density estimator by a chaining argument to get uniform convergence on a compact set. The following theorem is a direct corollary of Bickel and Rosenblatt's [16] convergence result of the norm of the deviation of the kernel density estimator to a double exponential law:

Theorem 3.20 (Bickel and Rosenblatt) *For f bounded and non-vanishing on a compact subset J included in the interior of supp(f), and a bandwidth sequence $h_n \to 0$, such that $nh_n^d \to \infty$, $nh_n^d / \ln n \to \infty$,*

$$\sup_{x \in J} \left| \hat{f}_n(x) - E\hat{f}_n(x) \right| = O_p \left[\left(\frac{\ln n}{nh_n^d} \right)^{1/2} \right].$$

Therefore, for the choice of the bandwidth $h_n \simeq (\ln n/n)^{1/d+4}$ which realises the optimal trade-off between the bias and variance, one gets, by combining this result with Bochner's lemma 3.19 above, the following result in probability:

$$\sup_{x \in J} \left| \hat{f}_n(x) - f(x) \right| = O_p \left[\left(\frac{\ln n}{n} \right)^{2/(d+4)} \right]$$

which is the optimal speed in the minimax sense in the class of density functions with bounded second derivatives, according to Hasminskii [72].

Uniform almost sure convergence

Geffroy [59] and Bertrand-Retali [15] give necessary and sufficient conditions for the strong uniform consistency of the kernel density estimator (See also Bosq and Lecoutre [26]). We cite Stute's [127, 129] theorem along with the following rates:

Theorem 3.21 (Stute) *Let J be a compact subset of $\mathbb{R}^d$, included in the support of f.*

i) If the kernel K is of bounded support, and of finite variation (e.g. if K has bounded partial derivative of order two),

ii) if the density f is uniformly continuous on J, is bounded away from zero and infinity on J: $0 \leq m < f|_J < +\infty$,

iii) if the marginal densities f_i of f, $i = 1, \ldots, d$ are bounded away from zero and infinity on J,

iv) if the bandwidth $h_n \to 0$ satisfy $nh_n^d \to +\infty$, $\ln(1/h_n^d) = o(nh_n^d)$, and $\ln(1/h_n^d)/(\ln \ln n) \to +\infty$

then, with probability one,

$$\lim_{n\to\infty} \sup_{t\in J} \sqrt{\frac{nh_n^d}{2\ln h_n^{-d}}} \left| \frac{f_n(t) - Ef_n(t)}{\sqrt{f(t)}} \right| = \left(\int K^2 \right)^{1/2}$$

Remark 3.22 *if the last condition on the bandwidth is suppressed, the theorem remains valid with* lim *replaced with* $\overline{\lim}$*. With the usual choice of bandwidth* $h_n \simeq (\ln n/n)^{1/(d+4)}$ *to deal with the bias, one gets the almost sure uniform convergence of the kernel density estimator at the rate* $(\ln n/n)^{2/(d+4)}$*.*

3.5.3 Two Uniform approximation propositions

We present analogue extensions of propositions 3.16 and 3.17 to uniformity on a compact set. In order to get uniformity on sets and to be able to make use of theorems 3.20, 3.21 and lemma 3.19 above, the regularity and boundedness assumptions on the density g and c and their derivatives of section 3.3.1 will have to be slightly strengthened as follows:

Assumption A

(iii) Suppose the density g anc c are uniformly continuous and non-vanishing almost everywhere on a compact set $J := [a, b]$ and $D \subset (0,1) \times (0,1)$ included in the interior of $\text{supp}(g)$ and $\text{supp}(c)$, respectively.

Proposition 3.23 *Let the regularity assumptions* ***A*** *and* ***B*** *be satisfied, then, for a compact set* $D \subset (0,1)^2$*,* $a_n \to 0$ *and* $na_n^3/\ln n \to \infty$ *entails*

$$\sup_{(x,y)\in D} |\hat{c}_n(F_n(x), G_n(y)) - \hat{c}_n(F(x), G(y))| = O_P\left(\frac{1}{na_n^4} + \frac{\ln n}{n^{1/2}}\right)$$

$$\sup_{(x,y)\in D} |\hat{c}_n(F_n(x), G_n(y)) - \hat{c}_n(F(x), G(y))| = O_{a.s}\left(\frac{\ln\ln n}{na_n^4} + \frac{\ln n(\ln\ln n)^{1/2}}{n^{1/2}}\right)$$

Proof *We proceed as in the proof of proposition 3.17. Set*

$$W_{i,n}(u,v) := \nabla K\left(\frac{u - F(X_i)}{a_n}, \frac{v - G(Y_i)}{a_n}\right)$$

By Taylor expansions, we still have the decomposition

$$\begin{aligned}\Delta_n(x,y) = &\frac{Z_n^T(x,y)}{na_n^3}.\sum_{i=1}^n W_{i,n}(F(x),G(y)) \\ &+ \frac{Z_n^T(x,y)}{2na_n^4}\sum_{i=1}^n Z_{i,n}^T(x,y).\nabla^2 K\left(\frac{u-F(X_i)}{a_n},\frac{v-G(Y_i)}{a_n}\right) \\ &+ \frac{||Z_n||^2}{a_n^4}R_3\end{aligned}$$

with the remainder term $R_3 = O_{a.s.}(1)$ *uniformly. By bounding the* $\nabla^2 K$*, and using the properties of the Kolmogorov-Smirnov statistic, the last two terms are of order* $O_P\left(\frac{1}{na_n^4}\right)$*, or* $O_{a.s}\left(\frac{\ln\ln n}{na_n^4}\right)$*, uniformly in* x*,* y*. For the first term, by Cauchy-Schwarz inequality,*

$$\sup_{(x,y)\in D}|\frac{Z_n^T(x,y)}{na_n^3}.\sum_{i=1}^n W_{i,n}(F(x),G(y))| \leq ||Z_n|| \sup_{(x,y)\in D}\left\|\frac{1}{na_n^3}\sum_{i=1}^n W_{i,n}(F(x),G(y))\right\|$$

The convergence results of the kernel estimator $n^{-1}a_n^{-3}\sum_{i=1}^n W_{i,n}(u,v)$ *of the gradient of the density* $c(u,v)$ *can easily be derived from those of the kernel estimator (see Scott [119]). From the convergence results uniformly on a compact set of the latter obtained by e.g. Deheuvels [41] for the almost sure rates and Bickel and Rosenblatt [16] for the in probability rates, with the assumption that the gradient is uniformly bounded on* D *and that* $na_n^3/\ln n \to \infty$*, one gets that the uniform norm of the estimator of the gradient is an* $O_P(\ln n)$ *or an* $O_{a.s.}(\ln n)$ *. In turn,* $\sup_{(x,y)\in D}|\Delta_n(x,y)| = O_P(\ln n/n^{-1/2})$ *or* $O_{a.s.}(\ln n(\ln\ln n/n)^{1/2})$*. Thus the claimed result.*

Proposition 3.24 *Let the regularity assumptions* ***A*** *and* ***B*** *be satisfied, then, for a compact set* $D \subset (0,1)^2$*, and a bandwidth such as* $a_n \simeq \left(\frac{\ln n}{n}\right)^{1/6}$*, one has*

$$\sup_{(u,v)\in D}|\hat{c}_n(u,v) - c_n(u,v)| = O_{a.s.}\left(\left(\frac{\ln n}{n}\right)^{1/3}\right) \text{ or } O_P\left(\left(\frac{\ln n}{n}\right)^{1/3}\right)$$

Proof *For convenience, set* $||(x_1,\ldots,x_d)|| = \max_{1\leq j\leq d}|x_j|$*. Set* $D = [u_0,u_\infty]\times[v_0,v_\infty] \subset (0,1)^2$ *a compact subset where* $0 < u_0 \leq u_\infty < 1$ *and* $0 < v_0 \leq v_\infty < 1$*. We mimic the*

proof of proposition 3.16. We still have the additive decomposition,

$$\begin{aligned}\Delta(u,v) &= \Delta_1(u,v) + \Delta_2(u,v) \\ &= \Delta_{11}(u,v) + \Delta_{12}(u,v) + \Delta_2(u,v)\end{aligned}$$

with

$$\Delta_{11}(u,v) = \frac{1}{na_n^3}\sum_{i=1}^{n} Z_{i,n}.\left(W_{i,n}(u,v) - EW_{i,n}(u,v)\right)$$
$$\Delta_{12}(u,v) = \frac{1}{na_n^3}\sum_{i=1}^{n} Z_{i,n}.EW_{i,n}(u,v)$$

- ***Negligibility of*** Δ_2

  The proof remains the same:

  $$\sup_{(u,v)\in D}|\Delta_2(u,v)| = O_P(n^{-1}a_n^{-4}),\ \ or\ O_{a.s.}((\ln n)/(na_n^4)).$$

- ***Negligibility of*** Δ_{12}

  Recall that in the Taylor's expansion of the bias of the kernel estimator, the $O(.)$ is uniform in (u,v), therefore one gets that

  $$\sup_{(u,v)\in D}||EW_{i,n}(u,v) - a_n^3\nabla c(u,v)|| = O(a_n^5)$$

  Thus,

  $$\sup_{(u,v)\in D}|\Delta_{12}(u,v)| = O_P(1/\sqrt{n}),\ \ or\ O_{a.s.}((\ln\ln n/n)^{1/2}).$$

- ***Negligibility of*** Δ_{11}

  Define a covering of D by M_n^2 compact hypercubes D_k centered in (u_k,v_k),

  $$D_k = \{(u,v)\in D : ||(u,v)-(u_k,v_k)|| \le 1/M_n\}\ ,\ 1\le k\le M_n^2$$

  with

  $$\overset{\circ}{D}_k \cap \overset{\circ}{D}_{k'} = \emptyset\ ,\ 1\le k\ne k'\le M_n^2$$

One can write

$$\sup_{(u,v)\in D} |\Delta_{11}(u,v)| \leq \max_{1\leq k\leq M_n^2} \sup_{(u,v)\in D_k} |\Delta_{11}(u,v) - \Delta_{11}(u_k,v_k)|$$
$$+ \max_{1\leq k\leq M_n^2} |\Delta_{11}(u_k,v_k)|$$
$$:= (I) + (II)$$

- ***Negligibility of*** (I)

  For (I)*, by boundedness and Lipshitz assumption on the product kernel* K*, there exists a constant* C *such that,*

  $$||\nabla K(u,v) - \nabla K(u_k,v_k)|| \leq C||(u,v) - (u_k,v_k)||$$

  Therefore for $(u,v) \in D_k$*,*

  $$\left\|\nabla K\left(\frac{u-F(X_i)}{a_n}, \frac{v-G(Y_i)}{a_n}\right) - \nabla K\left(\frac{u_k-F(X_i)}{a_n}, \frac{v_k-G(Y_i)}{a_n}\right)\right\| \leq \frac{C}{M_n a_n}$$

  since K *is product-shaped. In turn, the same bound is valid by Jensen's inequality for the expectations of the difference, so that*

  $$(I) \leq \frac{2C||Z_n||}{M_n a_n^4}$$

  Setting $M_n = n^{1/2}a_n^{-3} \simeq n/\sqrt{\ln n}$ *for* $a_n \simeq (\ln n/n)^{1/6}$*, one has that* $(I) = o_{a.s.}\left(\sqrt{\frac{\ln n}{na_n^2}}\right)$ *or* $o_P((na_n^2)^{-1/2})$*.*

- ***Negligibility of*** (II)

  For the second term, set as before, $A_i(u,v) = W_{i,n}(u,v) - EW_{i,n}(u,v)$*, and majorize, for each* k*,*

  $$\begin{aligned}|\Delta_{11}(u_k,v_k)| &\leq \frac{||Z_n||}{na_n^3}\sum_{i=1}^n ||A_i(u_k,v_k)|| \\ &\leq \frac{||Z_n||}{na_n^3}\sum_{i=1}^n (||A_i(u_k,v_k)|| - E||A_i(u_k,v_k)|| + E||A_i(u_k,v_k)||) \\ &\leq \frac{||Z_n||}{na_n^3}\sum_{i=1}^n \eta_i(u_k,v_k) + \frac{||Z_n||}{na_n^3}\sum_{i=1}^n E||A_i(u_k,v_k)||\end{aligned}$$

where we have set $\eta_i(u_k, v_k) = ||A_i(u_k, v_k)|| - E||A_i(u_k, v_k)||$.

For the expectation term, as the product kernel is of finite variation, and with the assumption that the gradient of the copula density remains bounded on D*, one has that* $\max_{1\leq k\leq M_n^2} E||A_i(u_k, v_k)|| = O(a_n^3)$*. This yields that*

$$\max_{1\leq k\leq M_n^2} \frac{||Z_n||}{na_n^3} \sum_{i=1}^{n} E||A_i(u_k, v_k)|| = O_P(n^{-1/2}) \text{ , or } O_{a.s.}\left(\left(\frac{\ln\ln n}{n}\right)^{1/2}\right)$$

It remains to deal with the deviation term

$$\max_{1\leq k\leq M_n^2} \frac{||Z_n||}{na_n^3} \sum_{i=1}^{n} \eta_i(u_k, v_k)$$

We have

$$P\left(\max_{1\leq k\leq M_n^2} |\sum_{i=1}^{n} \eta_i(u_k, v_k)| > \varepsilon\right) \leq \sum_{k=1}^{M_n^2} P\left(|\sum_{i=1}^{n} \eta_i(u_k, v_k)| > \varepsilon\right)$$

and apply Hoeffding's inequality to the summand, to get that, for every $\varepsilon > 0$*,*

$$P\left(|\sum_{i=1}^{n} \eta_i(u_k, v_k)| > \varepsilon\sqrt{n\ln n}\right) \leq \exp\left(-\frac{\varepsilon^2 \ln n}{C}\right)$$

for a constant C *independent of* k*, which exists by the boundedness assumption on the gradient of the kernel. Thus,*

$$\begin{aligned} P\left(\max_{1\leq k\leq M_n^2} |\sum_{i=1}^{n} \eta_i(u_k, v_k)| > \varepsilon\sqrt{n\ln n}\right) &\leq M_n^2 \exp\left(-\frac{\varepsilon^2 \ln n}{C}\right) \\ &\leq \exp\left(\sqrt{2\ln M_n} - \frac{\varepsilon^2 \ln n}{C}\right) \end{aligned}$$

For $a_n \simeq (\ln n/n)^{1/6}$ *and* $M_n = n^{1/2}a_n^{-3} \simeq n/\sqrt{\ln n}$*,*

$$\exp\left(\sqrt{2\ln M_n} - \frac{\varepsilon^2 \ln n}{C}\right) \approx \exp\left(-\frac{\varepsilon^2 \ln n}{C}\right) = \frac{1}{n^{\varepsilon^2/C}}$$

which is absolutely summable for an ε *large enough. Therefore,*

$$\max_{1\leq k\leq M_n^2} |\sum_{i=1}^{n} \eta_i(u_k, v_k)| = O_{a.co.}\left(\sqrt{n\ln n}\right)$$

and eventually,

$$\frac{||Z_n||}{na_n^3} \max_{1\leq k\leq M_n^2} |\sum_{i=1}^{n} \eta_i(u_k, v_k)| = O_{a.s.}\left(\frac{\sqrt{\ln n \ln\ln n}}{na_n^3}\right)$$

for the choice $a_n \simeq (\ln n/n)^{1/6}$.

Recollecting all elements gives the claimed result with the given choice of a_n.

□

Chapter 4

Discussions, implementation and comparisons

Abstract : This chapter complements the asymptotic study of the quantile-copula conditional density estimator of the preceding chapter by providing discussions related to its finite sample performance. In section 4.1, we conduct a discussion on the estimation of the marginals and present some possible modifications of the estimator by using substitutes for the empirical c.d.f. In particular, based on efficiency considerations, we advocate the use of smoothed estimators of the empirical c.d.f., such as the kernel or Bernstein polynomial one. We then establish some heuristic connections with the related local bandwidth kernel estimators. At last, we show how to modify the proposed estimator to take into account additional information on the explanatory variable, i.e. when its marginal distribution is parametric. In section 4.2, we investigate the possible issues involved in the implementation of the estimator in a small sample setting. In particular, in addition to possible infinities of the copula densities at the corners if the empirical c.d.f. are used, we discuss on methods to decrease the bias and choose the bandwidths. In section 4.3, we perform a simulation study and compare, both theoretically and numerically, the per-

formance of the proposed estimator to its competitors. We then hint a possible area of special interest.

4.1 A discussion on the marginals and suggested modifications of the estimator

In this section, we conduct a theoretically-motivated discussion on the estimation of the marginal cumulative distribution functions F and G and suggest some modifications of the conditional density estimator.

4.1.1 On the asymptotic efficiency of the empirical transformations

Recall that the proposed conditional density estimator writes

$$\hat{f}(y|x) = \hat{g}(y)\hat{c}_n(F_n(x), G_n(y))$$

where the copula-related part of the estimator

$$\hat{c}_n(F_n(x), G_n(y)) = (na_nb_n)^{-1} \sum_{i=1}^{n} K_1\left(\frac{F_n(X_i) - F_n(x)}{a_n}\right) K_2\left(\frac{G_n(Y_i) - G_n(y)}{b_n}\right)$$

is the estimator of the copula density, evaluated at the approximate point $(F_n(x), G_n(y))$, based on the approximate sample $(F_n(X_i), G_n(Y_i))$. Since this copula density part of the estimator works doubly with an approximation of the true c.d.f. F and G, it is of great interest to assess the impact of the choice of the estimators of F and G on the performance of the final estimator. This is the purpose of the present subsection, which thereby provides suggested modifications $\hat{F}_n$ and $\hat{G}_n$.

A reminder on Uniform Unbiased Minimum Variance (UMVU) estimation

A discussion of the choice of the statistics to use in the quantile-copula estimator can be based on the concepts of sufficient statistics, in the spirit of Lehmann and Casella [92]. Indeed, recall that the sufficient statistic for estimating a continuous cumulative distribution function F is the set of order statistics $T = (X_{1,n} < \ldots < X_{n,n})$ and that T is complete for the set of distributions F which admits a density. Therefore, one can show that the efficient estimator of F, in the sense of the Unbiased Minimum Variance (UMVU), obtained by means of the Rao-Blackwellization device, is the empirical c.d.f. $F_n(.)$, see [92].

