Estimation de régularité locale

Rémi Servien

Université Montpellier II

–SCIENCES ET TECHNIQUES DU LANGUEDOC–

T H È S E

pour obtenir le grade de

Docteur de l'Université Montpellier II

Discipline	:	Mathématiques appliquées
Ecole Doctorale	:	**I**nformation, **S**tructures, **S**ystèmes
Formation Doctorale	:	Biostatistique

Estimation de régularité locale

par

Rémi SERVIEN

Soutenue publiquement et obtenue avec mention **très honorable** le **12/03/2010** devant le jury composé de :

C. ABRAHAM	Professeur, SupAgro Montpellier	Examinateur
A. BERLINET	Professeur, Université Montpellier II	Directeur de Thèse
G. BIAU	Professeur, Université Paris VI	Examinateur
A. MAS	Professeur, Université Montpellier II	Président
B. PELLETIER	Professeur, Université Rennes II	Rapporteur
P. SARDA	Professeur, Université Toulouse III	Rapporteur

Remerciements

Pendant la majeure partie de ma thèse, j'attendais avec impatience ce moment libérateur où j'écrirais les remerciements. Pour n'importe quel doctorant débutant cette étape marque en effet la fin de la rédaction et une grosse marche supplémentaire gravie vers le doctorat. Et même si on se rend compte, au fur et à mesure de l'avancée de notre travail, que l'obtention de la thèse n'est qu'une étape supplémentaire franchie et non une fin en soi, ce n'est pas sans émotions que je me lance dans cette tâche en m'excusant par avance d'éventuels oublis.

Je tiens tout d'abord à exprimer toute ma gratitude à Alain Berlinet qui m'a accompagné, aidé et corrigé tout au long de cette thèse. Il m'a appris à la fois la rigueur scientifique et l'ouverture d'esprit et je reste impressionné par l'étendue de ces connaissances, tant théoriques qu'appliquées. J'espère que notre collaboration ne s'arrêtera pas avec cette thèse.

Merci infiniment à Bruno Pelletier et Pascal Sarda qui m'ont fait l'honneur d'accepter de rapporter cette thèse, malgré un emploi du temps que je devine chargé. Je remercie de même Christophe Abraham, Gérard Biau et André Mas qui ont spontanément accepté de participer à mon jury.

Je remercie également l'ensemble des membres du département de Mathématiques, et particulièrement les gens avec qui j'ai collaboré pour mes vacations comme Christian Lavergne, Catherine Trottier ou Cyrille Joutard, ainsi que Nicolas Molinari avec qui j'ai fait mon monitorat-conseil au sein du CHU de Nîmes.

J'adresses mes plus sincères remerciements à Pierre Cartigny et l'ensemble de l'UMR ASB du campus ENSAM-INRA de Montpellier. J'ai eu la chance de travailler dans des conditions privilégiées et j'ai pu rencontrer là-bas des gens formidables avec qui j'ai pu nouer des liens solides.

Je remercie également le CNRS et la région Languedoc-Roussilon qui ont financé cette thèse.

Sur un plan plus personnel, je commencerai avec une pensée particulière pour les doctorants, ATER et Maîtres de conférences avec qui j'ai tissé de forts liens d'amitié. Je me suis demandé si une avalanche de prénoms était ici nécessaire. Je pense qu'elle l'est donc à Soffana, Guillaume, Olivier, Benoît, Hilde, Gwladys, Elamine, Julien, Khader, Kevin, Leslie, Ahmad, Chady, Nadia, Romain, Rémy, Afaf, Chloé, Véra, Virginie, Guillemette etc .... Merci.

Il aura bien mérité un paragraphe pour lui tout seul, mon co-bureau pendant ces 3 ans de travail, Thomas. Pour tes corrections (à part l'orthographe), tes réponses à la fameuse question bête du vendredi soir, tous nos échanges en général, le screenquizz (dont je remercie également les participants pour leur aide précieuse comme heffy, vaudou, worm, magnum, bebert, mimo, boule, Forest, Heart, mast, node, syd, Gilles, grbill ...), Michel Delpech et les mongolfières (du plus au moins sérieux), un grand merci.

Je n'aurais pas pu arriver là sans l'amour que me portent mes parents et sans la totale liberté qu'ils m'ont laissé dans mes études malgré mes changements de direction. Merci Maman, Papa mais je n'oublie pas non plus mes grands-parents ainsi que Benja et Manon.

Je souhaite également remercier tous mes amis qui n'ont pas encore été cités notamment le GRC dans son ensemble, les Insaliens ou les Gruissannais. Je remercie plus particulièrement ceux qui ne m'ont pas trop invité à des soirées, me laissant du temps pour travailler, ainsi que ceux qui assisteront à ma soutenance sans y comprendre un traître mot.

Et, enfin, comment ne pas finir en remerciant Christelle. Pour avoir su m'épauler pendant les moments difficiles et m'avoir apporté du bonheur au quotidien, je te remercie du fond du coeur.

Table des matières

Introduction générale **1**
1.1 Cadre général 1
1.2 Normalité asymptotique d'estimateurs de la densité 7
1.3 Estimation du mode pour des densités non continues 9
1.4 Estimateurs de l'indice de régularité utilisant des estimateurs de la fonction de répartition 10
Bibliographie de l'Introduction 12

I Normalité asymptotique d'estimateurs de la densité **15**
1.1 Introduction 17
1.2 Estimateur des k_n-plus proches voisins 20
1.2.1 Définition de l'estimateur 20
1.2.2 Conditions Nécessaires et Suffisantes de Normalité Asymptotique 22
1.3 Nouvelle définition pour l'indice de régularité 25
1.3.1 Fonctions de répartition C^1 25
1.3.2 Densité C^0 28
1.3.3 Discontinuité du second ordre 30
1.3.4 Nouvelle définition 31
1.4 Application à l'estimateur des k_n-plus proches voisins 32
1.5 L'histogramme 33
1.5.1 Construction de l'estimateur 33
1.5.2 Résultat sur la Normalité Asymptotique 34
1.6 Simulations 35
1.6.1 Indice de régularité 35
1.6.2 Estimation de la densité 39
Bibliographie 41

II Estimation du mode pour des densités non continues 43

1.1 Introduction . . . 45
1.2 Convergence . . . 47
1.2.1 Hypothèses et Notations . . . 47
1.2.2 Convergence de θ_n . . . 49
1.2.3 Vitesse de convergence de θ_n . . . 49
1.2.4 Lien entre l'indice de pic et l'indice de régularité . . . 51
1.3 Preuves . . . 52
1.4 Simulations . . . 54
1.4.1 Étude de f . . . 54
1.4.2 Calcul du mode . . . 55
1.4.3 Étude du point $x = 0$ . . . 56
1.4.4 Vérification des hypothèses . . . 56
1.4.5 Résultats . . . 56
Bibliographie . . . 62

III Estimateurs de l'indice de régularité utilisant des estimateurs de la fonction de répartition 63

1 Estimateurs de l'indice de régularité 65
1.1 Introduction . . . 65
1.2 Résultats de convergence . . . 67
1.2.1 L'estimateur empirique . . . 67
1.3 Preuves . . . 68
1.4 Simulations . . . 71
1.4.1 Estimateur des k_n-plus proches voisins . . . 71
1.4.2 Comparaison des estimateurs de l'indice de régularité . 77
Bibliographie . . . 85