This approach therefore advocates the use of the empirical distribution functions for the estimation of the c.d.f. F and G. However, it is well known that the UMVU point of view, which consists in reducing the class of all possible estimators to those which are unbiased estimators and searching within this subclass those with the minimum variance, may be questionable. Indeed, apart from the fact that no unbiased estimators may exist in some situations, one can see from the bias and variance decomposition of an estimator T of a given function of the parameter $g(\theta)$,

$$E_\theta[T - g(\theta)]^2 = (E_\theta[T] - g(\theta))^2 + Var_\theta[T]$$

that an overall minimum of this sum of two functions can be reached without being either a minimum of the first term (zero bias) or of the second term (zero variance). In other words, there may also exist nontrivial biased estimators, but with an overall error measured in the Mean Square sense smaller than that of the UMVU estimator: it is known e.g., since as long as Aggarwal [5], that, for a continuous F, in the class of estimators invariant under the group of one to one, monotone transformations of the real numbers onto themselves which leaves the sample values X_i, $i = 1, \ldots, n$, invariant, the following

risk

$$R(\hat{F}, F) := E \int (\hat{F}(x) - F(x))^2 dF(x)$$

is minimized by the following step estimator

$$F_n^B(x) := \frac{nF_n(x) + 1}{n + 2}$$

In particular, F_n^B is minimax invariant and dominates the empirical distribution function F_n.

This is also precisely the case for smooth estimates of the c.d.f. where, in this nonparametric setup, one has to admit bias, as is shown below.

Asymptotic efficiency of the kernel smoothed estimator of the c.d.f.

For that purpose, let's introduce a kernel smoothed variant of the empirical distribution function by,

$$\hat{F}(x) = n^{-1} \sum_{i=1}^{n} L\left(\frac{X_i - x}{h}\right) \tag{4.1}$$

where L is a c.d.f obtained from the probability density function l, by $L(x) = \int_{-\infty}^{x} l(z)dz$. That is to say, $\hat{F}$ is the primitive of the nonparametric density estimator $\hat{f}$ with kernel l and bandwidth h, $\hat{F}(x) = \int_{-\infty}^{x} \hat{f}(y)dy$. Assuming classical regularity conditions, this smoothed c.d.f estimator $\hat{F}$ has the following asymptotic properties (see e.g. Li and Racine [93], Reiss [109], Azzalini [10]),

$$E\hat{F}(x) = F(x) + 1/2 m_2(K) h^2 F^{(2)}(x) + o(h^2)$$
$$Var[\hat{F}(x)] = n^{-1} F(x)[1 - F(x)] - h n^{-1} f(x) 2 \int tL(t)l(t)dt + o(hn^{-1})$$

Regarding rates of convergence, note that the situation is very different from kernel density estimation : the smoothed empirical distribution function retains the same $\sqrt{n}$ rate of convergence as its unsmoothed cousin. Indeed, although this estimator is smoothed, the bias introduced by the smoothing process does not deteriorate the $\sqrt{n}$ consistency property of the unsmoothed estimator, for a choice of the bandwidth optimal in the MISE

sense of order $h \simeq n^{-1/3}$. To compare its efficiency to its unsmoothed counterpart F_n, one can combine the two previous equations and evaluate the difference of the Mean Square Error of the two estimators as in [110] p.263,

Theorem 4.1 (Reiss) *Assume l is a differentiable kernel of compact support $[-1,1]$, with vanishing first moment and F has 2 derivatives such that, in a neighbourhood of x, $|F^{(2)}| \leq C$, then, uniformly over the bandwidth $h \in (0,1)$,*

$$\left| E(\hat{F}(x) - F(x))^2 - E(F_n(x) - F(x))^2 + 2h/nF'(x) \int tl(t)L(t)dt \right| \leq h^4 AC^2 + O(h^2/n) \tag{4.2}$$

where A is constant depending on the kernel l.

Following the discussion of Reiss [110] p.263, the kernel estimator thus has a lower error in the Mean square error sense, for a small enough bandwidth h and a n large enough, as soon as $\int xl(x)L(x)dx > 0$; a condition which is automatically fulfilled for a non-negative, symmetric kernel. Further analysis can be carried on in terms of the Hodges and Lehmann's deficiency concept (see [75]): introduce

$$i(n) = \min\{m : E(F_m(x) - F(x))^2 \leq E(\hat{F}(x) - F(x))^2\}$$

that is to say the smallest integer m such that F_m has the same or better performance than $\hat{F}$. The previous inequality 4.2 states that $i(n)/n \to 1$ and $i(n) - n \to \infty$. In other words, first order comparison states that one has the same asymptotic efficiency between the kernel estimator of the d.f. and the empirical d.f., whereas second order comparison entails that F_n is asymptotically deficient w.r.t. $\hat{F}_n$; as $i(n) - n$ stands for the number of observations being wasted if we use the sample c.d.f. instead of the kernel estimator.

Implication for the conditional density estimator

As a consequence of this heuristic discussion, it seems sensible to modify the proposed conditional density estimator by replacing the empirical distribution functions F_n and G_n

by the smoothed estimates $\hat{F}$ and $\hat{G}_n$ of the distribution functions F and G. This doubly smoothed quantile copula estimator is thus defined as

$$\hat{f}(y|x) = \hat{g}(y)(na_nb_n)^{-1} \sum_{i=1}^{n} K_1 \left(\frac{\hat{F}(X_i) - \hat{F}(x)}{a_n} \right) K_2 \left(\frac{\hat{G}(Y_i) - \hat{G}(y)}{b_n} \right) \tag{4.3}$$

To justify this claim, one needs to ensure that the smoothed estimator of the c.d.f. enjoys a law of the iterated logarithm with a fast enough rate compared to those of the kernel density estimators. Uniform convergence at rate fast enough and Chung-Smirnov property of this smoothed empirical process are maintained, as shown by, e.g. [108]. Therefore, the convergence results of the doubly smoothed quantile copula conditional density estimator can be obtained without substantial modifications from its unsmoothed version, as shown in corollary 4.3 below.

We believe this modification could be beneficial from a practical point of view : if we choose a bandwidth of order $n^{-1/3}$, equation 4.2 shows that both variance and mean squared error are reduced by an amount of size $n^{-4/3}$, which, for small sample sizes, can be numerically similar to their asymptotic order, n^{-1}. Moreover, the smoothed version has a less wiggly behaviour, and the resulting conditional density estimator obtained is thus more graphically appealing.

On the negative side, one may claim that this estimator introduces yet another smoothing parameter, which thus has to be chosen, whereas the empirical distribution function is parameter free. More importantly, the desired property of deficiency is obtained by a suitable choice of the smoothing bandwidth, the latter being made with the knowledge of the regularity of the distribution function. As with classical kernel density estimation, one can partially alleviate these issues with data-dependent methods for choosing the bandwidth, such as plug-in or cross-validation. See Sarda [118] , Altman and Léger [8], Bowman et al. [27] and Tenreiro [133] for details on the strategy for choosing the bandwidths. On the positive side, note from those papers that the bandwidth selection issue for kernel c.d.f. estimation seems to be less stringent than for kernel density estimation.

Marginals with bounded support and Bernstein Polynomial Estimation of the c.d.f.

In the case where the marginal distributions F and G of X and Y have compact support, the smoothing of the empirical distribution function by the previous kernel method introduces bias on the boundaries of the support, and the estimator becomes inconsistent (see e.g. [143, 119, 121, 45]. It is therefore advisable to use a different smoothing method, which is the purpose of this paragraph.

Assume in the following, and without loss of generality (see remark 4.2 below), that the support of the c.d.f. F is $[0, 1]$. The Bernstein estimator is based on Bernstein's [14] constructive proof of the famous Weierstrass theorem: let the Bernstein's polynomial be defined as,

$$P_{k,m}(x) = C_m^k x^k (1-x)^{m-k}$$

and the Bernstein approximation of order m of a continuous function $f : [0, 1] \to \mathbb{R}$ by,

$$B_{g,m}(x) = \sum_{k=0}^{m} f(k/m) P_{k,m}(x)$$

Then, Bernstein's theorem [14] states that

$$||B_{g,m} - f||_\infty \to 0$$

as $m \to \infty$. In a stochastic setting, Babu, Canty and Chaubey [11] and Prakasa Rao [106], following the work of Vitale [139] for density estimation, introduced similarly the Bernstein polynomial estimator of F as,

$$\hat{F}_{m,n}(x) := \sum_{k=0}^{m} F_n(k/m) P_{k,m}(x) \tag{4.4}$$

with $m = m_n \to \infty$ as $n \to \infty$. Note that $\hat{F}_{m,n}$ has the following interesting features:

- it is a genuine cumulative distribution function, i.e. it is right-continuous, $0 \leq \hat{F}_{m,n}(x) \leq 1$, for $x \in [0, 1]$, and $\hat{F}_{m,n}(0) = 0$, $\hat{F}_{m,n}(1) = 1$;
- it is a polynomial of degree m and is thus computationally easy to implement;

- its smoothing parameter, the degree m of the polynomial, is an integer whereas the smoothing parameter of the kernel estimator is a real, thus simplifying the optimisation step in the tuning of the bandwidths;

- it has less variance than the kernel smoother.

Remark 4.2 *Note that the Bernstein estimator can be extended to the more general settings where the support is known to be in* $[a, b]$, $[0, \infty)$ *or even* $[-\infty, \infty]$*: by transforming the variables by e.g.* $X \leftarrow \frac{X-a}{b-a}$, $X \leftarrow \frac{X}{1+X}$, $X \leftarrow 1/2 + \pi^{-1} \tan^{-1}(X)$ *respectively, the same analysis applies on the transformed variables, as explained in subsection 4.1.2.*

As for the kernel estimator of the c.d.f, asymptotic bias and variance expansions give rises to an optimal choice of the m_n parameter, for which the Bernstein Polynomial estimator dominates $F_n(x)$ asymptotically at every point in $x \in (0, 1)$. Refer to Leblanc [90] for the proof of the asymptotic deficiency of the empirical c.d.f. with respect to the Bernstein's polynomial. Moreover, it also satisfy the law of the iterated logarithm of Chung-Smirnov, [106]. Therefore, as for the kernel c.d.f. estimator, the rate of convergence in supremum norm is fast enough to validates the use of this estimator in the construction of the quantile-copula estimator, as imbedded in the following corollary.

Corollary 4.3 *Let* $\hat{F}$ *and* $\hat{G}$ *be either the kernel c.d.f. estimator (4.1), or the Bernstein c.d.f. estimator (4.4), of* F *and* G *respectively. Assume their bandwidths be chosen in such a way that the Chung-Smirnov property is verified by these estimators. Then, the conclusions of theorem 3.7 remain valid for the estimator (4.3) of the conditional density with* F_n *and* G_n *replaced by* $\hat{F}$ *and* $\hat{G}$*.*

Proof *It is identical to the proof of theorem 3.7, with* F_n *and* G_n *replaced by* $\hat{F}$ *and* $\hat{G}$*, and the Chung-Smirnov property of* F_n *and* G_n *in lemmas 3.11 and 3.12 replaced by their counterpart for* $\hat{F}$ *and* $\hat{G}$*. It is therefore omitted.*

4.1.2 On the connection with variable bandwidths density estimators

The smoothing of the distribution functions advocated in the preceding subsection gives us the occasion to make a slight digression and explore the connection between the copula part of the conditional density estimator and different variable bandwidth kernel estimates. Indeed, with the smoothed distribution function estimators $\hat{F}_n$ and $\hat{G}_n$, the copula part of the conditional density estimator writes:

$$\hat{c}_n(\hat{F}_n(x), \hat{G}_n(y)) = (na_nb_n)^{-1} \sum_{i=1}^{n} K_1\left(\frac{\hat{F}_n(X_i) - \hat{F}_n(x)}{a_n}\right) K_2\left(\frac{\hat{G}_n(Y_i) - \hat{G}_n(y)}{b_n}\right)$$

Note that this smooth modification of the empirical c.d.f. allows to make a more refined analysis of the asymptotic behaviour of the estimator: instead of bounding roughly the terms $F_n(X_i) - F(X_i)$ or $F_n(x) - F(x)$ by the infinite norm $||F_n - F||_\infty$, since $\hat{F}$ is now a differentiable function, one can make a Taylor expansion of $\hat{F}(X_i) - \hat{F}(x)$ by

$$\hat{F}(X_i) - \hat{F}(x) = \hat{f}(x)(X_i - x) + o(|X_i - x|) \tag{4.5}$$

or

$$\hat{F}(x) - \hat{F}(X_i) = \hat{f}(X_i)(x - X_i) + o(|X_i - x|) \tag{4.6}$$

or even at higher orders, assuming the required degree of smoothness. In turn, the copula density part of the estimator becomes,

$$(na_nb_n)^{-1} \sum_{i=1}^{n} K_1\left(\frac{X_i - x}{a_n/\hat{f}(X_i)}\right) K_2\left(\frac{Y_i - y}{b_n/\hat{g}(Y_i)}\right),$$

or,

$$(na_nb_n)^{-1} \sum_{i=1}^{n} K_1\left(\frac{X_i - x}{a_n/\hat{f}(x)}\right) K_2\left(\frac{Y_i - y}{b_n/\hat{g}(y)}\right),$$

using respectively 4.5 or 4.6. Both forms shows that the doubly smoothed copula estimator is approximately like a kernel estimator with an adaptive local bandwidth $a_n/\hat{f}(X_i)$ or $a_n/\hat{f}(x)$, also called the sample smoothing or balloon respectively generalised kernel estimator by Terrell and Scott [134].

This fact is to be related with the work on transformed kernel density estimators in the spirit of [77, 78, 116, 143] in the univariate case. Indeed, let T be a smooth and increasing function and set $Z_i = T(X_i)$. The density

$$f_Z(z) = \frac{f[T^{-1}(z)]}{T'[T^{-1}(z)]}$$

can be estimated by a kernel estimate $\hat{f}_Z(z) = (nh)^{-1}\sum_{i=1}^{n} K((z-Z_i)/h)$. By proceeding by back transforming the data, one gets an estimator of the density f of X as

$$\hat{f}_T(x) = (nh)^{-1}\sum_{i=1}^{n} K((T(x)-T(X_i))/h)T'(x)$$

By choosing T close to F, one can reduce the bias. With $T = \hat{F}_1$ where $\hat{F}_1(x) = \int_{-\infty}^{x} f_1(d)du$ and $\hat{f}_1$ is a kernel density estimator with bandwidth h_1, Hössjer and Ruppert [77] and Ruppert and Cline [116] show that the obtained estimator

$$\hat{f}_{\hat{F}_1}(x) = (nh)^{-1}\sum_{i=1}^{n} K((\hat{F}_1(x)-\hat{F}_1(X_i))/h)\hat{f}_1(x)$$

has the same rate of convergence as a fourth order kernel, and thus significantly reduces the bias. The smoothness of $\hat{F}_1$ allows to make a Taylor expansion of

$$\hat{F}_1(x) - \hat{F}_1(X_i)) \approx (x - X_i)\hat{f}_1(x)$$

and yields an estimator which is equivalent to a local bandwidth kernel density estimator

$$\hat{f}_{\hat{F}_1}(x) \approx (nh)^{-1}\sum_{i=1}^{n} K((x-X_i)\hat{f}_1(x)/h)\hat{f}_1(x)$$

which is itself essentially a k-nearest neighbour estimator, as noted in Silverman [121] section 5.2. Note that the copula part of the conditional density estimator does not have the multiplicative term $T'(x)$ in $\hat{f}_T(x) = (nh)^{-1}\sum_{i=1}^{n} K((T(x)-T(X_i))/h)T'(x)$. Further enquiries with the k-nearest neighbour estimator are developed in chapter 5.

Remark 4.4 *This discussion allows to motivate the use of yet another alternative estimator for the cumulative distribution functions, the Swanepoel and Van Graan [2005] and*

Janssen, Swanepoel and Van Graan [2007]'s reduced bias kernel estimators.

$$\tilde{F}_g(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{\hat{F}_g(x) - \hat{F}_g(X_i)}{h}\right)$$
$$\check{F}_{h,g}(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\hat{f}_g^{1/2}(X_i)\right)$$

where $\hat{f}_g(x) = \frac{1}{ng}\sum_{i=1}^{n} k\left(\frac{x-X_i}{g}\right)$ *is a kernel density estimator with kernel* k *and bandwidth* g*, and* $\hat{F}_g$ *its primitive.*

4.1.3 Estimator with parametric margins

The proposed estimator of the conditional density of chapter 3 is fully nonparametric, in the sense that both the estimators of the marginal density of Y, the empirical tranformations F_n and G_n, and the estimator of copula density are nonparametric. One can also consider a semi-parametric framework in which some of the marginal distributions are assumed to belong to a parametric family of cumulative distribution functions $\{F(.,\theta), \theta \in \Theta\}$, indexed by a parameter $\theta \in \Theta \subset \mathbb{R}^d$. Assume θ_0 is the true parameter. One can substitute in the quantile-copula estimator the empirical distribution function F_n by

$$\tilde{F}_n(x) = F(x, \hat{\theta}_n),$$

where $\hat{\theta}_n \to \theta$ is a root-n consistent estimator of θ (obtained e.g. by maximum likelihood estimation). This leads to the semi-parametric estimator,

$$\hat{f}_{\theta_n}(y|x) = \hat{g}(y)\hat{c}_n(\tilde{F}_n(x), \hat{G}(y)) \tag{4.7}$$

where $\hat{c}_n(.,.)$ is the kernel copula density estimator based on the approximate data $(\tilde{F}_n(X_1), \hat{G}(Y_1)), \ldots, (\tilde{F}_n(X_n), \hat{G}(Y_n))$. This setup can be useful in particular to incorporate supplementary information or constraints on the model, when the Statistician has some insight on the kind of distribution followed by the marginals. This can occur when one deals e.g. with failure-time distributions or more generally in survival analysis, where

exponential families of distributions are prominent.

Uniform consistency of the plugged distribution function estimator yields without further work the consistency of this semi-parametric estimator of the conditional density, as exemplified in the proposition below:

Proposition 4.5 *Assume $F(.,.)$ is Lipschitz in its second argument in the sense that there exists a non-negative function $C(x,\theta)$, such that*

$$|F(x,\theta) - F(x,\theta_0)| \leq C(x,\theta_0)|\theta - \theta_0|, \quad x \in \mathbb{R}, \theta, \theta_0 \in \Theta$$

with $\sup_{x\in\mathbb{R}} C(x,\theta_0) < \infty$, and that $\hat{\theta}_n$ is a $\sqrt{n}$ consistent estimator of θ, i.e.

$$\sqrt{n}(\hat{\theta}_n - \theta) = O_P(1)$$

Then, with hypothesis similar to that of theorem 3.5, its conclusion applies for the semi parametric quantile-copula conditional density estimator 4.7.