2 Estimation de la fonction de répartition : revue bibliographique 87
2.1 Introduction . . . 87
2.2 Un estimateur naturel : la fonction de répartition empirique . 89
2.3 Estimateurs par lissage local . . . 94
2.4 Estimateur à noyau . . . 95
2.5 Estimateurs splines . . . 97
2.6 Les Support Vector Machines . . . 98
2.7 Le level-crossing . . . 99
2.8 Les Systèmes de Fonctions Itérées . . . 100

2.9 D'autres estimateurs 102
2.10 Fonction de répartition conditionnelle 103
2.11 Données biaisées 105
2.12 Conclusion 107
2.13 Simulations 107
Bibliographie 115

Conclusion et perspectives **117**

Introduction générale

Le sujet principal de cette thèse est lié au problème général de dérivation des mesures (Rudin [15], Dudley [8]). Il trouve ses motivations dans l'étude de problèmes d'estimation quand les conditions de régularité habituelles ne sont pas vérifiées. En effet, de nombreux théorèmes de convergence font intervenir des hypothèses de continuité qui ne sont en pratique pas toujours satisfaites. Nous utilisons donc des conditions moins contraignantes permettant de plus d'étudier la régularité de la densité associée à la mesure considérée.
Un paramètre α_x appelé *indice de régularité* apparaît lorsqu'on essaie d'étudier localement le comportement d'une fonction de densité dérivée d'une mesure quelconque. Ce paramètre de régularité étant fortement local, son estimation est difficile. Nous nous attacherons dans cette thèse à étudier certains problèmes d'estimation non paramétrique où cet indice intervient et à trouver différents estimateurs convergents de α_x.

1.1 Cadre général

Pour $d \geq 1$, définissons $\mathcal{B}(\mathbb{R}^d)$ le champ borélien de $\mathbb{R}^d$. Nous considérons μ une mesure de probabilité sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Notons λ la mesure de Lebesgue sur $\mathbb{R}^d$ muni d'une norme notée $||.||$. Soit x un point de $\mathbb{R}^d$, δ un réel positif et $B_\delta(x)$ la boule ouverte de centre x et de rayon δ. Afin de mesurer le comportement local de $\mu(B_\delta(x))$ par rapport à $\lambda(B_\delta(x))$ nous pouvons considérer le quotient de ces deux mesures. Ainsi, si pour x fixé la limite suivante

$$\boxed{f(x) = \lim_{\delta \to 0} \frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))}} \tag{1.1}$$

existe, alors x est appelé *point de Lebesgue* de la mesure μ. Si μ est absolument continue par rapport à λ nous pouvons sélectionner parmi toutes les densités

obtenues à partir de μ, une densité particulière f, qui satisfait (1.1) en tout point où cette limite existe. Il est important de noter que la notion de point de Lebesgue est plus large que la notion de continuité. Elle permet donc d'élargir certains résultats en diminuant les contraintes sur les fonctions à estimer. Dans ce contexte, Berlinet et Levallois [4] définissent un point ρ-régulier de la mesure μ comme un point de Lebesgue x de μ tel que

$$\boxed{\left|\frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} - f(x)\right| \leq \rho(\delta),} \tag{1.2}$$

où ρ est une fonction mesurable telle que $\lim_{\delta\downarrow 0}\rho(\delta) = 0$.
Par exemple, si $d = 1$ et si la mesure μ a une densité f avec une dérivée f' bornée par une constante quelconque C_x dans le voisinage de x, alors nous avons $\rho(\delta) = C_x\delta$ et x est ρ-régulier. Il est aussi clair que, si f est une fonction localement hölderienne en x avec un exposant α_x, cela implique $\rho(\delta) = C_x/(\alpha_x + 1)\delta^{\alpha_x}$. De plus, il est possible de trouver des exemples de mesures ρ-régulières mais avec un mauvais comportement local de la densité, comme des discontinuités du second ordre. Par exemple la fonction f_1 (Figure 1.1), définie pour $x \in [-1, 1] \neq 0$ par

$$f_1(x) = \frac{\sqrt{|x|} - \cos(1/x) + 2x\sin(1/x) + 2}{c},$$

avec $c = 4 + 4/3 + 2\sin(1)$. Pour cette densité, 0 est un point où nous n'avons ni limite à droite ni limite à gauche. Il est cependant possible de démontrer que le rapport entre μ_1 la mesure associée à f_1 et λ vaut $2/c$ en 0. Par conséquent, 0 est bien un point de Lebesgue de μ_1. Pour d'autres exemples, nous renvoyons le lecteur à l'article de Berlinet et Levallois [4].

Précisons que la fonction ρ de (1.2) n'est pas unique et dépend de la norme choisie sur $\mathbb{R}^d$. Il est par ailleurs possible d'aller plus loin que la relation (1.2) et de considérer qu'en x, point de Lebesgue de la mesure μ, nous ayons

$$\boxed{\frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} = f(x) + C_x\delta^{\alpha_x} + o(\delta^{\alpha_x}) \text{ quand } \delta \downarrow 0,} \tag{1.3}$$

où C_x est une constante différente de 0 et α_x un nombre réel strictement positif que nous appellerons *indice de régularité*. Ces constantes sont alors

uniques et, trivialement, cette relation implique la ρ-régularité en x avec $\rho(\delta) \sim C_x \delta^{\alpha_x}$. Cette relation définissant l'indice de régularité joue un rôle central tout au long de cette thèse.

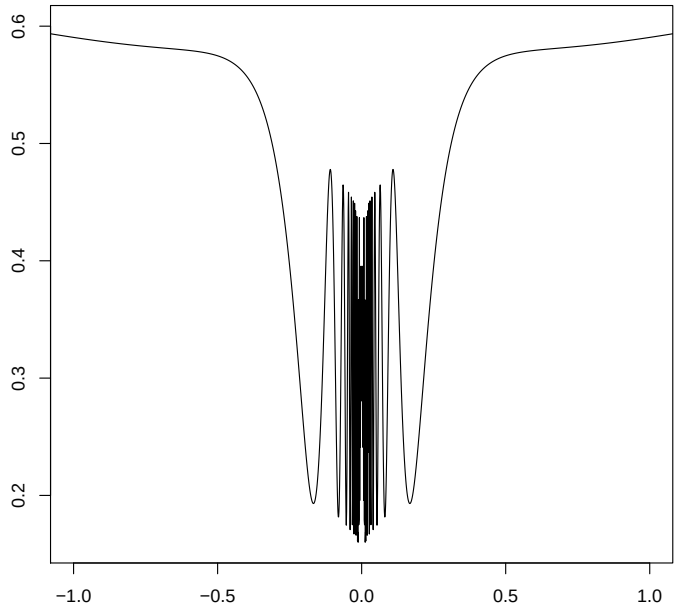


FIGURE 1.1 – Graphique de la densité f_1.

Notons que l'indice α_x reflète le degré de régularité de la mesure μ par rapport à la mesure de Lebesgue λ. En effet, plus α_x sera grand, plus la dérivée de μ sera lisse autour du point x. En reprenant rapidement l'exemple de la fonction f_1 précédente, nous avions remarqué que 0 était bien un point de Lebesgue de μ. Il est également possible de démontrer que son indice de régularité vaut 1/2.

La connaissance de cet indice est intéressante en pratique pour étudier le comportement local de la mesure. En effet, il nous donne d'importantes indications sur le caractère plus ou moins lisse d'une mesure autour du point

x. Il est important de noter que α_x existe dans le cas d'une densité non nécessairement continue, ceci nous garantissant un large cadre de travail. Il intervient également dans différents problèmes d'estimation intimement liés au caractère lisse ou non lisse de la mesure. Nous pouvons citer notamment l'estimation du nombre optimal de voisins pour l'estimateur des k_n-plus proches voisins de la densité que nous développons ci-dessous.

Un problème souvent rencontré en statistique est l'estimation d'une densité f ou de sa mesure de probabilité μ (à valeurs dans $\mathbb{R}^d$) à partir d'un échantillon de variables aléatoires réelles $X_1, ..., X_n$ indépendantes et identiquement distribuées (i.i.d.) et de même loi μ. Un estimateur de la densité simple et facile à mettre en oeuvre est l'estimateur des k_n-plus proches voisins défini de la façon suivante. Soit $(k_n)_{n\geq 1}$ une suite d'entiers positifs, l'estimateur des k_n-plus proches voisins de f au point x est alors

$$f_{k_n}(x) = \frac{k_n}{n\lambda(\overline{B}_{k_n}(x))},$$

où $\overline{B}_{k_n}(x)$ est la plus petite boule fermée de centre x contenant au moins k_n points de l'échantillon. L'entier k_n joue donc le rôle d'un paramètre de lissage. Plus explicitement, plus k_n sera grand plus f_{k_n} sera lisse.

En analyse discriminante, Fix and Hodges [9] ont introduit la règle de classification basée sur les plus proches voisins (voir aussi plus récemment Devroye, Györfi et Lugosi [7]). Son application à l'estimation de la densité fut étudiée par la suite par Loftsgaarden et Quesenberry [10] et Moore et Yackel [11]. Pour une introduction complète sur le sujet, nous renvoyons le lecteur au livre de Bosq et Lecoutre [5] ainsi qu'à l'article de Berlinet et Levallois [4].

Le choix du nombre de voisins k_n est un problème difficile, d'autant plus si la densité à estimer est non lisse. Dans ce cas, van Es [16] a obtenu des résultats pour le choix de la fenêtre de l'estimateur à noyau en utilisant la validation croisée.

Beirlant, Berlinet et Biau [3] ont présenté une approche nouvelle en optimi-

sant l'erreur quadratique moyenne

$$\Delta_n(x) = E\left(f_{k_n}(x) - f(x)\right)^2,$$

où l'espérance se calcule sur l'échantillon $X_1, ..., X_n$. La valeur de k_n minimisant $\Delta_n(x)$ en un point de Lebesgue x vérifiant (1.3) est alors

$$k_n^* = \left(\frac{dV_d^{2\alpha_x/d}}{2\alpha_x C_x^2} f^{2+2\alpha_x/d}(x)\right)^{d/(d+2\alpha_x)} n^{2\alpha_x/(d+2\alpha_x)},$$

où V_d est le volume de la boule unité de dimension d et C_x et α_x les constantes de l'égalité (1.3). Nous remarquons que k_n^* augmente avec α_x, ce qui n'est pas surprenant du fait du lien entre α_x et la mesure de probabilité.

Dans le même article, Beirlant, Berlinet et Biau proposent un estimateur $f_{k_n}^{(a)}$ pour minimiser le biais de l'estimateur des k_n-plus proches voisins. Cet estimateur est de la forme

$$f_{k_n}^{(a)}(x) = f_{k_n}(x) - g(\alpha_x, d, f_{k_n}(x)),$$

g étant une fonction facile à calculer lorsque ses paramètres sont connus. L'indice de régularité α_x tient une part importante dans ce terme correctif qui améliore sensiblement l'estimation de la densité comme nous pouvons le voir sur l'exemple de la Figure 1.2 ci-dessous.

L'estimation $f_{k_n}^{(a)}(0)$ est ici meilleure et beaucoup moins dépendante du choix du nombre de voisins k_n que l'estimation classique $f_{k_n}(0)$. Cependant, l'estimateur de α_x proposé dans cet article et utilisé dans cette estimation (pour α_x dans $f_{k_n}^{(a)}(0)$ nous choisissons la médiane de son estimation pour des valeurs de k_n entre 150 et 500) reste insatisfaisant au point de vue des simulations, comme nous le verrons en détail dans le troisième chapitre de cette thèse. Il est en effet très sensible à la valeur de k_n choisie. Un meilleur estimateur de α_x permettrait donc d'augmenter l'efficacité de $f_{k_n}^{(a)}$ sans condition de continuité nécessaire.

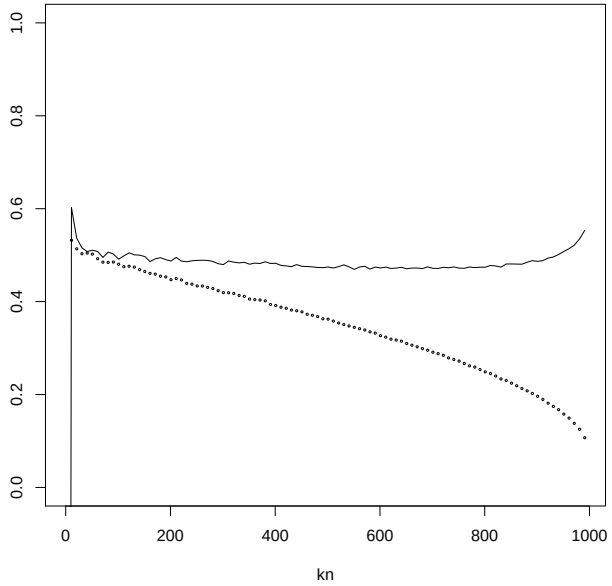


FIGURE 1.2 – Estimations de la densité $f(x) = 0.5\exp(|x|)$ en $x = 0$ par $f_{k_n}^{(a)}$ (trait plein) et f_{k_n} (en pointillés) pour $n = 1000$.

L'indice de régularité tient donc une place importante dans ce problème d'estimation fonctionnelle. Il nous fournit également d'importantes indications sur le comportement d'une mesure autour d'un point de Lebesgue. Afin de l'étudier, nous avons décider de structurer notre travail en trois parties.

La première partie, intitulée **Normalité asymptotique d'estimateurs de la densité**, nous permettra d'étudier un nouveau problème d'estimation fonctionnelle où l'indice de régularité intervient. En effet, nous étudions les lois asymptotiques de l'estimateur des k_n-plus proches voisins de la densité et de l'histogramme. Nous remarquons que l'indice de régularité apparaît dans ces problèmes.

Dans la seconde partie, qui se nomme **Estimateur du mode de densités non continues**, nous définissons un estimateur du mode ne nécessitant pas de conditions de continuité sur la densité étudiée. Nous examinons également

comment l'indice de régularité intervient dans cette problématique.