Proof *The Lipschitz hypothesis and convergence of the estimator yields uniform consistency of the distribution function, i.e. $\sup_{x\in\mathbb{R}} |\tilde{F}_n(x) - F(x,\theta_0)| = O_P(1/\sqrt{n})$. The conclusion follows by using this result instead of the Chung-Smirnov property in the proof of 3.5.*

4.1.4 Further remarks

On the asymptotic efficiency of the estimation of the copula density

Similarly to the discussion on the estimation of the margins, a similar one on the efficiency of the estimator of the copula density could be based on the concepts of sufficiency and of invariance of a statistical model w.r.t. to a group of transformations, as in [92]. Indeed, copula functions are known to be invariant with respect to monotone increasing continuous transformations (see e.g. [101]). Therefore, the search of a sufficient statistic for the estimation of a copula is reduced to the search of a maximal invariant of the group of

monotone increasing continuous transformation, which are the ranks. By transforming the data by the empirical distribution functions, the copula density part of the proposed estimator is partially based on the normalised ranks $F_n(X_i) = \text{rank}(X_i)/n$, and thus appears to be based on the minimal sufficient statistic. We leave open for further research such questions on efficiency and just mention some of the related work on semi-parametric efficiency, see [70, 85, 135, 36, 61].

On discrete marginal distributions

For non continuous data, a transformation of the data does not really change its distribution but simply the sample space. As an example, take X binary-valued with $P(X = 0) = P(X = 1) = 0.5$, then $X_i^* := F_n(X_i)$ is approximately equal to 1 or $1/2$ for $X = 1$ or 0 respectively, i.e. with probability one-half each and the distribution of the transformed data is not reduced to a prescribed uniform one. As a consequence, our approach to conditional density does not extend to the case of discrete data or more generally to the case of discontinuous marginals.

4.2 Practical implementation of the estimator

In this section we address the issues related to the practical implementation of the proposed estimator to achieve a good finite sample performance.

4.2.1 On the infinities at the corners and the approximate observations

Let's recall the following elementary remark that whenever $(U_1, \ldots, U_n)$ is a sequence of i.i.d uniform $[0, 1]$ random variables, then they are in the open interval $(0, 1)$ with probability 1. In particular, since F and G are continuous, the pseudo-variables $F(X_1), \ldots, F(X_n)$

and $G(Y_1), \ldots, G(Y_n)$ are in $(0,1)$ with probability 1.

To the contrary, the approximated sample $\{(F_n(X_1), \ldots, F_n(X_n)\}$ and $\{(G_n(X_1), \ldots, G_n(X_n)\}$, obtained by estimating the unknown c.d.f. F and G by the empirical counterparts, does not lie in this set. Indeed, since $\{(F_n(X_1), \ldots, F_n(X_n)\}$ and $\{(G_n(X_1), \ldots, G_n(X_n)\}$ are equal to the normalised ranks $\{\frac{rank(X_i)}{n}\}$ and $\{\frac{rank(Y_i)}{n}\}$, the transformations F_n and G_n establish a one-to-one correspondence from the original samples $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ to the set $\{1/n, 2/n, \ldots, 1\}^2$. In other words, one of the approximate observation (the one corresponding to the largest order statistic) is pushed to the extreme right end value 1, and this, with probability 1. Therefore, when it comes to approximate these pseudo-variables by approximate ones, this feature is not in favour of using the empirical cumulative distribution functions as estimates.

In addition, it is well known that many copula densities are unbounded at the corners $(0,0)$ and $(1,1)$, see e.g. [101] or [83]. Coupled with the fact the estimation of a density on its boundary by the kernel method is somehow harder than in the interior of its support (see section 4.2.2 below), this also motivates the substitution, at least in the practical implementations, of the empirical c.d.f. by another estimate. An usual and simple modification encountered in the literature is to use a rescaled version such as

$$F_n^R(x) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x} = \frac{n}{n+1} F_n(x)$$

and similarly for G_n. Note that this slight modification still retains the computationally important feature of being a rank-based estimator. We point the fact that this motivation for using substitutes of the empirical c.d.f. comes also in addition to the arguments developed in subsection 4.1.1 in favour of Aggarwal's version [5] of the empirical c.d.f., or smoothed kernel estimates $\hat{F}$ and $\hat{G}$.

4.2.2 Boundary bias correction

As noted in, e.g. [93], chapter 1, it is a well-known fact that whenever the support of a density has a finite boundary, the kernel smoothing method becomes inconsistent on that boundary. This is precisely the case for the estimation of the copula density which is supported on the compact square $[0,1]^2$. Although the previous subsection 4.2.1 shows that, provided a suitable modification of the empirical c.d.f. is implemented, one is never on the boundary of the quantile density, in practice, the biases near the boundaries can no longer be neglected in a finite sample setting. To that end, we discuss below some possible modifications of the estimator to take into account this issue.

Bias reduction by transformed kernels

Building on the idea of transforming the data, one can use it in a reversed way as advocated by Charpentier [33] by using the following elegant trick. The idea is to use transformed kernels to estimate the copula density: Let Φ a *known* c.d.f. of a continuous differentiable on $\mathbb{R}$ with ϕ its p.d.f. assumed to be strictly positive. Φ^{-1} is therefore a one-to-one mapping from $[0,1]$ to $\mathbb{R}$. Let $\tilde{X} = \Phi^{-1}(U)$, $\tilde{Y} = \Phi^{-1}(V)$, the density of $\tilde{X}, \tilde{Y}$ is

$$f_{\tilde{X},\tilde{Y}}(\tilde{x},\tilde{y}) = \phi(\tilde{x})\phi(\tilde{y})c(\Phi(\tilde{x}),\Phi(\tilde{y}))$$

and can be easily estimated by the standard kernel estimator $\hat{f}_{\tilde{X},\tilde{Y}}(\tilde{x},\tilde{y})$ (or any other technique), which is now free of boundary bias. The conditional density estimator is then obtained by back-transformation as

$$\hat{f}(y|x) = \hat{g}(y)\frac{1}{\phi(\Phi^{-1}(F_n(x)))\phi(\Phi^{-1}(G_n(y)))}\hat{f}_{\tilde{X},\tilde{Y}}(\Phi^{-1}(F_n(x)),\Phi^{-1}(G_n(y))).$$

This algorithm can be seen by the chain of transformations,

$$X_1,\ldots,X_n,x \xrightarrow{F} U_1^*,\ldots,U_n^*,u^* \xrightarrow{\Phi^{-1}} \tilde{X}_1,\ldots,\tilde{X}_n,\tilde{x}$$

$$Y_1,\ldots,Y_n,y \xrightarrow{G} V_1^*,\ldots,V_n^*,v^* \xrightarrow{\Phi^{-1}} \tilde{Y}_1,\ldots,\tilde{Y}_n,\tilde{y}$$

Bias reduction by boundary kernels

A common solution in the literature is to use higher order or boundary-corrected kernels to improve the results on the copula density estimator. As an example, one can correct a kernel K *à la Gasser-Muller* [58] by K^c as follows:

$$K^c(x,y) = \begin{cases} K(\frac{x-y}{a}) / \int_{-x/a}^{\infty} K(v)dv & \text{if } x \in [0,a] \\ K(\frac{x-y}{a}) & \text{if } x \in [a, 1-a] \\ K(\frac{x-y}{a}) / \int_{-\infty}^{(1-x)/a} K(v)dv & \text{if } x \in [1-a, 1] \end{cases}$$

Bias reduction by Beta kernels

Chen [35] advocates to use Beta kernels to remove boundary bias. Recall that the density of the Beta(p,q) distribution is

$$\beta_{p,q}(t) = \frac{(1-t)^{q-1}t^{p-1}}{B(p,q)} \mathbb{1}_{0<t<1}$$

for $t \in [0,1]$, $p, q > 0$, where $B(p,q)$ denotes the Beta Euler function. The Beta kernel is defined as

$$K_{x,b}(t) = \beta_{x/b+1,(1-x)/b+1}(t) = \frac{t^{x/b}(1-t)^{(1-x)/b}}{B(x/b+1,(1-x)/b+1)} \mathbb{1}_{(0\leq t\leq 1)}$$

To improve the bias of the previous estimator, Chen also considered a modified Beta kernel defined as follows by

$$K^*_{x,b}(t) = \begin{cases} \beta_{x/b,(1-x)/b}(t) & \text{if } x \in (2b, 1-2b) \\ \beta_{\rho(x,b),(1-x)/b}(t) & \text{if } x \in [0, 2b] \\ \beta_{x/b,\rho(1-x,b)}(t) & \text{if } x \in [1-2b, 1] \end{cases}$$

where $\rho(x,b) = 2b^2 + 2.5 - \sqrt{4b^4 + 6b^2 + 2.25 - x^2 - x/b}$.

With either modification of the kernel, the copula density pseudo estimator is now defined as $c_n(u,v) = n^{-1}\sum_{i=1}^n K_{u,b}(U_i)K_{v,b}(V_i)$ and the conditional density estimate becomes

$$\hat{f}(y|x) = \left(\frac{1}{nh_n}\sum_{i=1}^n K_0\left(\frac{y-Y_i}{h_n}\right)\right)\left(\frac{1}{n}\sum_{i=1}^n K_{F_n(x),a_n}(F_n(X_i))K_{G_n(y),a_n}(G_n(Y_i))\right)$$

This is the solution we implemented in the simulations. See e.g. Gustafsonn, Hagmann,

Nielsen, Scaillet [64] for an application of Beta kernels to loss distributions.

4.2.3 On bandwidth selection

As mentioned earlier in chapter 3, the performance of non-parametric estimation techniques depends crucially on the bandwidths parameters, whose optimal choice depends itself on some underlying unknown features of the density, see e.g. [143] for a discussion. In this section, we discuss the specific features of the bandwidth selection problem in the conditional density context and sketch some possible strategies based on the relevant literature in the domain.

Bandwidth selection strategies for ratio-shaped conditional density estimators

For ratio-shaped estimators as those mentionned in sections 3.1 and 4.3 below, we denote by h_X and h_Y their smoothing parameters in the respective x and y direction. Bandwidth selection for ratio-shaped conditional density estimators is more complicated than for density estimation. Indeed, conditional density involves multivariate densities, and the complexity of the bandwidth selection procedure is increased with the dimensionality of the data. Moreover, apart from the higher dimensional nature of the problem, conditioning requires to localise the X data near x before smoothing in the y direction. Therefore, bandwidth selection in the X and Y part of the vector (X, Y) are interrelated and one has to perform a simultaneous selection for h_X and h_Y.
We review below some strategies proposed in the literature for ratio-shaped estimators. Bashtannyk and Hyndman [13] propose a practical method for conditional density estimation which is an iterative combination of these methods.

- Rule of thumb methods

  A simple ad-hoc method is to make assumptions on the model underpinning the densities. Bashtannyk and Hyndman [13] propose the following rule-of-thumb method

based on the assumption that the conditional density is Normal with a linear regression, $f(y|x) \approx \mathcal{N}(d_0 + d_1 x, \sigma^2)$ and $f(x) \approx \mathcal{N}(\mu, \nu^2)$.

- Two step regression methods

  Fan, Yao and Tong [52] propose a mixed-method in a two step procedure : first, select h_y by the Normal reference rule, and second, with this h_y selected, choose a data-dependent method based from nonparametric regression such as Plug-in or Cross-validation.

- Least-squares cross validation

  As advocated by Fan and Yim [53] and Hall, Racine and Li [68], it consists in considering the following integrated square error,

$$I = \iint [\hat{f}(y|x) - f(y|x)]^2 f(x) dy dx$$

  which is expanded as

$$\begin{aligned} I &= \iint \hat{f}^2(y|x) f(x) dy dx - 2 \iint \hat{f}(y|x) f(y|x) f(x) dy dx \\ &+ \iint f^2(y|x) f(x) dy dx \\ &:= I_1 - 2I_2 + I_3 \end{aligned}$$

  where the last term I_3 is independent of the bandwidth. Then I_1 and I_2 are estimated by

$$\begin{aligned} \hat{I}_1 &= \frac{1}{n} \sum_{i=1}^{n} \int \hat{f}_{-i}(y|X_i) dy \\ \hat{I}_2 &= \frac{1}{n} \sum_{i=1}^{n} \hat{f}_{-i}(Y_i|X_i) \end{aligned}$$

  respectively, where $\hat{f}_{-i}(y|x)$ is the leave-one out estimator based on the sample $(X_j, Y_j), j = 1, \ldots, n, j \neq i$. For example, for the double kernel estimator of section

3.1, the leave-one out estimator writes,

$$\hat{f}_{-i}(y|x) = \sum_{j=1, j\neq i}^{n} \frac{K\left(\frac{x-X_j}{h_X}\right) K\left(\frac{y-Y_j}{h_Y}\right)}{K\left(\frac{x-X_j}{h_X}\right)}$$

By setting

$$CV(h_X, h_Y) = \hat{I}_1 - 2\hat{I}_2$$

the cross-validation method is to select simultaneously h_X^{CV},h_Y^{CV} such that

$$(h_X^{CV}, h_Y^{CV}) = \arg\inf CV(h_X, h_Y)$$

Youndjé Sarda and Vieu [147] and [148] give optimality results for the cross-validation of the double kernel conditional density estimator with the same bandwidth, i.e. for $h_X = h_Y$.

- Bootstrap Method

  Hall, Wolff and Yao [69], apply a bootstrap-local modelling approach, which can be decomposed in the following steps:

  1. First, a polynomial regression model $Y_i = a_0 + a_1 X_i + \ldots + a_k X_i^k + \sigma\epsilon_i$ is fitted to the data, where ϵ_i is standard normal and $a_0, \ldots, a_k, \sigma$ are estimated from the data and k is selected by the Akaike Information Criterion.

  2. Second, a parametric estimator $\check{f}(y|x)$, based on the model, is computed. A bootstrap sample $Y_1^*, \ldots, Y_n^*$ based on the given observations $X_1, \ldots, X_n$ is generated by Monte-Carlo simulation from the fitted parametric estimate $\check{f}(y|x)$.

  3. From the simulated observations, a bootstrap estimate $\hat{f}^*(y|x)$ is computed from $\hat{f}(y|x)$, where (X_i, Y_i) is replaced by X_i, Y_i^*.

  4. The bandwidths are selected in order to minimise the bootstrap estimate of the absolute deviation error of $\hat{f}(y|x)$, i.e.

  $$(h_X^*, h_Y^*) = \arg\inf E[(|\hat{f}^*(y|x) - \check{f}(y|x)|)|\{X_i, Y_i\}]$$

Bandwidth selection for the quantile-copula conditional density estimator

- Theoretical optimal local bandwidths:

  In the pointwise case, we recall that the calculation of the Asymptotic Mean Square Error (AMSE) of the conditional density estimator in claim 3.12 shows that the main term is

$$E_0 = \frac{a_n^4 g^2(y) \left(B_k(c,x,y)\right)^2}{4} + \frac{g(y) f(y|x) ||K||_2^2}{n a_n^2} + o\left(a_n^4 + \frac{1}{n a_n^2}\right)$$

  with $B_K(c,x,y) := m_2(K_1)\frac{\partial^2 c(F(x),G(y))}{\partial u^2} + m_2(K_2)\frac{\partial^2 c(F(x),G(y))}{\partial v^2}$. Therefore, by differentiating with respect to a_n and solving the equation $\partial E_0/\partial a_n = 0$, one gets the optimal bandwidth balancing the bias and variance,

$$a_{op}(x,y) = n^{-1/6}\left(\frac{2 f(y|x)||K||_2^2}{g(y)\left(B_k(c,x,y)\right)^2}\right)^{1/6}$$

  One note that the terms involving the h_n bandwidth do not appear in the AMSE, since they are negligible to first order compared to the above term. However, inspection of the decomposition of $\hat{f}_n(y|x) - f(y|x)$ shows that the next term to $g(y)[c_n(F(x),G(y)) - c(F(x),G(y))]$ is $c(F(x),G(y))[\hat{g}_n(y) - g(y)]$, which is asymptotically negligible compared to the former term. Squaring and taking expectation, the AMSE of the estimator is thus a linear combination of the AMSE of $c_n(F(x),G(y))$ and of the AMSE of $\hat{g}_n(y)$

$$MSE[\hat{f}_n(y|x)] = g^2(y) MSE[c_n(F(x),G(y))] + c^2(F(x),G(y)) MSE[\hat{g}_n(y)]$$

  with

$$MSE[\hat{g}_n(y)] = o(MSE[c_n(F(x),G(y))])$$

  plus remaining terms which are negligible. In a small sample setting, choosing the bandwidth h_n which minimizes the AMSE of $g_n(y)$ will thus lead to a better overall minimising of the AMSE of $f(y|x)$. Therefore, from a theoretical point of view, the bandwidths choices for a_n and h_n can be done independently, in a first

order approximation. For the Y density part, the optimal bandwidth is the optimal bandwidth in univariate density estimation and is given by

$$h_{op}(y) = n^{-1/5} \left(\frac{g(y)||K_0^2||^2}{[g''(y)m_2(K_0)]^2} \right)^{1/5}$$

- A separate bandwidth selection strategy:

  We can therefore propose the following bandwidth selection procedure:

  1. choose h_n which minimises the AMSE of the estimator $\hat{g}_n(y)$ of the density g of Y,
  2. choose a_n which minimises the AMSE of the pseudo-estimator $c_n(F(x), G(y))$ of the copula density $c(F(x), G(y))$.

  A practical data-dependent method is therefore to select h by cross-validation on the X data alone, then, for a_n, use bivariate cross-validation on the pseudo data $F_n(X_1), \ldots, F_n(X_n)$ and $G_n(Y_1), \ldots, G_n(Y_n)$.

  We note this approach is somehow different from competitors that have to choose the optimal bandwidths simultaneously, and propose sequential bandwidth selection procedures, by first picking a choice for the first bandwidth, and then choosing the optimal second bandwidth given this choice.

4.3 Simulations and comparison with other estimators

4.3.1 Presentation of alternative estimators

For convenience, we recall below the definition of other estimators of the conditional density encountered in the literature and summarize their bias and variance properties. We will note the bias of the ith estimator $\hat{f}_n^i(y|x)$ by E_i and its variance by V_i.