Nous nous attachons dans la troisième partie, intitulée **Estimateurs de l'indice de régularité à l'aide d'estimateurs de la fonction de répartition**, à obtenir différents estimateurs de l'indice de régularité. Cette partie se divise en deux chapitres. Le premier est consacré à l'obtention et la comparaison de plusieurs estimateurs de l'indice de régularité en utilisant des estimateurs de la fonction de répartition. Dans le deuxième, nous réalisons une revue bibliographique sur les différents estimateurs de la fonction de répartition afin que la classe d'estimateurs de l'indice de régularité soit potentiellement la plus large possible.

1.2 Normalité asymptotique d'estimateurs de la densité

En notant $B_n(x) = B(x, R_n(x))$ la plus petite boule fermée de centre x contenant au moins k_n points de l'échantillon, Berlinet et Levallois [4] démontrent le résultat suivant : sous les conditions de convergence de f_{k_n}

$$\lim_{n\to\infty} k_n = \infty \quad \text{et} \quad \lim_{n\to\infty} \frac{k_n}{n} = \infty,$$

si x est un point ρ-régulier de la mesure μ avec $f(x) > 0$, alors la condition

$$\sqrt{k_n}\rho(R_n(x)) \xrightarrow{P} 0$$

lorsque n tend vers l'infini implique la convergence en distribution de la variable aléatoire

$$T_n(x) = \sqrt{k_n}\frac{f_{k_n}(x) - f(x)}{f(x)}$$

vers une loi $\mathcal{N}(0, 1)$.

Ils obtiennent ensuite comme corollaire de ce résultat que sous les conditions de convergence évoquées précédemment et si de plus la densité f est lipschitzienne d'ordre $\alpha > 0$ alors la condition

$$\lim_{n\to\infty} \frac{k_n^{1+1/2\alpha}}{n} = 0$$

implique la convergence en distribution de $T_n(x)$ vers une loi $\mathcal{N}(0,1)$.
Une bonne estimation de l'indice de régularité est donc primordiale afin d'obtenir la vérification des conditions et, par conséquent, la normalité asymptotique de $T_n(x)$. En effet, comme nous pourrons le remarquer sur un exemple, en ne tenant pas compte de la spécificité de la mesure étudiée et de son indice de régularité il est possible de commettre une importante erreur d'estimation de la densité. Nous nous attachons tout d'abord dans ce chapitre à élargir les résultats de Berlinet et Levallois [4] sur l'estimateur des k_n-plus proches voisins. Nous obtenons une condition nécessaire et suffisante à la normalité asymptotique de la loi limite de $T_n(x)$ ainsi que l'expression de cette loi. Puis, après l'étude de différents exemples, nous donnons une nouvelle définition moins contraignante de l'indice de régularité. Nous définissons l'ensemble E_x par

$$\begin{aligned} E_x \;=\; & \Big\{\alpha \geq 0 \text{ tel que } \exists C>0, \exists \lambda_0>0, \text{ tels que } \forall I \in \mathcal{I}_x \text{ vérifiant } \lambda(I)<\lambda_0 \\ & \text{on ait } \left|\frac{\mu(I)}{\lambda(I)} - f(x)\right| \leq C\lambda(I)^\alpha \Big\} \end{aligned}$$

où $\mathcal{I}_x$ est un ensemble d'intervalles contenant x. S'il existe un réel α_x vérifiant

$$\alpha_x = \sup E_x$$

alors α_x sera l'indice de régularité de la mesure μ au point x. A l'aide de cette définition, nous obtenons une condition suffisante pour la normalité asymptotique de $T_n(x)$. Cette nouvelle définition n'utilisant pas de développements de boules centrées en x, le point d'estimation, elle nous permet également d'étudier un estimateur de la densité tel que l'histogramme. En notant $f_h(x)$ l'estimateur de la densité par la méthode de l'histogramme, nous énonçons des conditions suffisantes pour obtenir la normalité asymptotique de

$$H_n(x) = \sqrt{nh_n}\frac{f_h(x) - f(x)}{\sqrt{f(x)}}.$$

Enfin, nous testons nos résultats sur des données simulées.

1.3 Estimation du mode pour des densités non continues

Considérons une mesure de probabilité μ dans $\mathbb{R}^d$ à partir de laquelle nous obtenons une densité f. Nous nous intéressons au problème de l'estimation du mode θ de f à partir d'un échantillon i.i.d. $S_n = \{X_1, \ldots, X_n\}$ distribué selon μ. Formellement, le mode d'une densité est l'argument où la densité est maximisée, c'est-à-dire la valeur pour laquelle on approche le plus (voire on atteint) la borne supérieure essentielle de f. Ce problème a suscité une littérature considérable (Parzen [13], Nadaraya [12], Devroye [6]) comme nous pourrons le voir plus en détails dans l'introduction de cette seconde partie. Néanmoins, tous les estimateurs nécessitent des conditions de régularité forte, au minimum la continuité autour du mode θ. Ainsi, pour tout $x \in \mathbb{R}^d$, on définit l'estimateur à noyau de la densité f_{h_n}(Akaike[2], Rosenblatt [14], Parzen [13]) par

$$f_{h_n}(x) = \frac{1}{nh_n^d} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h_n}\right),$$

où k est un noyau et h_n la fenêtre de lissage strictement supérieure à 0 et telle que h_n tend vers 0 quand n tend vers l'infini. Abraham, Biau et Cadre [1] définissent un estimateur θ_n du mode par

$$\theta_n \in \arg\max_{S_n}(f_{h_n}),$$

plus précisément,

$$\theta_n \in \left\{ x \in S_n : f_{h_n}(x) = \max_{1 \le i \le n} f_{h_n}(X_i) \right\}.$$

Cet estimateur converge alors presque sûrement vers le mode θ si f est continue autour du mode. Ils obtiennent ensuite une bonne vitesse de convergence et un intervalle de confiance asymptotique. En nous appuyant sur leurs travaux, nous démontrons la convergence de θ_n sous des conditions ne portant pas sur la régularité de la densité f. Nous montrons ensuite que ces conditions sont vérifiées pour $\theta \in V$ où V est un intervalle où tout point est un point ρ-régulier. Nous obtenons également des intervalles de confiance asymptotiques sous certaines hypothèses supplémentaires, qui sont vérifiées pour θ

appartenant à un intervalle de points de Lebesgue admettant un indice de régularité. Enfin, nous terminons ce chapitre par une étude pratique sur des données simulées.

1.4 Estimateurs de l'indice de régularité utilisant des estimateurs de la fonction de répartition

Nous avons pu remarquer que l'indice de régularité intervient dans différents problèmes d'estimation non paramétrique. Cependant, il est difficile à estimer et le seul estimateur disponible à notre connaissance est celui de Beirlant, Berlinet et Biau [3] qui utilise l'estimateur des k_n-plus proches voisins de la densité. Ils définissent leur estimateur $\bar{\alpha}_{n,x}$, quelque soit $\tau > 1$, par

$$\bar{\alpha}_{n,x} = \frac{d}{\log \tau} \log \frac{f_{\lfloor \tau^2 k_n \rfloor}(x) - f_{\lfloor \tau k_n \rfloor}(x)}{f_{\lfloor \tau k_n \rfloor}(x) - f_{\lfloor k_n \rfloor}(x)},$$