1. **Double kernel estimator**: as defined in the introduction section of our paper by the following ratio,

$$\hat{f}_n^{(1)}(y|x) := \frac{\frac{1}{n}\sum_{i=1}^{n} K'_{h_1}(X_i - x)K_{h_2}(Y_i - y)}{\frac{1}{n}\sum_{i=1}^{n} K'_{h_1}(X_i - x)}.$$

where h_1 and h_2 are the bandwidths. One then has, see e.g. [81],

- Bias:

$$B_1 = \frac{h_1^2 m_2(K)}{2}\left(2\frac{f'(x)}{f(x)}\frac{\partial f(y|x)}{\partial x} + \frac{\partial^2 f(y|x)}{\partial x^2} + \left(\frac{h_2}{h_1}\right)^2 \frac{\partial^2 f(y|x)}{\partial y^2}\right) + o\left(h_1^2 + h_2^2\right)$$

- Variance:

$$V_1 = \frac{\|K\|_2^2 f(y|x)}{nh_1h_2 f(x)}\left(\|K\|_2^2 - h_2 f(y|x)\right) + o\left(\frac{1}{nh_1h_2}\right)$$

2. **Local polynomial estimator**: Set

$$R(\theta, x, y) := \sum_{i=1}^{n}\left(K_{h_2}(Y_i - y) - \sum\nolimits_{j=0}^{r}\theta_j(X_i - x)^j\right)^2 K'_{h_1}(X_i - x),$$

then the local polynomial estimator is defined as

$$\hat{f}_n^{(2)}(y|x) := \hat{\theta}_0,$$

where $\hat{\theta}_{xy} := (\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_r)$ is the value of θ which minimizes $R(\theta, x, y)$. This local polynomial estimator, although it has a superior bias than the kernel one, is no longer restricted to be non-negative and does not integrate to 1, except in the special case $r = 0$. From results of [52], we get for the local linear estimator (see also [51] p. 256),

- Bias:

$$B_2 = \frac{h_1^2 m_2(K')}{2}\frac{\partial^2 f(y|x)}{\partial x^2} + \frac{h_2^2 m_2(K)}{2}\frac{\partial^2 f(y|x)}{\partial y^2} + o(h_1^2 + h_2^2)$$

- Variance:

$$V_2 = \frac{||K||_2^2||K'||_2^2 f(y|x)}{nh_1h_2f(x)} + o\left(\frac{1}{nh_1h_2}\right)$$

3. **Local parametric estimator**: As in [82] and [51], set

$$R_1(\theta, x, y) := \sum_{i=1}^{n} \left(K_{h_2}(Y_i - y) - A(X_i - x, \theta)\right)^2 K'_{h_1}(X_i - x)$$

where $A(x, \theta) = l\left(\sum_{j=0}^{r} \theta_j (X_i - x)^j\right)$ and $l(.)$ is a monotonic function mapping $\mathbb{R} \mapsto \mathbb{R}^+$, e.g. $l(u) = \exp(u)$. Then,

$$\hat{f}_n^{(3)}(y|x) := A(0, \hat{\theta}) = l(\hat{\theta}_0).$$

- Bias:

$$B_3 = h_1^2 \eta(K') \left(\frac{\partial^2 f(y|x)}{\partial x^2} - \frac{\partial^2 A(0, \theta_{xy})}{\partial x^2}\right) + \frac{h_2^2 m_2(K)}{2} \frac{\partial^2 f(y|x)}{\partial y^2} + o(h_1^2 + h_2^2)$$

- Variance:

$$V_3 = \frac{\tau(K, K')^2 f(y|x)}{nh_1h_2f(x)} + o\left(\frac{1}{nh_1h_2}\right)$$

where η and τ are kernel dependent constants.

4. **Constrained local polynomial estimator**: A simple device to force the local polynomial estimator to be positive is to set $\theta_0 = \exp(\alpha)$ in the definition of R_0 to be minimized. The constrained local polynomial estimator $\hat{f}_n^4(y|x)$ is then defined analogously as the local polynomial estimator $\hat{f}_n^2(y|x)$. We have, as in [82] and [51]:

- Bias:

$$B_4 := h_1^2 \frac{m_2(K')}{2} \frac{\partial^2 f(y|x)}{\partial x^2} + h_2^2 \frac{m_2(K)}{2} \frac{\partial^2 f(y|x)}{\partial y^2} + o(h_1^2 + h_2^2)$$

- Variance:

$$V_4 = \frac{\|K\|_2^2 f(y|x)}{nh_1h_2f(x)} + o\left(\frac{1}{nh_1h_2}\right)$$

4.3.2 Asymptotic Bias and Variance comparison

All estimators have (hopefully) the same order $n^{-1/3}$ and $n^{-2/3}$ in their asymptotic bias and variance terms, for the usual bandwidths choice. The main difference lies in the constant terms which depend on unknown densities.

Bias: Contrary to all the alternative estimators whose bias involves derivatives of the full conditional density, one can note that our estimator's bias only involves the density of Y and the derivatives of the copula density. To make things more explicit, the terms involved, e.g. in the local polynomial estimator, write themselves as the sum of the derivatives of the conditional density,

$$h_n^{-2} B_2 \approx \frac{\partial^2 f(y|x)}{\partial x^2} + \frac{\partial^2 f(y|x)}{\partial y^2}$$

that is to say,

$$\begin{aligned} h_n^{-2} B_2 \approx\ & f'(x)g(y)\frac{\partial c(F(x),G(y))}{\partial u} + f^2(x)g(y)\frac{\partial^2 c(F(x),G(y))}{\partial u^2} \\ & + 2g'(y)g(y)\frac{\partial c(F(x),G(y))}{\partial v} + g^3(y)\frac{\partial^2 c(F(x),G(y))}{\partial v^2} \end{aligned}$$

whereas our $(g(y)/2)B_K(c,x,y)$ term, modulo the constants involved by the kernel, is written as

$$a_n^{-2} B_0 \approx g(y)\left(\frac{\partial^2 c(F(x),G(y))}{\partial u^2} + \frac{\partial^2 c(F(x),G(y))}{\partial v^2}\right).$$

It then becomes clear that we have a simpler expression, with less unknown terms, as is the case for competitors which do involve the density f and its derivative f' of X and the derivative g' of the Y density.

In a fixed bandwidth and asymptotic context, it seems difficult to compare further. Nonetheless, we believe this feature of our estimator would be practically relevant when it comes to choosing the bandwidths. Indeed, bandwidth selection is usually performed by minimizing local or global asymptotic error criteria such as Asymptotic Mean Square Error (AMSE) or Asymptotic Mean Integrated Square Error (AMISE), in which unknown

terms have to be estimated. Since in our approach, the asymptotic bias and variance involve less unknown terms, we expect that a higher accuracy could be obtained in this pre-estimation stage. Moreover, by having managed to separate the estimation problem of the marginal from the copula density, we could use known optimal data-dependent bandwidths selection procedures for density estimation such as cross validation, separately for the density of Y and for the copula density.

Remark 4.6 *As mentioned earlier in subsection 4.2.2, since the copula density c has a compact support* $[0,1]^2$*, our estimator may suffer from bias issues on the boundaries, i.e. in the tails of* X *and* Y*. In the tail of the distribution of* X*, this bias issue in the copula density estimator is balanced by the improved variance, as shown below.*

Variance: The variance of our estimator involves a product of the density $g(y)$ of Y by the conditional density $f(y|x)$,

$$na_n^2 V_0 \approx g(y)f(y|x) = g^2(y)c(F(x), G(y)), \tag{4.8}$$

whereas competitors involve the ratio of $f(y|x)$ by the density $f(x)$ of X

$$\frac{f(y|x)}{f(x)} = \frac{g(y)}{f(x)}c(F(x), G(y)). \tag{4.9}$$

As a consequence, if we compute the ratio of the asymptotic variances between our estimator (4.8) and a competitor (4.9), we get, neglecting the constants in the kernels,

$$\frac{V_0}{V_1} = f(x)g(y)$$

Since marginal densities are positive and integrate to unity, they are usually smaller than one, and so is the latter ratio for a wide range of x and y values. That is to say, the proposed estimator is, bias comparisons left aside, asymptotically more efficient than its competitors.

Moreover, it is a remarkable feature of the estimator we propose, that its variance does not involve directly $f(x)$, as is the case for the competitors, but only its contribution to Y,

through the copula density. This reflects the ability announced in the introduction of the copula representation to have effectively separated the randomness pertaining to Y alone, from the dependence structure of (X, Y). Moreover, our estimator also does not suffer from the unstable nature of competitors who, due to their intrinsic ratio structure, get an explosive variance for small value of the density $f(x)$, making conditional estimation difficult, e.g. in the tail of the distribution of X.

Remark 4.7 *To make estimators comparable, we have restricted ourselves to so-called fixed bandwidths estimators, i.e. nonparametric estimators where the bandwidths are of the generic form $h_n = bn^\alpha$ or $h_n = b(\ln n/n)^\alpha$ with α and b real numbers. Improved behavior for all the preceding estimators can be obtained with data-dependent bandwidths where $h_n = H_n(X_1, \ldots, X_n, x)$ can be functions of the location and of the data.*

4.3.3 Finite sample numerical simulation

In this subsection, we aim at validating the potential interest of the proposed procedure and complementing the asymptotic comparison of the previous subsection with a limited numerical illustration where, at least on one model, the proposed estimator leads to a better estimation procedure.

Model and implementation

We simulated a sample of $n = 100$ variables (X_i, Y_i), from the following model: X, Y is marginally distributed as $\mathcal{N}(0, 1)$ and linked via Frank Copula .

$$C(u, v, \theta) = \frac{\ln[(\theta + \theta^{u+v} - \theta^u - \theta^v)/(\theta - 1)]}{\ln \theta}$$

with parameter $\theta = 100$.

We implemented below the proposed estimator with some of its competitors and conducted a comparative analysis. To deal with the possible issues raised in section 4.2, we used the following modifications of the estimator:

- Infinities at the corners : we changed the empirical distribution functions F_n and G_n to $n/(n+1)F_n$ and $n/(n+1)G_n$ respectively.

- Boundary bias : we used the Beta kernels of Chen [35], as advocated in subsection 4.2.2.

- Bandwidths : for the selection of the a_n bandwidth of the copula density estimator, we applied Scott's Rule on the data $F_n(X_i)$. We used the Normal reference rule for the bandwidth h_n of the estimator $\hat{g}(y)$ of the density g of Y.

For all the other estimators, we used Epanechnikov kernels. For the matter of bandwidth selection, we decided, in order to get a first picture, to restrict ourselves to simple, fixed for all x, y, rule-of-thumb methods based on Normal reference rule .

Comparative results

We plotted the conditional density along with its estimations on the domain $x \in [-5, 5]$ and $y \in [-3, 3]$ on figure 4.1. A comparison plot at $x = 2$ is shown on figure 4.2.

Clipping and Estimation in the tails

As mentioned earlier, as the performance of the estimators depends on the performance of the bandwidths selection method, it is delicate to give a conclusive answer. However, we would like to illustrate at least one case where the proposed estimator clearly outperforms its competitors. Indeed, one major issue of alternative estimators already mentioned is their numerical explosion when the estimated density $\hat{f}(x)$ is close to zero. In particular, if the kernel is of compact support, the denominator is zero for the x whose distance from the closest X_i exceeds half the bandwidth times the length of the support, thereby allowing estimation only on a closed subset of X included in $[\min X_i, \max X_i]$. This is one of the reasons why simulation studies are often performed either with a marginal X density of bounded support and/or with a Gaussian kernel. Note that the problem

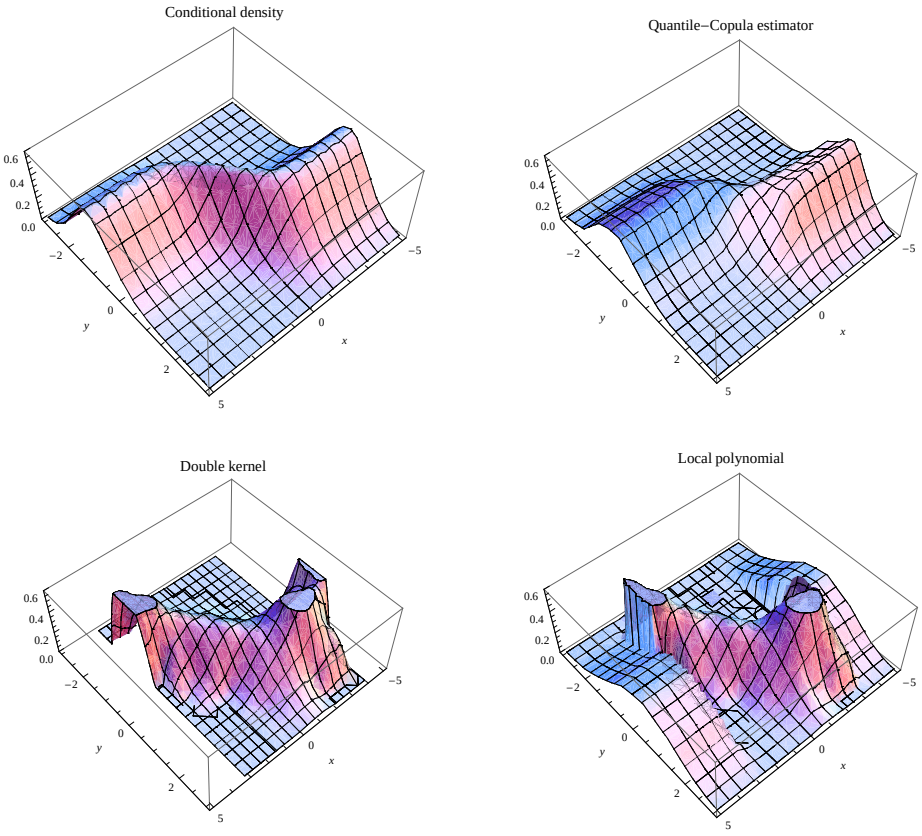

Figure 4.1: 3D Plots. From left to right, top to bottom: true density, quantile-copula estimator, double kernel, local polynomial (clipped).

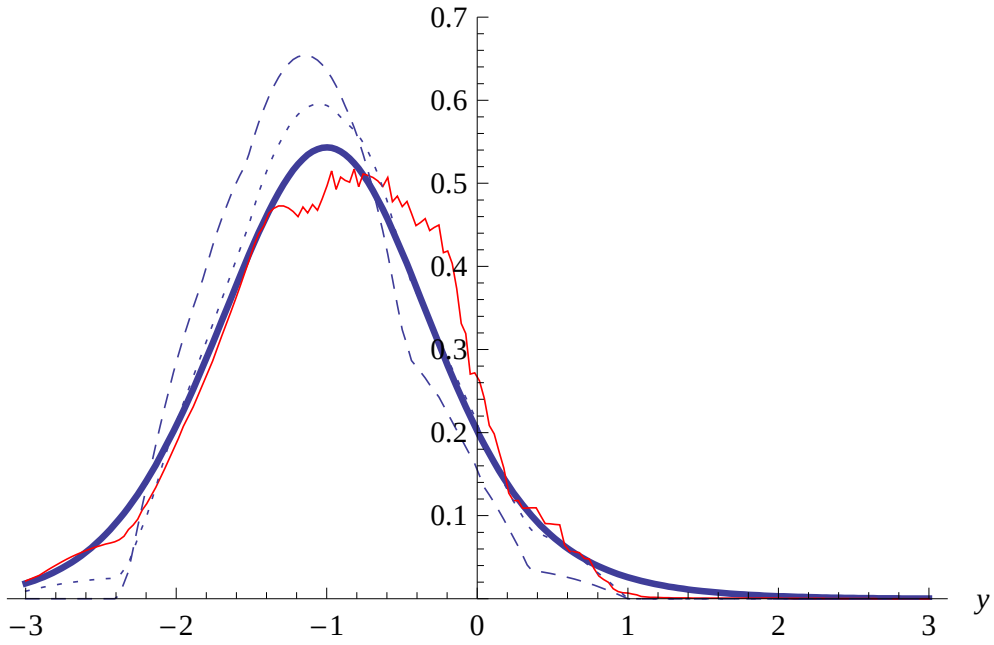

Figure 4.2: Comparison at x=2: conditional density=thick curve, quantile-copula=continuous line, double kernel=dotted curve, local polynomial=dashed curve.

remains with a Gaussian kernel since the estimated density can become quickly lower than the machine precision. To avoid this numerical explosion, the definition of the conditional density estimators have to be modified either by

$$\hat{f}(y|x) = \begin{cases} \frac{\hat{f}_{XY}(x,y)}{\hat{f}_X(x)} & \text{if } \hat{f}_X(x) > c \\ \hat{a}(y) & \text{if } \hat{f}_X(x) = 0 \end{cases} \quad \text{or by, } \hat{f}(y|x) = \frac{\hat{f}_{XY}(x,y)}{\max\{\hat{f}(x), c\}}$$

where $c > 0$ is an arbitrary amount of clipping, and $\hat{a}(.)$ is an arbitrary density estimator (usually chosen to be zero or $\hat{g}(y)$).

An illustration of these issues clearly appears in figure 4.1. The unclipped version of the double kernel estimator is unable to estimate the conditional density for $|x|$ roughly > 3, and the clipped version of the local polynomial estimator with $c = 0.00001$ and $\hat{a}(y) = \hat{g}(y)$ gives a wrong estimation in the tails, reflecting the arbitrary choices in the clipping decision. To the contrary, the quantile-copula estimator is surprisingly able to estimate the conditional density $f(y|x)$ at locations x where there is "no data", i.e. in

the tails of the distribution of X. A possible heuristic explanation of this apparently paradoxal phenomenon comes from the fact that the estimator is partially based on the ranks of X_i and Y_i. Therefore, it can recover "hidden" information on the density of X from the ordering of the pairs (X_i, Y_i). See Hoff [76] for a detailed explanation. We believe that this feature might be of potential interest for applications, e.g. in statistical inference of extreme values and rare events, see section 6.4.2 for a more detailed account on that perspective.

Obviously, we performed our numerical comparison on one specific model and a more detailed simulation is required to make more general statements. However, such a study, together with applications to specific practical fields is beyond the scope of this thesis and is left for future research. See chapter 6 for some proposed sketches in that direction.