si $[f_{\lfloor \tau^2 k_n \rfloor}(x) - f_{\lfloor \tau k_n \rfloor}(x)]/[f_{\lfloor \tau k_n \rfloor}(x) - f_{\lfloor k_n \rfloor}(x)] > 1$ et $\bar{\alpha}_{n,x} = 0$ sinon, $\lfloor . \rfloor$ étant la fonction partie entière. Ils obtiennent la convergence en probabilité et la normalité asymptotique de cet estimateur. Cependant, les simulations s'avèrent perfectibles. En nous appuyant sur cet article, nous déterminons dans la troisième partie différents estimateurs de l'indice de régularité en utilisant des estimateurs de la fonction de répartition. Ainsi, nous obtenons un nouvel estimateur convergent, sous certaines hypothèses, de l'indice de régularité

$$\alpha_{n,x} = \frac{d}{\log \tau} \log \frac{\varphi_{n,\tau^2 \delta_n}(x) - \varphi_{n,\tau \delta_n}(x)}{\varphi_{n,\tau \delta_n}(x) - \varphi_{n,\delta_n}(x)},$$

où

$$\varphi_{n,\delta_n}(x) = \frac{\mu_n(B_{\delta_n}(x))}{\lambda(B_{\delta_n}(x))}$$

avec μ_n la mesure empirique. Comme pour l'estimateur de Beirlant, Berlinet et Biau, cet estimateur prendra la valeur 0 si le quotient à l'intérieur du logarithme est inférieur à 1. Lors des simulations, nous utiliserons également un estimateur $\hat{\alpha}_{n,x}$ qui est obtenu en lissant la fonction de répartition empirique

par un estimateur à noyau.

Un estimateur convergent de la fonction de répartition peut donner, sous les conditions adéquates, un estimateur convergent de l'indice de régularité. Nous réalisons par conséquent une large revue bibliographique des estimateurs de la fonction de répartition dans le second chapitre de la deuxième partie. Ce travail a fait l'objet d'un article à paraître dans le *Journal de la Société Française de Statistique*.

Bibliographie

[1] C. Abraham, G. Biau, and B. Cadre. Simple estimation of the mode of a multivariate density. *The Canadian Journal of Statistics*, pages 23–34, 2003.

[2] H. Akaike. An approximation to the density function. *Annals of the Institute of Statistical Mathematics*, 6 :127–132, 1954.

[3] J. Beirlant, A. Berlinet, and G. Biau. Higher order estimation at lebesgue points. *Annals of the Institute of Statistical Mathematics*, 60 :651–677, 2008.

[4] A. Berlinet and S. Levallois. Higher order analysis at lebesgue points. In M. Puri, editor, *G. G. Roussas Festschrift - Asymptotics in Statistics and Probability*, pages 1–16. 2000.

[5] D. Bosq and J.-P. Lecoutre. *Théorie de l'Estimation Fonctionnelle*. Economica, 1987.

[6] L. Devroye. Recursive estimation of the mode of a multivariate density. *The Canadian Journal of Statistics*, 7 :159–167, 1979.

[7] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New-York, 1996.

[8] R. Dudley. *Real Analysis and Probability*. Chapman and Hall, New-York, 1989.

[9] E. Fix and J. J. Hodges. Discriminatory analysis, nonparametric discrimination : consistency properties. *USAF School of Aviation Medicine*, 1951.

[10] D. Loftsgaarden and C. Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, pages 1049–1051, 1965.

[11] D. Moore and J. Yackel. Large sample properties of nearest neighbour density function estimates. In S. Gupta and D. Moore, editors, *Statistical Decision Theory and Related Topics II*. Academic Press, New-York, 1977.

[12] E. Nadaraya. On non-parametric estimates of density functions an regression curves. *Theory of Probability and its Application*, 10 :186–190, 1965.

[13] E. Parzen. On the estimation of a probability density and mode. *The Annals of Mathematical Statistics*, 33 :1065–1076, 1962.

[14] M. Rosenblatt. Remarks on some non-parametric estimates of a density function. *The Annals of Mathematical Statistics*, 27 :832–837, 1956.

[15] W. Rudin. *Real and Complex Analysis*. McGraw-Hill, New York, 1987.

[16] C. van Es. Asymptotics for least squares cross-validation bandwidths in nonsmooth cases. *The Annals of Statistics*, pages 1647–1657, 1992.

Première partie

Normalité asymptotique d'estimateurs de la densité

Normalité asymptotique d'estimateurs de la densité

1.1 Introduction

Nous commençons ce chapitre en rappelant certaines définitions importantes aperçues lors de l'introduction. Nous considérons μ une mesure de probabilité sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, x un point de $\mathbb{R}^d$, δ un réel positif et $B_\delta(x)$ la boule ouverte de centre x et de rayon δ. Si pour x fixé la limite suivante

$$f(x) = \lim_{\delta \to 0} \frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} \tag{1.4}$$

existe, alors x est appelé un *point de Lebesgue* de la mesure μ. Dans ce contexte, Berlinet et Levallois [2] définissent un point ρ-régulier de la mesure μ comme un point de Lebesgue x qui vérifie

$$\left| \frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} - f(x) \right| \leq \rho(\delta), \tag{1.5}$$

où ρ est une fonction mesurable telle que $\lim_{\delta \to 0} \rho(\delta) = 0$.
Cette définition est ensuite utilisée dans l'étude de f_{k_n}, l'estimateur des k_n-plus proches voisins de la densité. En notant $B_n(x) = B(x, R_n(x))$ la plus petite boule fermée de centre x contenant au moins k_n points de l'échantillon, ils démontrent le théorème suivant :

Théorème 1.1.1 ***(Berlinet et Levallois [2])***
Sous les conditions de convergence de f_{k_n}

$$\lim_{n \to \infty} k_n = \infty \quad \text{et} \quad \lim_{n \to \infty} \frac{k_n}{n} = \infty,$$

si x est un point ρ-régulier de la mesure μ avec $f(x) > 0$, alors la condition

$$\sqrt{k_n}\rho(R_n(x)) \xrightarrow{P} 0$$

lorsque n tend vers l'infini implique la convergence en distribution de la variable aléatoire

$$T_n(x) = \sqrt{k_n}\frac{f_{k_n}(x) - f(x)}{f(x)}$$

vers une loi $\mathcal{N}(0,1)$.

Ils obtiennent ensuite comme corollaire de ce résultat que sous les conditions de convergence du Théorème 1.1.1 et si de plus la densité f est lipschitzienne d'ordre $\alpha > 0$ alors la condition

$$\lim_{n\to\infty} \frac{k_n^{1+1/2\alpha}}{n} = 0 \tag{1.6}$$

implique la convergence en distribution de $T_n(x)$ vers une loi $\mathcal{N}(0,1)$. Ils proposent ensuite de considérer la densité f_0 suivante, définie pour $x \in [-0.5, 0.5]$,

$$f_0(x) = 1 - \frac{\sqrt{2}}{3} + \sqrt{|x|}.$$

Pour $0 < t < y < 0.5$ nous avons

$$f_0(t) - f_0(y) = (t-y)\frac{1}{2\sqrt{\epsilon}}$$

où $\epsilon \in (x, y)$ alors que pour $y > 0$ nous avons

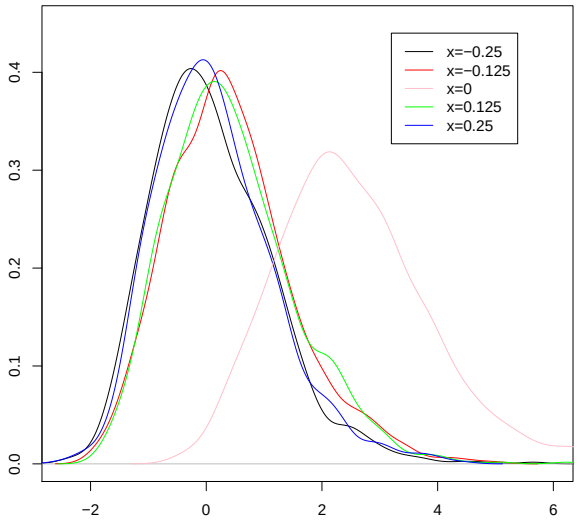
$$f_0(0) - f_0(y) = -\sqrt{y}.$$

La densité f_0 est donc lipschitzienne sur $[-0.5, 0.5]$ d'ordre $1/2$ en 0 et 1 partout ailleurs. Ceci implique donc la ρ-régularité en tout point $x \in [-0.5, 0.5]$ avec

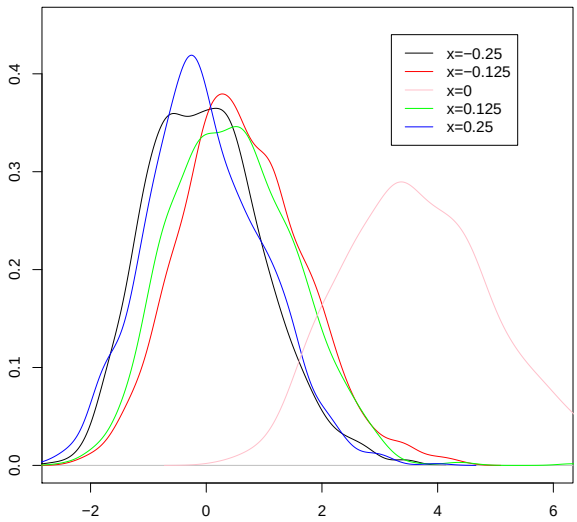
$$\rho(\delta) = \begin{cases} C_x\delta & \text{pour } x \neq 0 \\ C_x\delta^{1/2} & \text{pour } x = 0. \end{cases}$$

Si l'on prend $\alpha = 1$ dans la condition (1.6) nous obtenons alors la condition

$$\lim_{n\to\infty} \frac{k_n^{3/2}}{n} = 0.$$



(a) $n = 1000$



(b) $n = 10000$

FIGURE 1.3 – Estimations de la densité de $T_n(x)$ en différents points x avec $k_n = \sqrt{n}$.

Ainsi, en choisissant $k_n = \sqrt{n}$, nous vérifions la condition précédente et obtenons les estimations de la Figure 1.3, par la méthode du noyau, pour les densités des statistiques $T_n(x)$ estimées en différents points x.
Nous remarquons donc que, pour $x \neq 0$, la distribution de $T_n(x)$ a la forme générale de la loi $\mathcal{N}(0,1)$. Cependant, pour $x = 0$, nous en sommes très éloignés. En effet, comme nous avons $\alpha_0 = 1/2$, la condition (1.6) n'est pas vérifiée en 0 pour $k_n = \sqrt{n}$. En ne tenant pas compte de la spécificité de cette mesure et du changement brutal de l'indice de régularité, nous commettons une importante erreur sur l'estimation de f_0 en 0 due à un choix inadapté pour k_n. Une bonne estimation de l'indice de régularité est donc nécessaire en tout point de définition de la densité afin de déterminer avec précision le nombre de voisins k_n à utiliser et la loi limite de l'estimateur. Nous obtenons dans la suite de cette partie une condition nécessaire et suffisante pour la normalité asymptotique de l'estimateur des k_n-plus proches voisins de la densité et mettons en avant l'importance de l'indice de régularité dans cette loi limite.

1.2 Estimateur des k_n-plus proches voisins

1.2.1 Définition de l'estimateur

Soit $\{k_n, n \geq 1\}$ une suite d'entiers strictement positifs, nous rappelons que l'on définit l'estimateur des k_n-plus proches voisins de la densité par

$$f_{k_n}(x) = \frac{k_n}{n\lambda\left(\overline{B}_{k_n}(x)\right)}$$

où $\overline{B}_{k_n}(x) = B\left(x, R_n(x)\right)$ est la plus petite boule fermée de centre x contenant au moins k_n points de l'échantillon. L'entier k_n jouera donc un rôle crucial à la manière de la fenêtre h_n pour l'estimateur à noyau : si k_n est choisi trop grand l'estimateur sera trop lisse et inversement dans le cas contraire.
En analyse discriminante, Fix et Hodges [5] ont introduit la règle de classification basée sur les k_n-plus proches voisins (voir également Devroye, Györfi et Lugosi [4] sur ce sujet). L'application de cette règle à l'estimation de la densité est due à Loftsgaarden et Quesenberry [7] en tout point où la densité

est positive et continue. Ils démontrent également la convergence en probabilité de leur estimateur, en tout point de Lebesgue x, sous les hypothèses suivantes

$$\lim_{n\to\infty} k_n = \infty\,;\ \lim_{n\to\infty} \frac{k_n}{n} = 0. \tag{1.7}$$

Par la suite, Moore et Yackel [8] ont obtenu le résultat asymptotique suivant

$$\sqrt{k_n}\frac{f_{k_n}(x) - f(x)}{f(x)} \xrightarrow{L} N(0,1)$$

si f est continue et à dérivées bornées dans un voisinage de x avec $f(x) > 0$ et en ajoutant la condition

$$\lim_{n\to\infty} \frac{k_n}{n^{2/3}} = 0$$

aux conditions (1.7). Puis, nous avons vu que Berlinet et Levallois [2] ont utilisé la définition de ρ-régularité pour obtenir le même résultat de normalité asymptotique mais pour une densité f positive dans un voisinage de x, sous les conditions (1.7) et

$$\lim_{n\to\infty} \frac{k_n^{1+1/2\alpha}}{n} = 0,$$

lorsque nous avons $\rho(\delta) = C\delta^{\alpha}$. Comme nous l'avons remarqué dans l'introduction générale de cette thèse il est cependant possible de supposer qu'une relation plus précise que la ρ-régularité a lieu. Ainsi, il est possible de considérer qu'en x point de Lebesgue de la mesure μ nous ayons

$$\boxed{\frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} = f(x) + C_x\delta^{\alpha_x} + o(\delta^{\alpha_x}) \text{ quand } \delta \downarrow 0,} \tag{1.8}$$

où C_x est une constante différente de 0 et α_x un nombre réel strictement positif que nous appelons *indice de régularité*. Ces constantes sont alors uniques et il est clair que cette relation implique la ρ-régularité avec $\rho(\delta) = C_x\delta^{\alpha_x}$. Cet indice joue un rôle primordial dans les résultats qui suivent.