Chapter 5

Application of conditional density estimation to prediction

Abstract : In this chapter, we are interested in the prediction of a response variable Y given an observed value of an explanatory variable X and the knowledge of an i.i.d. training sample (X_i, Y_i), $i = 1, \ldots, n$. In this nonparametric setup, it is natural to build upon the estimator of the conditional density studied in Chapter 3 to derive point and set predictors, which are presented in section 5.1. Asymptotic consistency of the conditional mode and of the predictive sets are studied in section 5.2 and 5.3 respectively, together with a discussion regarding their implementation. For the more delicate case of the conditional mean of section 5.4, we prove its convergence for a limited case, present some alternative implementations and establish a connection with Yang [145] and Stute [128] estimators of the regression.

5.1 Nonparametric statistical approaches to prediction

In this section, we discuss the issue of predicting a response variable Y given an observed value of an explanatory variable X. Although, estimating the conditional density of Y allows to fully quantify the input of X on Y, it is sometimes desirable in practical applications to summarise it through a univariate characteristic of this conditional distribution, such as the "most likeky" value to appear or the "average". Moreover, one can also be interested to give predictive sets similar to confidence interval in estimation. This is the approach pursued below, which shows how to define point and interval predictors from the conditional density estimator.

5.1.1 Construction of Point predictors

To prevent repetitions with the discussions of chapter 1, we just briefly recall some possible Bayes (i.e. probabilistic) point predictors, among others, according to the distance or loss $L : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}^+$ considered to measure the performance of the prediction:

- for the squared loss, $L(x, y) = (x - y)^2$, the "average" is given by the regression function $m(x) = E[Y|X = x]$,
- for the $0 - 1$ loss, $L(x, y) = \mathbb{1}_{x \neq y}$, the "most likely value" is the conditional mode $\theta(x) := \arg\sup_y f(y|x)$,

Since we have an estimator of the conditional density, we can define the corresponding statistical point predictors by their empirical counterparts as follows:

- for the regression, one gets an estimator by $\hat{m}(x) := \int y \hat{f}_n(y|x)dy$
- for the conditional mode, one gets an estimator by $\hat{\theta}(x) := \arg\sup_y \hat{f}_n(y|x)$

The two aforementioned statistical point predictors are studied below in subsections 5.2 and 5.4 respectively.

5.1.2 Predictive intervals and level sets

Regarding predictive sets, one is interested in defining a region of the sample space covering a specified probability, i.e. to define a set $\mathcal{C}_\alpha(x)$ such that

$$P(Y \in \mathcal{C}_\alpha(x)|X = x) = \alpha$$

In the present context, there are numerous way to construct such predictive intervals or sets covering a specified conditional probability, e.g., by

- the interval symmetric around the mean;
- the interval symmetric around the median;
- the interval between the $\frac{1-\alpha}{2}$ and $\frac{1+\alpha}{2}$ quantiles;
- the interval of shortest length;
- the interval that minimizes the probability of covering a given family of sets.
- et caetera...

Hyndman [80] provides a detailed discussion of the issues involved to define such a probability region in the unconditional case. In the conditional case we are interested, the approach is similar but the coverage region is dependent on a fixed, chosen x. In particular, to incorporate features such as multimodality, a level set approach called Highest Density Region (HDR) is advocated. See also the minimum length predictor by Polonik in [104] and Fan and Yao [51]. This highest density region approach also allows to give an informative and convenient graphical display of the predicted regions by drawing confidence bands with, e.g. 50 % and 99% coverage probability, as shown in [81].

In our present case, and following this HDR approach, we can use the estimator of the conditional density as a building block to define the set predictor as

Definition 5.1 *The level set (probabilistic) predictor is the set $\mathcal{C}_\alpha$ consisting of points y,*

$$\mathcal{C}_\alpha := \{y : f(y|x) \geq f_\alpha\} \tag{5.1}$$

where f_α is the largest constant such as the prediction set has coverage probability α,

$$P(Y \in C_\alpha(x)|X = x) \geq \alpha \tag{5.2}$$

$\mathcal{C}_\alpha$ is constructed in such a way that, (see [80]),

1. it minimises the volume (understood as the Lebesgue measure) among all sets with a given coverage probability;

2. every point y inside the set has a conditional probability density at least as large as every point outside the region.

In case of multimodality, $\mathcal{C}_\alpha(x)$ takes the form of an union of possibly disjoint intervals, say $\mathcal{C}_\alpha(x) = \bigcup I_\alpha(x)$, where each $I_\alpha(x) := [y_\alpha, y^\alpha]$, with $y_\alpha \leq y^\alpha$. Each extremity of these subintervals is such that $f(y_\alpha|x) = f(y^\alpha|x) = f_\alpha$. A plug-in strategy to define the corresponding statistical predictor is discussed in section 5.3.

5.2 Prediction by the conditional mode

5.2.1 Asymptotic properties of the conditional mode predictor

We follow the approaches of [86, 47, 48] and Parzen [102]. We fix S a compact subset of $\mathbb{R}$. In order to assure the existence of the desired object, we assume that $f(y|x)$ is such that:

(R) There exists an $\eta > 0$, an unique $y_0 \in S$ such that $f(.|x)$ is strictly increasing on $(y_0 - \eta, y_0)$, and strictly decreasing on $(y_0, y_0 + \eta)$.

Under this assumption, the problem of maximizing $f(y|x)$ on $S' = (y_0 - \eta, y_0 + \eta)$ has a unique solution, which is exactly y_0. Therefore, the conditional mode $\theta(x) = \arg\sup_{y \in S'} f(y|x)$ is defined and unique.

It is estimated by a plug-in estimate $\hat{\theta}(x) = \arg\sup_{y \in S'} \hat{f}(y|x)$. We have the following consistency result of the conditional mode predictor:

Proposition 5.2 *if $f(.|x)$ follows assumption (R), and the conditions for uniform consistency of the conditional density on a compact set of theorem 3.18, then,*

$$\hat{\theta}(x) \overset{a.s.}{\to} \theta(x).$$

Proof *Let k, n be integers and, set the index of the conditional mode estimator $\hat{\theta}(x)$ by $\hat{\theta}_n(x)$ to be more specific. By assumption (1), $f(.|x)$ is continuous and strictly increasing on $(\theta(x) - \eta, \theta(x))$. therefore, the inverse function $f^{-1}(.|x)$ exists and is continuous. Thus, by continuity of the latter at the point $f(\theta(x)|x)$, for any $\varepsilon > 0$,*

$$\exists \delta_1(\varepsilon) > 0, \forall y \in (\theta(x) - \eta, \theta(x)), |f(y|x) - f(\theta(x)|x)| \leq \delta_1(\varepsilon) \Rightarrow |y - \theta(x)| \leq \varepsilon$$

Similarly,

$$\exists \delta_2(\varepsilon) > 0, \forall y \in (\theta(x), \theta(x) + \eta), |f(y|x) - f(\theta(x)|x)| \leq \delta_2(\varepsilon) \Rightarrow |y - \theta(x)| \leq \varepsilon$$

So that,

$$\exists \delta(\varepsilon) > 0, \forall y \in (\theta(x) - \eta, \theta(x) + \eta), |f(y|x) - f(\theta(x)|x)| \leq \delta(\varepsilon) \Rightarrow |y - \theta(x)| \leq \varepsilon$$

By construction, $\hat{\theta}_k(x) \in (\theta(x) - \eta, \theta(x) + \eta)$, so that,

$$\exists \delta(\varepsilon) > 0, |f(\hat{\theta}_k(x)|x) - f(\theta(x)|x)| \leq \delta(\varepsilon) \Rightarrow |\hat{\theta}_k(x) - \theta(x)| \leq \varepsilon$$

and finally,

$$\exists \delta(\varepsilon) > 0, P(\sup_{k \geq n} |\hat{\theta}_k(x) - \theta(x)| > \varepsilon) \leq P(\sup_{k \geq n} |f(\hat{\theta}_k(x)|x) - f(\theta(x)|x)| > \delta(\varepsilon)) \quad (5.3)$$

On the other hand, it comes from the triangle inequality that

$$\begin{aligned}\left|f(\theta(x)|x) - f(\hat{\theta}_k(x)|x)\right| &\leq \left|\hat{f}_k(\theta(x)|x) - f(\theta(x)|x)\right| \\ &+ \left|\hat{f}_k(\hat{\theta}_k(x)|x) - f(\hat{\theta}_k(x)|x)\right| \\ &\leq 2 \sup_{y\in(\theta(x)-\eta,\theta(x)+\eta)} \left|\hat{f}_k(y|x) - f(y|x)\right|\end{aligned}$$

and uniform almost sure convergence of the conditional mode estimator on a compact set of theorem 3.18 entails that

$$\forall\delta > 0, \lim_{n\to\infty} P\left(\sup_{k\geq n} \sup_{y\in(\theta(x)-\eta,\theta(x)+\eta)} \left|\hat{f}_k(y|x) - f(y|x)\right| > \delta\right) = 0,$$

thus $\hat{\theta}(x) \overset{a.s.}{\to} \theta(x)$ *by equation (5.3).*

5.2.2 A remark on the practical implementation of the conditional mode predictor

Set $\mathcal{S}_{Y|X} = \{y : f(y|x) > 0\}$ the support of the conditional density. In practice, the search of the conditional mode can be difficult and time-consuming to implement. Indeed, as the conditional mode estimator is defined as the maximizer of $\hat{f}(y|x)$, i.e. $\hat{\theta}(x) = \arg\sup_{y\in\mathcal{S}_{Y|X}} \hat{f}(y|x)$, one has *a priori* to compute the estimator of the conditional density on a large number of y values in $\mathcal{S}_{Y|X}$ to find the largest value of the estimated conditional density.

Therefore, we would like to mention a method to ease the computation of the conditional mode predictor, proposed in the papers by Abraham, Biau, Cadre [3, 4]. An alternative is to maximize the estimator on the Y data $D_n := \{y_1, \ldots, y_n\}$, i.e. to set $\tilde{\theta}(x) = \arg\max_{y\in D_n} \hat{f}(y|x)$. The maximisation is thus performed on a set of finite cardinality, and can be quickly implemented. According to the asymptotics developed in these papers, one has that $\tilde{\theta}(x) - \hat{\theta}(x) \overset{a.s.}{\to} 0$ as $n \to \infty$, under suitable regularity conditions.

5.3 Prediction by intervals

5.3.1 Determination of the level by a density quantile approach

We follow the approach proposed by Hyndman [80]. In order to determine the set predictor, a first step is to determine, for a given coverage probability α, the corresponding cut-off level f_α of equation (5.2). To that purpose, assume x is fixed. For Y with conditional density $f(y|x)$, define the random variable $Z = f(Y|x)$. Then,

$$Y \in \mathcal{C}_\alpha \Leftrightarrow f(Y|x) \geq f_\alpha \Leftrightarrow Z \geq f_\alpha$$

Therefore, $P(Y \in \mathcal{C}_\alpha) = \alpha \Leftrightarrow P(Z \geq f_\alpha) = \alpha$. So f_α is the $1-\alpha$ quantile of Z. It thus can be estimated by the sample quantile from a set of i.i.d. observations $Z_1, \ldots, Z_n$ from the distribution of $Z = f(Y|X = x)$.

As $f(y|x)$ is unknown, it has to be estimated. Therefore, one can propose two practical approaches to determine the level of the level-set:

1. A Bootstrap technique for estimating f_α is to generate a i.i.d. pseudo-sample $(\hat{Y}_1, \ldots, \hat{Y}_N)$ from the estimated distribution $\hat{f}_n(y|x)$ of $f(y|x)$. Then, $(\hat{Z}_1, \ldots, \hat{Z}_N) := (\hat{f}_n(\hat{Y}_1|x), \ldots, \hat{f}_n(\hat{Y}_N|x))$ will be a i.i.d. pseudo-sample from the distribution of Z. The level f_α is estimated by the sample quantile of the Z_i as

   $$\hat{f}_\alpha := \hat{Z}_{j_\alpha, N}$$

   with $j_\alpha = \lfloor (1-\alpha)N \rfloor$ and where $\hat{Z}_{j,N}$ denotes the jth order statistic of the sample $\hat{Z}_1, \ldots, \hat{Z}_N$.

2. A more direct approach, especially if n is large, is to use the same set of observations $(Y_1, \ldots, Y_n)$, and to calculate the quantile from the synthetic sample $\tilde{Z} = (\tilde{Z}_1, \ldots, \tilde{Z}_n) := (\hat{f}_n(Y_1|x), \ldots, \hat{f}_n(Y_n|x))$. The estimated value is defined analogously by

   $$\hat{f}_\alpha := \tilde{Z}_{j\alpha, n}$$

Remark 5.3 *Set f_α^* the sample quantile obtained from the i.i.d sample $Z_i^* = (f(Y_i|X = x))$, where the conditional density is supposed to be known. The Glivenko-Cantelli result on the sample quantile function (see e.g. Shorack and Wellner [120]) entail that $f_\alpha^* \overset{a.s}{\to} f_\alpha$. Now, as $\hat{f}_n(y|x)$ converge uniformly to the true $f(y|x)$, we conjecture that $|\hat{f}_\alpha - f_\alpha^*| \overset{a.s.}{\to} 0$ as in the proof of proposition 5.2. See the arguments of [80].*

5.3.2 Calculation of predictive intervals

A natural plug-in estimate of the predictive set $\mathcal{C}_\alpha(x)$ defined by equation (5.1), would be to set

$$\mathcal{C}_{\alpha,n}(x) := \{y : \hat{f}_n(y|x) \geq \hat{f}_\alpha\}$$

where $\hat{f}_\alpha$ is the above mentioned estimate of the level f_α. Practically, recall that $\mathcal{C}_\alpha(x)$ is made up of the different subintervals $I_\alpha(x) = [y_\alpha, y^\alpha]$. The corresponding statistical interval estimate $\hat{I}_\alpha(x) = [\hat{y}_\alpha, \hat{y}^\alpha]$ with $\hat{y}_\alpha \leq \hat{y}^\alpha$ is then obtained by solving for y the equation $\hat{f}_n(y|x) = \hat{f}_\alpha$, i.e.

$$\hat{y}_\alpha = \hat{f}_n^{-1}(\hat{f}_\alpha|x) \text{ and } \hat{y}^\alpha = \hat{f}_n^{-1}(\hat{f}_\alpha|x)$$

In the following, we assume the existence of these inverses, that is to say we consider that the level is reasonably chosen.

Convergence of the estimated predictive intervals is then a natural consequence of the convergence of the conditional density estimator, as shown in the following proposition.

Proposition 5.4 *Assume $\hat{f}_\alpha \overset{a.s.}{\to} f_\alpha$. Then $\hat{y}_\alpha \overset{a.s}{\to} y_\alpha$ and $\hat{y}^\alpha \overset{a.s}{\to} y^\alpha$, thus $\lambda(\mathcal{C}_{\alpha,n} \Delta \mathcal{C}_\alpha) \overset{a.s.}{\to} 0$.*

Proof *We do the proof only for $\hat{y}_\alpha$, the proof for $\hat{y}^\alpha$ being similar. Introduce the estimate y_α^* of y_α, had we known the true value f_α, i.e.*

$$\hat{f}_n(y_\alpha^*|x) = f_\alpha$$

Then,

$$P(|\hat{y}_\alpha - y_\alpha| > \epsilon) \leq P(|\hat{y}_\alpha - y_\alpha^*| > \epsilon/2) + P(|y_\alpha^* - y_\alpha| > \epsilon/2)$$

Since $\hat{f}_n^{-1}(.|x)$ *is continuous at* y_α*, for every* $\epsilon > 0$*, there exists a* $\delta_\epsilon > 0$*, such that* $|\hat{f}_n(y|x) - \hat{f}_n(y_\alpha|x)| \leq \delta_\epsilon/2$ *implies* $|y - y_\alpha| \leq \epsilon$*. In particular, for* $y = y_\alpha^*$*, there exists a* δ_ϵ *such that*

$$\begin{aligned} P(|y_\alpha^* - y_\alpha| > \epsilon/2) &\leq P(|\hat{f}_n(y_\alpha^*|x) - \hat{f}_n(y_\alpha|x)| > \delta_\epsilon) \\ &\leq P(|f_\alpha - \hat{f}_n(y_\alpha|x)| > \delta_\epsilon) \\ &\leq P(|f(y_\alpha|x) - \hat{f}_n(y_\alpha|x)| > \delta_\epsilon) \end{aligned}$$

and almost sure convergence of the conditional density estimator yields almost sure convergence of the y_α^* *to* y_α*.*

Similarly, by continuity of $\hat{f}_n^{-1}(.|x)$ *at* y_α^**, there exists* $\delta'_\epsilon > 0$*, such that*

$$P(|\hat{y}_\alpha - y_\alpha^*| > \epsilon/2) \leq P(|\hat{f}_n(\hat{y}_\alpha|x) - \hat{f}_n(y_\alpha^*|x)| > \delta'_\epsilon)$$

and almost sure convergence of $\hat{f}_\alpha \overset{a.s.}{\to} f_\alpha$ *means that* $|\hat{f}_n(\hat{y}_\alpha|x) - \hat{f}_n(y_\alpha^*|x)| \overset{a.s.}{\to} 0$*, yielding* $\hat{y}_\alpha - y_\alpha \overset{a.s.}{\to} 0$*.*

5.4 Prediction by the conditional mean

5.4.1 Consistency of the conditional mean predictor in the bounded case

Assume the support of Y is included in a compact set S. Then, the consistency of the conditional mean predictor derived by integration of the quantile-copula estimator of the conditional density,

$$\hat{m}_1(x) = \int y \hat{f}_n(y|x) dy$$

is a direct corollary of the uniform consistency of the conditional density estimator.

Proposition 5.5 *If* Y *is of bounded support, then* $\int y \hat{f}_n(y|x) dy \overset{a.s.}{\to} E(Y|X = x)$

Proof *The following chain of inequalities holds:*

$$\left| \int y\hat{f}_n(y|x)dy - \int yf(y|x)dy \right| \leq \int_S |y||\hat{f}_n(y|x) - f(y|x)|dy$$
$$\leq \sup_{y\in S} |\hat{f}_n(y|x) - f(y|x)| \int_S |y|dy$$

and the last integral is finite since S is a bounded set. Then, theorem 3.18 yields the desired result.

Remark 5.6 *Note that the assumption of Y having a bounded support is rather restrictive, excluding standard cases such as e.g. the Normal one. The feature that unbounded cases are more difficult to handle are typical of nonparametric regression, see e.g. [66] chapter 10 and 23 and Bosq [21] p. 72 for an illustration. However, in practice, for a finite sample, one can aways assume the Y_i sample is in a compact set. Moreover, we believe the previous result can be extended to the unbounded case by splitting the integral in the proof above on an increasing sequence S_n of compact sets and on its complement S_n^c. To handle the first term of this decomposition, we would require to extend the results on the uniform consistency of the conditional density estimator on an increasing sequence S_n of compact sets, as in Bosq [21] corollary 2.2 and theorem 3.3. To handle the remaining integral term on S_n^c, we would require a moment assumption on Y. We leave open the proof of such a result for further research.*

5.4.2 On the implementation of the conditional mean predictor

Regarding its implementation, note that the conditional mean predictor is constructed as an integrated functional of an estimator of the density. Alternatively to integration over the whole domain of $\hat{m}_1(x) = \int y\hat{f}_n(y|x)dy$, computed in practice by classical numerical integration techniques, several related predictors can be implemented, in the spirit of Gyorfi and Van der Meulen [67]:

- instead of computing the estimator for every y value and integrate it over the entire

space $\mathbb{R}$, one can compute directly the integral on the data set of Y. In other words, as

$$m(x) = \int_{\mathbb{R}} yg(y)c(F(x), G(y))dy = E[Yc(F(x), G(Y))]$$

if one estimate the expectation w.r.t Y by the empirical average on the Y_is and replace the unknown c, F, G by their respective estimators, one gets

$$\begin{aligned}\hat{m}_2(x) &= \frac{1}{n}\sum_{i=1}^{n} Y_i \hat{c}_n(F_n(x), G_n(Y_i)) \\ &= \frac{1}{n}\sum_{i=1}^{n} Y_i \hat{c}_n\left(F_n(x), \frac{rank(Y_i)}{n}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} Y_{i,n} \hat{c}_n\left(F_n(x), \frac{i}{n}\right)\end{aligned}$$

where $Y_{i,n}$ is the ith order statistic.

- Remark one can also obtain the estimator by resampling from the estimated distribution or sample splitting techniques. We leave for further research such an approach.

5.4.3 On the asymptotic equivalence with Stute's smooth k-Nearest Neighbour regression estimator

At last we explore the connection between the conditional mean predictor and Yang and Stute's [145, 127] smoothed k-Nearest neighbour estimator of the regression.

Relation between the Double kernel and Nadaraya-Watson estimators

Indeed, let's recall a well-known connection between the double-kernel estimator of the conditional density and the Nadaraya-Watson estimator of the regression function. Remember that the double kernel estimator of the conditional density writes

$$\hat{f}^{DK}(y|x) = \frac{\sum_{i=1}^{n} a_n^{-2} K\left(\frac{x-X_i}{a_n}\right) K\left(\frac{y-Y_i}{a_n}\right)}{\sum_{i=1}^{n} a_n^{-1} K\left(\frac{x-X_i}{a_n}\right)}.$$

A natural plug-in approach to estimate the regression function $m(x) = E[Y|X = x] = \int yf(y|x)dy$ would be to integrate the double kernel estimator and define $\hat{m}^{DK}(x)$ as

$$\hat{m}^{DK}(x) := \int y\hat{f}^{DK}(y|x)dy = \frac{\sum_{i=1}^n a_n^{-2} K\left(\frac{x-X_i}{a_n}\right) \int_{-\infty}^{\infty} yK\left(\frac{y-Y_i}{a_n}\right) dy}{\sum_{i=1}^n a_n^{-1} K\left(\frac{x-X_i}{a_n}\right)}$$

By the change of variable formula, the above integral is equal to

$$a_n \int_{-\infty}^{\infty} (Y_i + ta_n)K(t)dt = a_n Y_i$$

by the properties of a symmetric kernel K. Therefore, the plug-in estimator reduces to

$$\hat{m}^{DK}(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x-X_i}{a_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{a_n}\right)} = m^{NW}(x)$$

which is the classical Nadaraya-Watson estimator of the regression function.

Presentation of the k-Nearest Neighbour estimators of the regression

Before establishing the connection, we recall the different k-Nearest Neighbour estimators of the regression.

- the classical k-NN estimator of the regression has been introduced by Loftsgaarden and Queensbury [97]. Fix x, and reorder the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ according to the increasing values of $|X_i - x|$ as

  $$(X_{(1,n)}(x), Y_{(1,n)}(x)), \ldots, (X_{(n,n)}(x), Y_{(n,n)}(x))$$

  Then, the estimator is

  $$\tilde{m}^{NN}(x) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y_{(i,n)}(x)$$

  For the density, an estimator is defined analogously as

  $$\hat{f}(x) = \frac{\mu_n[B(x, R(k,x))]}{\lambda[B(x, R(k,x))]}$$

  where μ_n stands for the empirical measure, $B(x, \epsilon)$ the ball of radius ϵ centered at x, and $R(k,x)$ is the distance from x to the k_nth nearest of $X_1, \ldots, X_n$. Stone [124]

showed its universal consistency. See also Liero [95]. Moore and Yackel [98] defined a generalised version as

$$\hat{f}(x) = \frac{1}{nR(k,x)} \sum_{i=1}^{n} K\left(\frac{x - X_i}{R(k,x)}\right)$$

- Yang [145] and Stute [128]'s smoothed version of the k-NN estimator is defined as

$$m^{NN}(x) = \frac{1}{na_n} \sum_{i=1}^{n} Y_i K\left(\frac{F_n(x) - F_n(X_i)}{a_n}\right)$$

The fact that m^{NN} is a smoothed version of the k-NN estimator can be seen with the kernel $K = \mathbb{1}_{[-1/2,1/2]}$. In this case, $m^{NN}(x_0)$ is the average number of Y_i for which, when $X_i \geq x_0$ (say), there exists no more than $k_n := na_n/2$ X_j values with $x_0 \leq X_j$ and such that $X_j < X_i$.

- with the preceding discussions in section 4.1 on the asymptotic deficiency of the empirical c.d.f. with respect to its kernel smoothed version in mind, we can define the doubly smoothed Nearest-Neighbour estimator of the regression function as

$$\hat{m}^{NN}(x) = \frac{1}{na_n} \sum_{i=1}^{n} Y_i K\left(\frac{\hat{F}_n(x) - \hat{F}_n(X_i)}{a_n}\right) \tag{5.4}$$

where $\hat{F}_n$ and $\hat{G}_n$ are the kernel smoothed estimator of the c.d.f. F and G, respectively.

A heuristic connection with the smoothed Rank nearest neighbour estimator

We explore below a heuristic asymptotic connection between the estimator of the regression function $\hat{m}_1(.)$ derived from integration of the quantile-copula estimator of the conditional density and the doubly smoothed nearest-neighbour estimator $\hat{m}^{NN}(.)$ of (5.4) discussed above.

To that end, set $H_n(x,y) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x; Y_i \leq y}$ the bivariate empirical distribution function. The doubly smoothed Nearest-Neighbour estimator of the regression function can

be written in the following integrated form,

$$\hat{m}^{NN}(x_0) = \frac{1}{a_n} \int yK\left(\frac{\hat{F}_n(x_0) - \hat{F}_n(x)}{a_n}\right) dH_n(x,y) \tag{5.5}$$

Similarly, the integrated estimator of the regression function can also be written in the following integrated form as

$$\hat{m}_1(x_0) = \frac{1}{a_n^2} \int_{-\infty}^{\infty} \int y_0 \hat{g}_n(y_0) K\left(\frac{\hat{F}_n(x_0) - \hat{F}_n(x)}{a_n}\right) \times K\left(\frac{\hat{G}_n(y_0) - \hat{G}_n(y)}{a_n}\right) dH_n(x,y)dy_0$$

which can also be written, by Fubini, as

$$\hat{m}_1(x_0) = \frac{1}{a_n} \int I_n(y) K\left(\frac{\hat{F}_n(x_0) - \hat{F}_n(x)}{a_n}\right) dH_n(x,y) \tag{5.6}$$

with

$$I_n(y) := \frac{1}{a_n} \int_{-\infty}^{\infty} y_0 \hat{g}_n(y_0) K\left(\frac{\hat{G}_n(y_0) - \hat{G}_n(y)}{a_n}\right) dy_0$$

Under this integral representation, the two estimators have the same shape, with $I_n(y)$ in (5.6) instead of y in (5.5). The fact that $I_n(y)$ is asymptotically close to y is proved in the following lemma.

Lemma 5.7 *In addition to the assumptions **A,B** of section 3.3.1, assume the density g of Y is such that $0 < g < \infty$, and $g' < \infty$ in a neighbourhood of y, then for a choice of the bandwidth as in theorem 3.5,*

$$|I_n(y) - y| = o_{a.s.}(1)$$

Proof *Since the kernel is of compact support, say $[-1,1]$ w.l.o.g., the integral above is restricted to the set of y_0 values in $S_n(y) := \{y_0 : |\hat{G}_n(y) - \hat{G}_n(y_0)| < a_n\}$. With the fact that $\hat{G}_n$ is a smooth c.d.f. and that the kernel is positive, the application $y_0 \to t = \frac{\hat{G}_n(y_0) - \hat{G}_n(y)}{a_n}$ is a diffeomorphism from $S_n(y)$ to $(-1,1)$, and one gets by the change of*

variable formula,

$$I_n(y) = \int_{-1}^{1} \hat{G}_n^{-1}[\hat{G}_n(y) + ta_n]K(t)dt$$

Define the deterministic counterpart of I_n *by*

$$J_n(y) := \int_{-1}^{1} G^{-1}[G(y) + ta_n]K(t)dt$$

Step 1 : Approximation of I_n ***by*** J_n***.***

We have, almost surely, that

$$\begin{aligned}|I_n(y) - J_n(y)| &\leq \int_{-1}^{1} |\hat{G}_n^{-1}[\hat{G}_n(y) + ta_n] - G^{-1}[\hat{G}_n(y) + ta_n]|K(t)dt \\ &+ \int_{-1}^{1} |G^{-1}[\hat{G}_n(y) + ta_n] - G^{-1}[G(y) + ta_n]|K(t)dt \\ &\leq ||\hat{G}_n^{-1} - G^{-1}||_\infty \int_{-1}^{1} K(t)dt + \int_{-1}^{1} \frac{|\hat{G}_n(y) - G(y)|}{g \circ G^{-1}(\zeta_{y,n}(t))}K(t)dt\end{aligned}$$

where $\zeta_{y,n}(t)$ *lies between* $\hat{G}_n(y) + ta_n$ *and* $G(y) + ta_n$*, for* $g \neq 0$ *in a neighbourhood of* y*.*

By the first mean value theorem, there exists a $c \in (-1, 1)$ *such that*

$$\int_{-1}^{1} \frac{|\hat{G}_n(y) - G(y)|}{g \circ G^{-1}(\zeta_{y,n}(t))}K(t)dt = \frac{|\hat{G}_n(y) - G(y)|}{g \circ G^{-1}(\zeta_{y,n}(c))} \int_{-1}^{1} K(t)dt$$

Since $||\hat{G}_n - G|| \overset{a.s.}{\to} 0$ *and* $\zeta_{y,n}(c)$ *is at most between* $\hat{G}_n(y) \pm a_n$ *and* $G(y) \pm a_n$*, we get that* $\zeta_{y,n}(c) \overset{a.s.}{\to} G(y)$*. By continuity,*

$$\frac{1}{g \circ G^{-1}(\zeta_{y,n}(c))} \overset{a.s.}{\to} \frac{1}{g \circ G^{-1} \circ G(y)} = \frac{1}{g(y)}$$

provided $g \neq 0$ *in a neighbourhood of* y*. In turn,*

$$\frac{|\hat{G}_n(y) - G(y)|}{g \circ G^{-1}(\zeta_{y,n}(c))} \int_{-1}^{1} K(t)dt = \frac{|\hat{G}_n(y) - G(y)|}{g(y)}(1 + o_{a.s.}(1))$$

and the following inequality holds,

$$|I_n(y) - J_n(y)| \leq ||\hat{G}_n^{-1} - G^{-1}||_\infty + \frac{||\hat{G}_n - G||_\infty}{g(y)}(1 + o_{a.s.}(1)). \tag{5.7}$$

Step 2 : Approximation of J_n***.***

By Taylor expansion, there exist a $\eta = \eta_{n,y}(t)$ *between* $G(y)$ *and* $G(y) + ta_n$ *such that*

$$G^{-1}(G(y) + ta_n) = y + \frac{ta_n}{g(y)} - \frac{t^2 a_n^2}{2} \frac{g' \circ G^{-1}(\eta)}{g^3 \circ G^{-1}(\eta)}$$

provided $0 < g < \infty$, *and* $g' < \infty$ *in a neighbourhood of* y. *By the properties of the second order kernel,* J_n *thus writes*

$$J_n(y) = y - a_n^2/2 \int_{-1}^{1} t^2 \frac{g' \circ G^{-1}(\eta)}{g^3 \circ G^{-1}(\eta)} K(t)dt.$$

By the first mean value theorem, there exists a $d \in (-1, 1)$ *such that*

$$\int_{-1}^{1} t^2 \frac{g' \circ G^{-1}(\eta)}{g^3 \circ G^{-1}(\eta)} K(t)dt = \frac{g' \circ G^{-1}(\eta_{n,y}(d))}{g^3 \circ G^{-1}(\eta_{n,y}(d))} \int_{-1}^{1} t^2 K(t)dt.$$

By assumptions on the second-order kernel, $\int_{-1}^{1} t^2 K(t)dt = C < \infty$, *for a constant, say,* C. *Moreover, since* $\eta_{n,y}(d)$ *is at most between* $G(y)$ *and* $G(y) \pm a_n$, *one gets that* $\sup_y |\eta_{n,y}(d) - G(y)| \leq a_n$, *thus* $\eta_{n,y}(d) \to G(y)$ *as* $n \to \infty$. *By continuity,* $\frac{g' \circ G^{-1}(\eta_{n,y}(d))}{g^3 \circ G^{-1}(\eta_{n,y}(d))} \to \frac{g'(y)}{g^3(y)}$ *as* $n \to \infty$. *Therefore,*

$$J_n(y) = y - \frac{a_n^2}{2} \frac{g'(y)}{g^3(y)} C + o_{a.s}(a_n^2). \tag{5.8}$$

Step 3 : Conclusion

By combining equations (5.7) and (5.8),

$$|I_n(y) - y| \leq ||\hat{G}_n^{-1} - G^{-1}||_\infty + \frac{||\hat{G}_n - G||_\infty}{g(y)}(1 + o_{a.s.}(1)) + \frac{a_n^2}{2} \left| \frac{g'(y)}{g^3(y)} \right| C + o_{a.s}(a_n^2)$$

and the law of the iterated logarithm entails that the first two terms are of order $(\ln \ln n/n)^{1/2}$, *which are negligible compared to* a_n^2. *In turn,*

$$|I_n(y) - y| = O_{a.s.}(a_n^2)$$

which is more than must be proved.

This lemma thus allows to give some substance to the heuristic connection between the two estimators, which have approximately the same shape.

On conditional empirical distribution function

A more general way to view the resemblances and dissimilarities between the quantile copula estimator and Stute's one, is to have an approach based on conditional empirical process. The conditional empirical process indexed by the function ϕ is defined as

$$\mathbb{G}_\phi(y|x) = \sum_{i=1}^{n} w_{ni}(x, X_1, \ldots, X_n)\phi(Y_i, y)$$

where $\{w_{ni}\}_{i=1}^n$ is a sequence of weights. In particular, for the class of functions $\phi = I_{Y_i \leq y}$, the conditional distribution function can be written as

$$\hat{F}(y|x) = \sum_{i=1}^{n} w_{ni}(x, X_1, \ldots, X_n) I_{Y_i \leq y}$$

- With the Nadara-Watson weights defined as

  $$w_{ni} = \frac{K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)}$$

  the Nadaraya-Watson regression and double kernel conditional density estimators are defined respectively as

  $$\hat{f}(y|x)^{DK} = \int K_h(y-t)d\hat{F}(t|x)$$
  $$\text{and } \hat{m}^{DK} = \int t d\hat{F}(t|x)$$

- In Stute's nearest neighbour approach, the weights are different,

  $$w'_{ni} = \frac{1}{a_n} K\left(\frac{F_n(x) - F_n(X_i)}{a_n}\right)$$

  and the estimators are defined analaguously, as

  $$\hat{f}(y|x) = \int K_a(y-t)d\hat{F}(t|x)$$
  $$\text{and } \hat{m}(x) = \int t d\hat{F}(t|x)$$

  Notice that the Nadaraya-Watson weights sum to 1, whereas the nearest neighbour weights asymptotically sum up to 1.

- In the quantile copula approach, the weights are also w'_{ni} as Stute's approach, but the density and regression estimators are defined in a slightly different manner, as

$$\hat{f}(y|x) = \int \hat{g}(y) K_a(G_n(y) - G_n(t)) d\hat{F}(t|x)$$
$$\text{and } \hat{m}(x) = \int y\hat{g}(y) d\hat{F}(y|x)$$

Remark 5.8 *With y replaced by 1 in the integral, the same argument shows, by noting that the density of the uniform variable $F(X)$ is 1, that*

$$\int_{-\infty}^{\infty} \hat{f}(y|x)dy \approx \frac{1}{na_n} \sum_{i=1}^{n} K\left(\frac{\hat{F}_n(x) - \hat{F}_n(X_i)}{a_n}\right) \overset{a.s.}{\to} 1$$

that is to say the quantile-copula density estimator is asymptotically a density, in the sense it integrates to unity.

Chapter 6

Perspectives and possible applications

In this conclusive section, we would like to sketch some perspectives for further research and possible applications of the proposed estimator. These are developed in a more or less lengthy manner, depending on the current degree of advancement of our ongoing research. In section 6.1, we briefly present how the proposed estimator could be refined by using some more sophisticated methods of estimation of the copula density. We mention some possible extensions to the multivariate case and to the dependent framework in section 6.2. General guidelines on estimating the conditional cumulative distribution functions are given in section 6.3. At last, possible fields of application such as extremes or missing data where we believe the proposed estimator could be an interesting starting point are presented in section 6.4.

6.1 Variants and mathematical refinements of the conditional density estimator

Considering the well known fact that, under regularity conditions, many nonparametric estimators can be written as kernel estimates (see Bosq and Bleuez [25] and Terell and Scott [134]), we conjecture that a whole family of possible estimators can be built by changing the method of estimation of the margins and/or of the copula density. We believe that a general theorem could be written, e.g. in the spirit of the delta sequences estimators of Walter and Blum [142] and Liero [94].