1.2.2 Conditions Nécessaires et Suffisantes de Normalité Asymptotique

Théorème 1.2.1 *Si x est un point de Lebesgue où (1.8) est vérifié avec $f(x) > 0$, et sous les conditions (1.7), la variable aléatoire*

$$T_n(x) = \sqrt{k_n}\frac{f_{k_n}(x) - f(x)}{f(x)}$$

converge en loi si et seulement si la suite

$$\left(\frac{k_n^{1+1/2\alpha_x}}{n}\right)$$

a une limite finie κ. Lorsque cette condition est vérifiée, la loi asymptotique de $T_n(x)$ est

$$\mathcal{N}\left(\frac{C_x\kappa^{\alpha_x}}{2^{\alpha_x}}\left(\frac{1}{f(x)}\right)^{\alpha_x+1}, 1\right).$$

Preuve :
La démonstration de ce théorème découle immédiatement des Lemmes 1.2.3 et 1.2.4 ci-dessous. □

Dans les preuves des lemmes ci-dessous, nous utilisons la décomposition suivante de $T_n(x)$:

$$T_n(x) = a_n(b_n + c_n)$$

où

$$\begin{aligned} a_n &= \frac{k_n}{n\mu(B_n(x))} \\ b_n &= \sqrt{k_n}\frac{1}{f(x)}\left(\frac{\mu(B_n(x))}{\lambda(B_n(x))} - f(x)\right) \\ \text{et } c_n &= \frac{n}{\sqrt{k_n}}\left(\frac{k_n}{n} - \mu(B_n(x))\right) \end{aligned}$$

Cette décomposition a été introduite par Moore et Yackel [8] qui ont prouvé le lemme suivant :

Lemme 1.2.1 *Dans la décomposition précédente de $T_n(x)$, on a, sous les conditions de convergence (1.7),*

$$a_n \xrightarrow{P} 1 \quad et \quad c_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Nous énonçons maintenant les lemmes permettant d'établir le théorème.

Lemme 1.2.2 *Sous les conditions de convergence (1.7) et si (1.8) est vérifée, nous avons*

$$\lim_{n\to\infty} \left(\frac{n}{k_n}\right)^{\alpha_x} \frac{1}{\sqrt{k_n}} b_n = \frac{C_x}{2^{\alpha_x}} \left(\frac{1}{f(x)}\right)^{\alpha_x+1}.$$

Preuve :
De (1.8) il résulte que la fonction

$$\Phi(\delta) = \left(\frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} - f(x)\right) \frac{1}{\delta^{\alpha_x}}$$

est telle que

$$\lim_{\delta\to 0^+} \Phi(\delta) = C_x.$$

En introduisant cette fonction Φ dans l'expression de b_n, on obtient

$$b_n = \Phi(R_n) \left(\frac{2nR_n}{k_n}\right)^{\alpha_x} \frac{1}{2^{\alpha_x} f(x)} \frac{k_n^{\alpha_x+1/2}}{n^{\alpha_x}}$$

et

$$\left(\frac{n}{k_n}\right)^{\alpha_x} \frac{1}{\sqrt{k_n}} b_n = \Phi(R_n) \left(\frac{2nR_n}{k_n}\right)^{\alpha_x} \frac{1}{2^{\alpha_x} f(x)}.$$

Or, sous les conditions (1.7), nous avons

$$R_n \xrightarrow{P} 0 \quad \text{et} \quad \frac{k_n}{2nR_n} = f_{k_n}(x) \xrightarrow{P} f(x).$$

On en déduit que

$$\Phi(R_n) \xrightarrow{P} C_x$$

et la conclusion du lemme en découle. □

Lemme 1.2.3 *(Condition suffisante)*
Sous les hypothèses du Théorème 1.2.1, si

$$\lim_{n\to\infty} \frac{k_n^{1+1/2\alpha_x}}{n} = \kappa$$

alors

$$T_n(x) = \sqrt{k_n}\frac{f_{k_n}(x) - f(x)}{f(x)}$$

est de loi asymptotique $\mathcal{N}\left(\frac{C_x\kappa^{\alpha_x}}{2^{\alpha_x}}\left(\frac{1}{f(x)}\right)^{\alpha_x+1}, 1\right)$.

Preuve :
Les hypothèses et le Lemme 1.2.2 entraînent que

$$\lim_{n\to\infty} b_n = \frac{C_x\kappa^{\alpha_x}}{2^{\alpha_x}}\left(\frac{1}{f(x)}\right)^{\alpha_x+1} \quad \text{en probabilité}$$

et le Lemme 1.2.1 permet de conclure. □

Lemme 1.2.4 *(Condition nécessaire)*
Sous les hypothèses du Théorème 1.2.1, si la suite $\left(\frac{k_n^{1+1/2\alpha_x}}{n}\right)$ *n'est pas convergente alors la variable aléatoire*

$$T_n(x) = \sqrt{k_n}\frac{f_{k_n}(x) - f(x)}{f(x)}$$

ne converge pas en loi.

Preuve :
Si la suite $\left(\frac{k_n^{1+1/2\alpha_x}}{n}\right)$ n'est pas bornée, alors il existe une sous-suite $\left(\frac{k_{q(n)}^{1+1/2\alpha_x}}{q(n)}\right)$ qui tend vers l'infini donc

$$\forall M > 0, P\left(\left|b_{q(n)}\right| > M\right) \underset{n\to\infty}{\to} 1$$

et (b_n), comme par conséquent (T_n), ne converge pas en loi.
Si la suite $\left(\frac{k_n^{1+1/2\alpha_x}}{n}\right)$ est bornée et non convergente alors il est possible d'extraire deux sous-suites convergeant vers des limites différentes l_1 et l_2. Par application du Lemme 1.2.1, les sous-suites correspondantes de (T_n) convergent en loi respectivement vers $\mathcal{N}(l_1, 1)$ et $\mathcal{N}(l_2, 1)$. (Z_n) ne converge donc pas en loi. □

1.3 Nouvelle définition pour l'indice de régularité

Beaucoup d'estimateurs de la densité requièrent un développement pour un rapport de mesures d'ensembles non centrés au point d'estimation x et n'étant pas forcément des boules. La première définition (1.8) est donc inutilisable dans ces cas précis. De plus, il est également intéressant de définir une nouvelle notion qui soit intermédiaire entre la définition de l'indice de régularité et la ρ-régularité. Par exemple, si on considère la densité

$$f_1(x) = \frac{2 - \cos(1/x) + 2x\sin(1/x)}{c}$$

définie sur $\mathbb{R}$ pour $x \in [-1,1]\backslash 0$ avec $c = 4 + 2\sin 1$ et μ_1 sa mesure de probabilité associée.
La fonction de répartition F_1 associée à la densité f_1 est

$$F_1(x) = \begin{cases} 0 & \text{pour } x < -1 \\ 1/2 + (2x + x^2\sin(1/x))/c & \text{pour } -1 \leq x \leq 1 \\ 1 & \text{pour } x > 1. \end{cases}$$

Cette densité est différentiable en tout point de $[-1,1]$ sauf en 0 où elle n'a ni limite à droite ni limite à gauche. On a cependant

$$\lim_{h\to 0} \frac{F_1(h) - F_1(-h)}{2h} = \frac{2}{c}.$$

0 est donc un point de Lebesgue de μ_1 et, en posant $f_1(0) = 2/c$, la discontinuité du second ordre est toujours présente. Maintenant,

$$\frac{\mu_1([-h,h])}{2h} - f_1(0) = \frac{1}{c} h \sin\left(\frac{1}{h}\right)$$

et en tout point de $[-1,1]$ nous avons ρ-régularité avec $\rho(\delta) = \delta/c$ alors qu'en 0 nous n'avons pas d'indice de régularité.
Afin de fixer une nouvelle définition englobant le maximum de cas, nous allons étudier quelques exemples.