Among possible other nonparametric estimators of the copula density based on the pseudo-data (U_i, V_i), one can mention,

- **Projection estimates on Wavelet bases :** A possible area of future research would be to use the copula density estimator based on wavelets of Genest, Masiello and Tribouley [60], which has the appealing feature of being free of boundary bias. Its adaptive version is based on thresholding procedures in Autin al. [9]. The goal is to considerably diminish the regularity hypothesis on the densities, and to build an adaptive estimator by selecting automatically the smoothing parameters to fit optimally a large class of densities, e.g. Besov spaces. See also the work of Lacour [87] with model selection techniques *à la* Birgé and Massart [18, 12].

- **Local Polynomial estimators :** Another approach would be to estimate the copula density by the Local polynomial method of Fan and Gijbels [50]. Its extension in the multivariate case has been investigated by Abdous and Bensaid [1]. See also Abdous and Ghoudi [2]. Recall that the local polynomial estimator of the copula function is defined as the minimiser of the following score,

$$L_n(x) = \int_{[0,1]^2} K_H(u-x)[C_n(u) - P(u-x)]^2 du$$

where $C_n(.) = n^{-1}\sum_{i=1}^{n} \mathbb{1}_{(F_n(X_i),G_n(Y_i))\leq .}$ is the empirical cumulative distribution function based on the approximate observations, $K_H(u) = |H|^{-1}K(H^{-1}u)$ a bivariate kernel with bandwidth matrix H, P is a multivariate polynomial of given degree, and x and u are bivariate in that context. Note that the polynomial has to be of degree at least one to estimate the copula density and that one can replace the empirical c.d.f C_n by any smooth estimate converging a.s. to it.

An alternative approach would be to replace C_n by the empirical measure $\mu_n(.,.) = n^{-1}\sum_{i=1}^{n} \delta_{F_n(X_i)}(.)\delta_{G_n(Y_i)}(.)$ to estimate directly the copula density and its derivatives.

The local polynomial method, when the integral to be minimised is restricted to the support of the function to be estimated, has the advantage of being free of boundary bias, thus correcting the bias issue of the kernel method.

- **Semi-parametric locally parametric estimator of the copula density :** Eventually, another possible approach which would be to make a compromise between the parametric and non parametric estimation methods by implementing locally parametric non parametric methods *à la* Hjort and Jones [74] and Loader [96].

  Consider a family of densities $\{f(.,\theta), \theta \in \Theta\}$. Introduce the local likelihood

  $$L_n(x,\theta) = \int K_h(t-x)w(x,t,\theta)[dF_n(t) - f(t,\theta)dt]$$

  The local likelihood estimator is defined as $\hat{\theta}(x) = \arg\max L_n(x,\theta)$, and gives an estimator of the density $f(x,\hat{\theta}(x))$ which is the local best approximation of the density by the parametric family $\{f(.,\theta), \theta \in \Theta\}$. The weight function can be chosen to be e.g. $w(x,t,\theta) = \frac{\partial}{\partial\theta}\log f(t,\theta)$, to get a local version of the maximum likelihood estimation. We believe this setup can be adapted to estimate the copula density and/or the marginal distribution.

Indeed, for the copula density, one would define the pseudo likelihood,

$$L_n(u,\theta) = \int K_h(t-u)w(u,t,\theta)[dC_n(t) - d_t c(t,\theta)]$$

where $t, u \in \mathbb{R}^2$ are bidimensionnal and $C_n(.) = n^{-1}\sum_{i=1}^n \mathbb{1}_{(F_n(X_i),G_n(Y_i))\leq .}$ is the empirical distribution function based on the approximate observations. However, if this approach is used for the copula density, the possible low rate of convergence of the copula density estimator may yield the analysis of the conditional density estimator difficult.

- **Beta Kernels with a data-driven bandwidth selection :** The goal is to build an efficient estimator in the finite sample setting. To do so, an asymptotic analysis with the Beta kernels (to alleviate the bias issues) is to be done to allow to establish a data-dependent smoothing selection procedure.

6.2 Extensions of the proposed estimator

6.2.1 Extension to the dependent case

In the mixing framework, analogues of the Chung-Smirnov property and convergence results of the kernel density estimator do exist. By coupling arguments as in [111], one should be able to extend the estimator in the dependent framework. A case of particular interest is when the X, Y variables corresponds to the X_n, X_{n+1} of a stationary Markov chain, which gives an estimate of the transition density. Such an estimate should serve as a building block to make inference, tests and prediction in fields such as e.g. econometrics.

6.2.2 Extension to the multivariate case

Unfortunately, this copula approach by the probability integral transform breaks down as soon as the dimension of the input variable $X = (X_1, \ldots, X_d)$ is higher than 1. Indeed,

assuming the densities exists, Sklar's [122] formula for the densities writes

$$f_{Y,X_1,\dots,X_d}(y,x_1,\dots,x_d) = g(y)f_{X_1}(x_1)\dots f_{X_d}(x_d)c(G(y),F_{X_1}(x_1),\dots,F_{X_d}(x_d))$$

Since in general, $f_{X_1,\dots,X_d}(x_1,\dots,x_d) \neq f_{X_1}(x_1)\dots f_{X_d}(x_d)$, unless the trivial and uninteresting case where the components are independent, a product formula like (3.3) is not available.

However, by applying Sklar's formula to the components of X, one can write

$$f_{X_1,\dots,X_d}(x_1,\dots,x_d) = f_{X_1}(x_1)\dots f_{X_d}(x_d)c_X(F_{X_1}(x_1),\dots,F_{X_d}(x_d))$$

where we noted c_X the copula density of X. We can thus obtain the following formula for the conditional density in the multivariate case,

$$f_{Y|X_1,\dots,X_d}(x_1,\dots,x_d) = g(y)\frac{c(G(y),F_{X_1}(x_1),\dots,F_{X_d}(x_d))}{c_X(F_{X_1}(x_1),\dots,F_{X_d}(x_d))}$$

From this we can obtain an estimator by plugging estimates of each one of the quantities

$$\hat{f}_{Y|X_1,\dots,X_d}(x_1,\dots,x_d) = \hat{g}(y)\frac{\hat{c}(G_n(y),F_{1,n}(x_1),\dots,F_{d,n}(x_d))}{\hat{c}_X(F_{1,n}(x_1),\dots,F_{d,n}(x_d))}.$$

The fraction of copula densities estimator has the structure of a Nadaraya-Watson type estimator on the transformed approximate data. Although we still get an estimator which is a product of an estimate of the density of Y and a copula estimate term, we lose the overall product shape of the univariate case, as the copula term is now a ratio of estimators. As a consequence, it appears unclear whether an improvement over the classical estimators can be reached, and such a study is left for further research.

Note that, due to the curse of dimensionality, it is maybe less interesting to use a fully nonparametric approach. A possible compromise would be to use a single index model approach, i.e. to assume that there is a single d dimensional vector θ such that $f(y|x) = f(y| < \theta, x >)$.

6.3 Estimation of the conditional cumulative distribution function

The Quantile transform and copula representation approach which lead to the proposed esimator of the conditional density can also be used to estimate the conditional cumulative distribution function $F(y|x)$.

6.3.1 On two possible approaches

- Approach by integration of the conditional density : from the estimator of the conditional density, one can integrate to obtain an estimate of the conditional c.d.f.: $F_{Y|X}(x,y) = \int_{-\infty}^{y} f_{Y|X}(x,u)du$ entails $\hat{F}_{Y|X}(x,y) = \int_{-\infty}^{y} \hat{f}_{Y|X}(x,u)du$, and the latter integral can be computed by numerical integration. One can also get an approximate explicit formula by noting that,

$$F_{Y|X}(x,y) = \int_{-\infty}^{y} g(t)c(F(x),G(t)) = E[\mathbb{1}_{Y\leq y}c(F(x),G(Y))]$$

  and replacing the expectation by the empirical mean and the unknown c, F, G by their respective estimators $\hat{c}_n$, F_n, G_n:

$$\hat{F}_{Y|X}(x,y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i\leq y}\hat{c}_n(F_n(x),G_n(Y_i))$$

- Direct approach : from Sklar's copula formula, one has by a change of variable that

$$\begin{aligned} F_{Y|X=x}(x,y) &= \int_{-\infty}^{y} g(t)c(F(x),G(t))dt \\ &= \int_{0}^{G(y)} c(F(x),v))dv \end{aligned}$$

  To construct an estimator, one may replace the unknown quantities c, F, G by the

respective estimators $\hat{c}_n$, F_n, G_n to obtain

$$\begin{aligned}\hat{F}_{Y|X=x}(x,y) &= \int_0^{G_n(y)} \hat{c}_n(F_n(x), v))dv \\ &= \frac{1}{na_n^2}\sum_{i=1}^{n} K\left(\frac{F_n(x)-F_n(X_i)}{a_n}\right)\int_0^{G_n(y)} K\left(\frac{v-G_n(Y_i)}{a_n}\right)dv\end{aligned}$$

By setting $K^I(t) = \int_{-\infty}^{t} K(v)dv$, the latter estimator writes, after integration, as

$$\hat{F}_{Y|X=x}(x,y) = \frac{1}{na_n}\sum_{i=1}^{n} K\left(\frac{F_n(x)-F_n(X_i)}{a_n}\right)\left[K^I\left(\frac{G_n(y)-G_n(Y_i)}{a_n}\right) - K^I\left(\frac{-G_n(Y_i)}{a_n}\right)\right]$$

6.3.2 Application to point and interval prediction

From an estimate $\hat{F}(.|x)$ of the conditional cumulative distribution function $F(.|x)$, one can similarly to the plug-in approach of chapter 5 defines statistical predictors as follows

- for point prediction: define the median predictor $M(x)$ such that

$$\hat{F}_{Y|X=x}(M(x), x) = 1/2$$

- for interval prediction : the predictive interval is defined between two quantiles, e.g. 2.5% and 97.5%. Define the $0 < \alpha < 1$ quantile by $t_\alpha(x) = \inf\{y \in \mathbb{R}, F(y|x) \geq \alpha\}$. From an estimate $\hat{t}_\alpha(x) = \inf\{y \in \mathbb{R}, \hat{F}(y|x) \geq \alpha\}$, one can construct the interval $[\hat{t}_\alpha(x), \hat{t}_{1-\alpha}(x)]$, which is a $(1-2\alpha)$ predictive confidence band.

See also Bosq and Blanke [24], chapter 5, for the prediction of the conditional distribution and computation of predictive intervals when the marginal law is known.

6.4 Some possible practical applications

To bridge the gap between theory and practice, we would like to mention some possible practical fields where we think the proposed estimator could be beneficial.

6.4.1 Missing data

Suppose we have some missing values of the pairs (X, Y), in the sense that for part of the sample we only observe one of the two components. This could happen for example if a censoring mechanism occur. The sample can be decomposed in the following three parts,

$$D_{XY} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$
$$D_X = \{X_{n+1}, \ldots, X_{n+p}\}$$
$$D_Y = \{Y_{n+1}, \ldots, Y_{n+q}\}$$

where D_X and D_Y are the data where one component is missing. In other approaches, one has to scrap the sample parts D_X and D_Y, and perform the estimation of the conditional density only on D_{XY}, thus losing some potential valuable information on X and Y.
The proposed copula approach allows to make use of this partial information. Define the X and Y merged samples as

$$D'_X = \{X_1, \ldots, X_{n+p}\}$$
$$D'_Y = \{Y_1, \ldots, Y_{n+q}\}$$

Then the conditional density estimator can be built by

- estimating F and G on D'_X and D'_Y by $\hat{F}$ and $\hat{G}$ respectively,
- estimating g of Y by $\hat{g}$ on D'_Y,
- estimating the copula density on D_{XY}, after transforming this data set by the aforementioned $\hat{F}$ and $\hat{G}$.

6.4.2 Estimation of conditional density for rare events

The comparative analysis and numerical simulation of section 4.3 showed some possible promising results when the explanatory variable takes possibly large values. We believe our estimator could be a good starting point to design an estimator tailored to situations

where we are interested in assessing whether Y takes large values given the fact that X is large too.

The method of estimation would be to combine results and method of standard extreme theory together with this nonparametric approach, and could be briefly sketched as follows:

- replace the empirical c.d.f. F_n and G_n by some estimators of the cumulative distribution functions F and G, suited to the estimation of tail probabilities such as Hill's [73] or Pickands' [103] estimator,
- for the copula density, the estimation is located around the corner $(1, 1)$ and one should use either a nonparametric estimator or use a model from multivariate extreme copula for $C(1 - u_n, v)$ with $u_n \to 0$.

Such a study could be of practical importance in the following fields

- Environmental applications : For preventing floods, one can be interested in understanding the impact of big waves occurring in windstorms on the water level, or the impact of the amount of rain on the flow of a river. See e.g. De Haan and Sinha [40]. Other possible applications could be energy production given wind strength for windmills, electrical consumption of households given temperature, impact of a given factor on pollution, etc...
- Insurance and finance applications : One can imagine this approach could be useful in finance, e.g. to detect the possibility of bankruptcy of an insurance company from a large claim, occurring with small frequency.
- Reliability applications : to assess how the failure of one component would impact the failure of another one, etc.

Bibliography

[1] Abdous, B., and Bensaid, E. Multivariate local polynomial fitting for a probability distribution function and its partial derivatives. *J. Nonparametr. Statist. 13*, 1 (2001), 77–94.

[2] Abdous, B., and Ghoudi, K. Non-parametric estimators of multivariate extreme dependence functions. *J. Nonparametr. Stat. 17*, 8 (2005), 915–935.

[3] Abraham, C., Biau, G., and Cadre, B. Simple estimation of the mode of a multivariate density. *Canad. J. Statist. 31*, 1 (2003), 23–34.

[4] Abraham, C., Biau, G., and Cadre, B. On the asymptotic properties of a simple estimate of the mode. *ESAIM Probab. Stat. 8* (2004), 1–11 (electronic).

[5] Aggarwal, O. P. Some minimax invariant procedures for estimating a cumulative distribution function. *Ann. Math. Statist. 26* (1955), 450–463.

[6] Algoet, P. Universal schemes for prediction, gambling and portfolio selection. *Ann. Probab. 20*, 2 (1992), 901–941.

[7] Algoet, P. H. The strong law of large numbers for sequential decisions under uncertainty. *IEEE Trans. Inform. Theory 40*, 3 (1994), 609–633.

[8] Altman, N., and Léger, C. Bandwidth selection for kernel distribution function estimation. *J. Statist. Plann. Inference 46*, 2 (1995), 195–214.

[9] Autin, F., Le Pennec, E., and Tribouley, K. Thresholding methods to estimate the copula density. *To appear* (2008).

[10] Azzalini, A. A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika 68*, 1 (1981), 326–328.

[11] Babu, G. J., Canty, A. J., and Chaubey, Y. P. Application of Bernstein polynomials for smooth estimation of a distribution and density function. *J. Statist. Plann. Inference 105*, 2 (2002), 377–392.

[12] Barron, A., Birgé, L., and Massart, P. Risk bounds for model selection via penalization. *Probab. Theory Related Fields 113*, 3 (1999), 301–413.

[13] Bashtannyk, D. M., and Hyndman, R. J. Bandwidth selection for kernel conditional density estimation. *Comput. Statist. Data Anal. 36*, 3 (2001), 279–298.

[14] Bernstein, S. Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités. *Commun. Soc. Math. Kharkow t. 13*, 2 (1912-1913), 1–2.

[15] Bertrand-Retali, M. Convergence uniforme d'un estimateur de la densité par la méthode du noyau. *Rev. Roumaine Math. Pures et Appliquées. 23*, 3 (1978), 361–385.

[16] Bickel, P. J., and Rosenblatt, M. On some global measures of the deviations of density function estimates. *Ann. Statist. 1* (1973), 1071–1095.

[17] Billingsley, P. *Convergence of probability measures.* John Wiley & Sons Inc., New York, 1968.

[18] Birgé, L., and Massart, P. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam.* Springer, New York, 1997, pp. 55–87.

[19] Blackwell, D. An analog of the minimax theorem for vector payoffs. *Pacific J. Math. 6* (1956), 1–8.

[20] Bosq, D. Estimation de la densité conditionnelle et de la régression. *C. R. Acad. Sci. Paris Sér. A-B 269* (1969), A661–A664.

[21] Bosq, D. *Nonparametric statistics for stochastic processes*, second ed., vol. 110 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1998. Estimation and prediction.

[22] Bosq, D. Sufficiency and efficiency in statistical prediction. *Statist. Probab. Lett. 77*, 3 (2007), 280–287.

[23] Bosq, D. A note on asymptotic parametric prediction. *J. Statist. Plann. Inference accepted* (2008).

[24] Bosq, D., and Blanke, D. *Inference and prediction in large dimensions*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 2007.

[25] Bosq, D., and Bleuez, J. Étude d'une classe d'estimateurs non-paramétriques de la densité. *Ann. Inst. H. Poincaré Sect. B (N.S.) 14*, 4 (1978), 479–498 (1979).

[26] Bosq, D., and Lecoutre, J.-P. *Théorie de l'estimation fonctionnelle*. Collection Economie et Statistiques Avancées. Economica, Paris, 1987.

[27] Bowman, A., Hall, P., and Prvan, T. Bandwidth selection for the smoothing of distribution functions. *Biometrika 85*, 4 (1998), 799–808.

[28] Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. *Time series analysis*, third ed. Prentice Hall Inc., Englewood Cliffs, NJ, 1994. Forecasting and control.

[29] Brockwell, P. J., and Davis, R. A. *Time series: theory and methods*, second ed. Springer Series in Statistics. Springer-Verlag, New York, 1991.

[30] Caires, S., and Ferreira, J. A. On the non-parametric prediction of conditionally stationary sequences. *Stat. Inference Stoch. Process. 8*, 2 (2005), 151–184.

[31] Carroll, R. J., and Ruppert, D. *Transformation and weighting in regression.* Monographs on Statistics and Applied Probability. Chapman and Hall, New York, 1988.

[32] Cesa-Bianchi, N., and Lugosi, G. *Prediction, Learning and Games.* Cambridge University Press, New York, 2006.

[33] Charpentier, A. *Dependence structure and limiting results: some applications in finance and insurance.* Phd Thesis, Katholieke Universiteit Leuven, 2006.

[34] Chatfield, C. *The analysis of time series*, sixth ed. Chapman & Hall/CRC Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL, 2004. An introduction.

[35] Chen, S. X. Beta kernel estimators for density functions. *Comput. Statist. Data Anal. 31*, 2 (1999), 131–145.

[36] Chen, X., Fan, Y., and Tsyrennikov, V. Efficient estimation of semiparametric multivariate copula models. *J. Amer. Statist. Assoc. 101*, 475 (2006), 1228–1240.

[37] Collomb, G., Härdle, W., and Hassani, S. A note on prediction via estimation of the conditional mode function. *J. Statist. Plann. Inference 15*, 2 (1987), 227–236.

[38] Cover, T. M. Open problems in information theory. *IEEE USSR Joint Workshop on Information Theory* (1975), 35–36.

[39] De Gooijer, J. G., and Zerom, D. On conditional density estimation. *Statist. Neerlandica 57*, 2 (2003), 159–176.

[40] de Haan, L., and Sinha, A. K. Estimating the probability of a rare event. *Ann. Statist. 27*, 2 (1999), 732–759.

[41] Deheuvels, P. Conditions nécessaires et suffisantes de convergence ponctuelle presque sûre et uniforme presque sûre des estimateurs de la densité. *C. R. Acad. Sci. Paris Sér. A 278* (1974), 1217–1220.

[42] Deheuvels, P. La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance. *Acad. Roy. Belg. Bull. Cl. Sci. (5) 65*, 6 (1979), 274–292.

[43] Deheuvels, P. A Kolmogorov-Smirnov type test for independence and multivariate samples. *Rev. Roumaine Math. Pures Appl. 26*, 2 (1981), 213–226.

[44] Delyon, B. Limit theorem for mixing processes. *Tech. Rept. 546. IRISA, Rennes I* (1990).

[45] Devroye, L., and Lugosi, G. *Combinatorial methods in density estimation.* Springer Series in Statistics. Springer-Verlag, New York, 2001.

[46] Doukhan, P. *Mixing*, vol. 85 of *Lecture Notes in Statistics.* Springer-Verlag, New York, 1994. Properties and examples.

[47] Eddy, W. F. Optimum kernel estimators of the mode. *Ann. Statist. 8*, 4 (1980), 870–882.

[48] Eddy, W. F. The asymptotic distributions of kernel estimators of the mode. *Z. Wahrsch. Verw. Gebiete 59*, 3 (1982), 279–290.

[49] Efromovich, S. Conditional density estimation in a regression setting. *Ann. Statist. 35*, 6 (2007), 2504–2535.

[50] Fan, J., and Gijbels, I. *Local polynomial modelling and its applications*, vol. 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1996.

[51] Fan, J., and Yao, Q. *Nonlinear time series*, second ed. Springer Series in Statistics. Springer-Verlag, New York, 2005. Nonparametric and parametric methods.

[52] Fan, J., Yao, Q., and Tong, H. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika 83*, 1 (1996), 189–206.

[53] Fan, J., and Yim, T. H. A crossvalidation method for estimating conditional densities. *Biometrika 91*, 4 (2004), 819–834.

[54] Fermanian, J.-D. Goodness-of-fit tests for copulas. *J. Multivariate Anal. 95*, 1 (2005), 119–152.

[55] Fermanian, J.-D., and O., S. Nonparametric estimation of copulas for time series. *Journal of Risk 5*, 4 (2003), 25–54.

[56] Fermanian, J.-D., Radulović, D., and Wegkamp, M. Weak convergence of empirical copula processes. *Bernoulli 10*, 5 (2004), 847–860.

[57] Ferraty, F., and Vieu, P. *Nonparametric functional data analysis*. Springer Series in Statistics. Springer, New York, 2006. Theory and practice.

[58] Gasser, T., and Müller, H.-G. Kernel estimation of regression functions. In *Smoothing techniques for curve estimation (Proc. Workshop, Heidelberg, 1979)*, vol. 757 of *Lecture Notes in Math.* Springer, Berlin, 1979, pp. 23–68.

[59] Geffroy, J. Sur la convergence uniforme des estimateurs d'une densité de probabilité. *Unpublished manuscript* (1974).

[60] Genest, C., Masiello, E., and Tribouley, K. Estimating copula densities through wavelets. *To appear* (2008).

[61] Genest, C., and Werker, B. J. M. Conditions for the asymptotic semiparametric efficiency of an omnibus estimator of dependence parameters in copula models. In *Distributions with given marginals and statistical modelling*. Kluwer Acad. Publ., Dordrecht, 2002, pp. 103–112.

[62] Gijbels, I., and Mielniczuk, J. Estimating the density of a copula function. *Comm. Statist. Theory Methods 19*, 2 (1990), 445–464.

[63] Gouriéroux, C., and Monfort, A. *Séries Temporelles et modèles dynamiques.* Économie et Statistiques Avancées. Economica, Paris, 1990.

[64] Gustafsonn, J., Hagmann, M., Nielsen, J., and Scaillet, O. Local transformation kernel density estimation of loss distributions. *Forthcoming in Journal of Business and Economic Statistics* (2007).

[65] Györfi, L., and Kohler, M. Nonparametric estimation of conditional distributions. *IEEE Trans. Inform. Theory 53*, 5 (2007), 1872–1879.

[66] Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. *A distribution-free theory of nonparametric regression.* Springer Series in Statistics. Springer-Verlag, New York, 2002.

[67] Györfi, L., and van der Meulen, E. C. On the nonparametric estimation of the entropy functional. In *Nonparametric functional estimation and related topics (Spetses, 1990)*, vol. 335 of *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.* Kluwer Acad. Publ., Dordrecht, 1991, pp. 81–95.

[68] Hall, P., Racine, J., and Li, Q. Cross-validation and the estimation of conditional probability densities. *J. Amer. Statist. Assoc. 99*, 468 (2004), 1015–1026.

[69] Hall, P., Wolff, R. C. L., and Yao, Q. Methods for estimating a conditional distribution function. *J. Amer. Statist. Assoc. 94*, 445 (1999), 154–163.

[70] Hallin, M., and Werker, B. J. M. Semi-parametric efficiency, distribution-freeness and invariance. *Bernoulli 9*, 1 (2003), 137–165.

[71] Härdle, W., Müller, M., Sperlich, S., and Werwatz, A. *Nonparametric and semiparametric models*. Springer Series in Statistics. Springer-Verlag, New York, 2004.

[72] Has′minskiĭ, R. Z. A lower bound for risks of nonparametric density estimates in the uniform metric. *Teor. Veroyatnost. i Primenen. 23*, 4 (1978), 824–828.

[73] Hill, B. M. A simple general approach to inference about the tail of a distribution. *Ann. Statist. 3*, 5 (1975), 1163–1174.

[74] Hjort, N. L., and Jones, M. C. Locally parametric nonparametric density estimation. *Ann. Statist. 24*, 4 (1996), 1619–1647.

[75] Hodges, Jr., J. L., and Lehmann, E. L. Deficiency. *Ann. Math. Statist. 41* (1970), 783–801.

[76] Hoff, P. D. Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Statist. 1*, 1 (2007), 265–283.

[77] Hössjer, O., and Ruppert, D. Taylor series approximations of transformation kernel density estimators. *J. Nonparametr. Statist. 4*, 2 (1994), 165–177.

[78] Hössjer, O., and Ruppert, D. Asymptotics for the transformation kernel density estimator. *Ann. Statist. 23*, 4 (1995), 1198–1222.

[79] Huber, P. J. *Robust statistics*. John Wiley & Sons Inc., New York, 1981. Wiley Series in Probability and Mathematical Statistics.

[80] Hyndman, R. J. Computing and graphing highest density regions. *American Statistician 50* (1996), 120–126.

[81] Hyndman, R. J., Bashtannyk, D. M., and Grunwald, G. K. Estimating and visualizing conditional densities. *J. Comput. Graph. Statist. 5*, 4 (1996), 315–336.

[82] Hyndman, R. J., and Yao, Q. Nonparametric estimation and symmetry tests for conditional density functions. *J. Nonparametr. Stat. 14*, 3 (2002), 259–278.

[83] Joe, H. *Multivariate models and dependence concepts*, vol. 73 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1997.

[84] Kallianpur, G. Nonlinear stochastic filtering: a brief survey. *Tr. Semin. im. I. G. Petrovskogo 23* (2003), 206–218, 411.

[85] Klaassen, C. A. J., and Wellner, J. A. Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli 3*, 1 (1997), 55–77.

[86] Konakov, V. D. The asymptotic normality of the mode of multivariate distributions. *Teor. Verojatnost. i Primenen. 18* (1973), 836–842.

[87] Lacour, C. Adaptive estimation of the transition density of a Markov chain. *Ann. Inst. H. Poincaré Probab. Statist. 43*, 5 (2007), 571–597.

[88] Laksaci, A. Convergence en moyenne quadratique de l'estimateur à noyau de la densité conditionnelle avec variable explicative fonctionnelle. *Ann. I.S.U.P. 51*, 3 (2007), 69–80 (2008).

[89] Laksaci, A. Erreur quadratique de l'estimateur à noyau de la densité conditionnelle à variable explicative fonctionnelle. *C. R. Math. Acad. Sci. Paris 345*, 3 (2007), 171–175.

[90] Leblanc, A. Asymptotic efficiency of the Bernstein polynomial estimator of a distribution function. *Submitted manuscript* (2008).

[91] Lecoutre, J.-P., and Ould-Said, E. Estimation de la densité conditionnelle et de la fonction de hasard conditionnelle pour un processus fortement mélangeant avec censure. *C. R. Acad. Sci. Paris Sér. I Math. 314*, 4 (1992), 295–300.

[92] Lehmann, E. L., and Casella, G. *Theory of point estimation*, second ed. Springer Texts in Statistics. Springer-Verlag, New York, 1998.

[93] Li, Q., and Racine, J. S. *Nonparametric econometrics.* Princeton University Press, Princeton, NJ, 2007. Theory and practice.

[94] Liero, H. Strong uniform consistency of nonparametric regression function estimates. *Probab. Theory Related Fields 82*, 4 (1989), 587–614.

[95] Liero, H. A note on the asymptotic behaviour of the distance of the k_nth nearest neighbour. *Statistics 24*, 3 (1993), 235–243.

[96] Loader, C. R. Local likelihood density estimation. *Ann. Statist. 24*, 4 (1996), 1602–1618.

[97] Loftsgaarden, D. O., and Quesenberry, C. P. A nonparametric estimate of a multivariate density function. *Ann. Math. Statist. 36* (1965), 1049–1051.

[98] Moore, D. S., and Yackel, J. W. Consistency properties of nearest neighbor density function estimators. *Ann. Statist. 5*, 1 (1977), 143–154.

[99] Nadaraja, È. A. On a regression estimate. *Teor. Verojatnost. i Primenen. 9* (1964), 157–159.

[100] Nadaraya, È. A. *Nonparametric estimation of probability densities and regression curves*, vol. 20 of *Mathematics and its Applications (Soviet Series).* Kluwer Aca-

demic Publishers Group, Dordrecht, 1989. Translated from the Russian by Samuel Kotz.

[101] Nelsen, R. B. *An introduction to copulas*, second ed. Springer Series in Statistics. Springer, New York, 2006.

[102] Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Statist. 33* (1962), 1065–1076.

[103] Pickands, III, J. Statistical inference using extreme order statistics. *Ann. Statist. 3* (1975), 119–131.

[104] Polonik, W., and Yao, Q. Conditional minimum volume predictive regions for stochastic processes. *J. Amer. Statist. Assoc. 95*, 450 (2000), 509–519.

[105] Prakasa Rao, B. L. S. *Nonparametric functional estimation.* Probability and Mathematical Statistics. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, 1983.

[106] Prakasa Rao, B. L. S. Estimation of distribution and density functions by generalized Bernstein polynomials. *Indian J. Pure Appl. Math. 36*, 2 (2005), 63–88.

[107] Priestley, M. B., and Chao, M. T. Non-parametric function fitting. *J. Roy. Statist. Soc. Ser. B 34* (1972), 385–392.

[108] Radulović, D., and Wegkamp, M. Weak convergence of smoothed empirical processes: beyond Donsker classes. In *High dimensional probability, II (Seattle, WA, 1999)*, vol. 47 of *Progr. Probab.* Birkhäuser Boston, Boston, MA, 2000, pp. 89–105.

[109] Reiss, R.-D. Nonparametric estimation of smooth distribution functions. *Scand. J. Statist. 8*, 2 (1981), 116–119.

[110] Reiss, R.-D. *Approximate distributions of order statistics.* Springer Series in Statistics. Springer-Verlag, New York, 1989. With applications to nonparametric statistics.

[111] Rio, E. *Théorie asymptotique des processus aléatoires faiblement dépendants*, vol. 31 of *Mathématiques & Applications (Berlin) [Mathematics & Applications].* Springer-Verlag, Berlin, 2000.

[112] Rosenblatt, M. A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci. U. S. A. 42* (1956), 43–47.

[113] Rosenblatt, M. Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist. 27* (1956), 832–837.

[114] Rosenblatt, M. Conditional probability density and regression estimators. In *Multivariate Analysis, II (Proc. Second Internat. Sympos., Dayton, Ohio, 1968).* Academic Press, New York, 1969, pp. 25–31.

[115] Roussas, G. G. Nonparametric estimation of the transition distribution function of a Markov process. *Ann. Math. Statist. 40* (1969), 1386–1400.

[116] Ruppert, D., and Cline, D. B. H. Bias reduction in kernel density estimation by smoothed empirical transformations. *Ann. Statist. 22*, 1 (1994), 185–210.

[117] Rüschendorf, L. Asymptotic distributions of multivariate rank order statistics. *Ann. Statist. 4*, 5 (1976), 912–923.

[118] Sarda, P. Smoothing parameter selection for smooth distribution functions. *J. Statist. Plann. Inference 35*, 1 (1993), 65–75.

[119] Scott, D. W. *Multivariate density estimation.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons

Inc., New York, 1992. Theory, practice, and visualization, A Wiley-Interscience Publication.

[120] Shorack, G. R., and Wellner, J. A. *Empirical processes with applications to statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1986.

[121] Silverman, B. W. *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986.

[122] Sklar, M. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris 8* (1959), 229–231.

[123] Stein, E. M. *Singular integrals and differentiability properties of functions*. Princeton Mathematical Series, No. 30. Princeton University Press, Princeton, N.J., 1970.

[124] Stone, C. J. Consistent nonparametric regression. *Ann. Statist. 5*, 4 (1977), 595–645. With discussion and a reply by the author.

[125] Stone, C. J. Optimal rates of convergence for nonparametric estimators. *Ann. Statist. 8*, 6 (1980), 1348–1360.

[126] Stute, W. A law of the logarithm for kernel density estimators. *Ann. Probab. 10*, 2 (1982), 414–422.

[127] Stute, W. The oscillation behavior of empirical processes. *Ann. Probab. 10*, 1 (1982), 86–107.

[128] Stute, W. Asymptotic normality of nearest neighbor regression function estimates. *Ann. Statist. 12*, 3 (1984), 917–926.

[129] Stute, W. The oscillation behavior of empirical processes: the multivariate case. *Ann. Probab. 12*, 2 (1984), 361–379.

[130] Stute, W. Conditional empirical processes. *Ann. Statist. 14*, 2 (1986), 638–647.

[131] Stute, W. On almost sure convergence of conditional empirical distribution functions. *Ann. Probab. 14*, 3 (1986), 891–901.

[132] Tang, Y., and Ghosal, S. A consistent nonparametric Bayesian procedure for estimating autoregressive conditional densities. *Comput. Statist. Data Anal. 51*, 9 (2007), 4424–4437.

[133] Tenreiro, C. On the asymptotic behaviour of the ISE for automatic kernel distribution estimators. *J. Nonparametr. Stat. 15*, 4-5 (2003), 485–504.

[134] Terrell, G. R., and Scott, D. W. Variable kernel density estimation. *Ann. Statist. 20*, 3 (1992), 1236–1265.

[135] Tsukahara, H. Semiparametric estimation in copula models. *Canad. J. Statist. 33*, 3 (2005), 357–375.

[136] van der Vaart, A. W. *Asymptotic statistics*, vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[137] van der Vaart, A. W., and Wellner, J. A. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[138] Viennet, G. Inequalities for absolutely regular sequences: application to density estimation. *Probab. Theory Related Fields 107*, 4 (1997), 467–492.

[139] Vitale, R. A. Bernstein polynomial approach to density function estimation. In *Statistical inference and related topics (Proc. Summer Res. Inst. Statist. Inference for Stochastic Processes, Indiana Univ., Bloomington, Ind., 1974, Vol. 2; dedicated to Z. W. Birnbaum)*. Academic Press, New York, 1975, pp. 87–99.

[140] Volkonskiĭ, V. A., and Rozanov, Y. A. Some limit theorems for random functions. I. *Teor. Veroyatnost. i Primenen 4* (1959), 186–207.

[141] Wald, A. *Statistical Decision Functions.* John Wiley & Sons Inc., New York, N. Y., 1950.

[142] Walter, G., and Blum, J. Probability density estimation using delta sequences. *Ann. Statist. 7*, 2 (1979), 328–340.

[143] Wand, M. P., and Jones, M. C. *Kernel smoothing*, vol. 60 of *Monographs on Statistics and Applied Probability.* Chapman and Hall Ltd., London, 1995.

[144] Watson, G. S. Smooth regression analysis. *Sankhyā Ser. A 26* (1964), 359–372.

[145] Yang, S.-S. Linear functions of concomitants of order statistics with application to nonparametric estimation of a regression function. *J. Amer. Statist. Assoc. 76*, 375 (1981), 658–662.

[146] Youndjé, É. Propriétés de convergence de l'estimateur à noyau de la densité conditionnelle. *Rev. Roumaine Math. Pures Appl. 41*, 7-8 (1996), 535–566.

[147] Youndjé, É., Sarda, P., and Vieu, P. Estimateur à noyau d'une densité conditionnelle: choix de la fenêtre pour des observations dépendantes. *C. R. Acad. Sci. Paris Sér. I Math. 316*, 9 (1993), 935–938.

[148] Youndjé, É., Sarda, P., and Vieu, P. Validation croisée pour l'estimation non-paramétrique de la densité conditionnelle. *Publ. Inst. Statist. Univ. Paris 38*, 1 (1994), 57–80.

[149] Yu, Z. Edgeworth expansion for nearest neighbor-kernel estimate and random weighting approximation of conditional density. *Appl. Math. J. Chinese Univ. Ser. B 15*, 2 (2000), 167–172.