1.3.1 Fonctions de répartition C^1

Lemme 1.3.1 *Soit F une fonction de répartition C^1 et admettant $F^{(2)}$ bornée par $C_x > 0$ sur un voisinage I de x dans $\mathbb{R}$. Si nous notons μ la mesure de*

probabilité associée et f sa densité, on a alors, pour tout $\varepsilon > 0$,

$$\left|\frac{\mu(I)}{\lambda(I)} - f(x)\right| \leq \frac{C_x}{2}\lambda(I).$$

Preuve :
En utilisant un développement limité en $x - h_1$ avec $h_1 > 0$ nous avons,

$$F(x - h_1) = F(x) - h_1 F'(x) + \frac{h_1^2}{2}F^{(2)}(c^-)$$

avec $c^- \in [x - h_1, x[$. Avec le même type de développement en $x + h_2$ avec $h_2 > 0$ et avec $I = [x - h_1, x + h_2]$, on obtient

$$\mu(I) = \lambda(I)F'(x) + \frac{1}{2}\left[h_2^2 F^{(2)}(c^+) - h_1^2 F^{(2)}(c^-)\right]$$

avec $c^+ \in]x, x + h_2]$. Finalement, comme $F'(x) = f(x)$ et $F^{(2)}$ bornée par $C_x > 0$, on a

$$\left|\frac{\mu(I)}{\lambda(I)} - f(x)\right| \leq \frac{C_x}{2}\lambda(I).$$

□

Selon les cas, il sera bien entendu possible d'obtenir des inégalités avec des puissances de $\lambda(I)$ plus grandes.

Exemple 1 : la Loi Normale
Nous considérons la loi Normale $N(0, 1)$ définie sur $\mathbb{R}$ avec f_2 sa densité, μ_2 la mesure de probabilité associée et F_2 sa fonction de répartition donnée par

$$F_2(x) = 0.5 * \left(1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right)$$

avec

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}}\sum_{i=0}^{\infty}(-1)^i \frac{x^{2i+1}}{(2i+1)i!}.$$

Nous choisissons h_1 et g_1 deux entiers positifs avec $0 \in I = [-h_1, g_1]$ et $\lambda(I) < \lambda_0$. On a alors

$$\begin{aligned} F_2(g_1) - F_2(-h_1) &= \frac{1}{\sqrt{\pi}}\left(\text{erf}\left(\frac{g_1}{\sqrt{2}}\right) - \text{erf}\left(\frac{-h_1}{\sqrt{2}}\right)\right) \\ &= \frac{1}{\sqrt{\pi}}\sum_{i=0}^{\infty}(-1)^i \frac{g_1^{2i+1} + h_1^{2i+1}}{\sqrt{2}^{2i+1}(2i+1)i!}. \end{aligned}$$

On obtient alors

$$\frac{\mu_2(I)}{\lambda(I)} = \frac{1}{\sqrt{2\pi}} + \frac{1}{\sqrt{\pi}(g_1+h_1)} \sum_{i=1}^{\infty} (-1)^i \frac{g_1^{2i+1} + h_1^{2i+1}}{\left(\sqrt{2}\right)^{2i+1} (2i+1)i!}.$$

Comme $f_2(0) = \frac{1}{\sqrt{2\pi}}$, nous avons

$$\left| \frac{\mu_2(I)}{\lambda(I)} - f_2(0) \right| = \frac{1}{\sqrt{\pi}(g_1+h_1)} \left| \sum_{i=1}^{\infty} (-1)^i \frac{g_1^{2i+1} + h_1^{2i+1}}{\sqrt{2}^{2i+1} (2i+1)i!} \right|.$$

Clairement, le terme général de cette série est de la forme $(-1)^i v_i$ où (v_i) désigne une suite décroissante de nombres positifs convergeants vers 0. C'est donc une série alternée convergente et son premier terme nous donne un majorant en valeur absolue (Lelong-Ferand et Arnaudiès [6]). On a donc

$$\left| \frac{\mu_2(I)}{\lambda(I)} - f_2(0) \right| \leq \frac{1}{\sqrt{\pi}(g_1+h_1)} \frac{g_1^3 + h_1^3}{3\left(\sqrt{2}\right)^3}$$

et finalement, comme $(g_1^3 + h_1^3)/(g_1+h_1)^3 < 1$,

$$\left| \frac{\mu_2(I)}{\lambda(I)} - f_2(0) \right| \leq \frac{(g_1+h_1)^2}{6\sqrt{2\pi}} = \frac{\lambda(I)^2}{6\sqrt{2\pi}}.$$

Exemple 2 : la Loi Cauchy

Nous considérons la loi de Cauchy de paramètre $a = 1$ définie sur $\mathbb{R}$ avec f_3 sa densité, μ_3 la mesure de probabilité associée et F_3 sa fonction de répartition qui vaut

$$F_3(x) = \frac{1}{\pi} \arctan(x) + \frac{1}{2}$$

avec

$$\arctan(x) = \sum_{i=0}^{\infty} \frac{(-1)^i x^{2i+1}}{2i+1}.$$

Nous choisissons h_1 et g_1 deux entiers positifs avec $0 \in I = [-h_1, g_1]$ et $\lambda(I) < \lambda_0$. On a alors

$$F_3(g_1) - F_3(-h_1) = \frac{1}{\sqrt{\pi}} \sum_{i=0}^{\infty} \frac{(-1)^i}{2i+1} \left(g_1^{2i+1} + h_1^{2i+1} \right)$$

ce qui nous donne

$$\frac{\mu_3(I)}{\lambda(I)} = \frac{1}{\pi} + \frac{1}{\sqrt{\pi}\lambda(I)} \sum_{i=1}^{\infty} \frac{(-1)^i}{2i+1} \left(g_1^{2i+1} + h_1^{2i+1} \right).$$

Avec les mêmes arguments que dans l'exemple précédent cette série est une série alternée convergente et son premier terme nous donne un majorant en valeur absolue. On a donc, comme $f_3(0) = 1/\pi$ et en utilisant la même démonstration que précédemment,

$$\left| \frac{\mu_3(I)}{\lambda(I)} - f_3(0) \right| \leq \frac{1}{3\pi} \lambda(I)^2.$$

Exemple 3 : Fonction exponentielle
Nous considérons la densité f_4 qui est définie sur $[-1, 1]$ par

$$f_4(x) = \frac{\exp x}{e - e^{-1}}$$

et sa mesure associée μ_4. Nous choisissons h_1 et g_1 deux entiers positifs avec $0 \in I = [-h_1, g_1]$ et $\lambda(I) < \lambda_0$. On a alors

$$\frac{\mu_4(I)}{\lambda(I)} = \frac{1}{\lambda(I)(e - e^{-1})} \left(g_1 + \frac{g_1^2}{2} + o(g_1^2) - (-h_1 + \frac{h_1^2}{2} + o(h_1^2)) \right),$$

et, avec $f_4(0) = (e - e^{-1})^{-1}$,

$$\left| \frac{\mu_4(I)}{\lambda(I)} - f_4(0) \right| \leq \frac{1}{\lambda(I)(e - e^{-1})} \left(\frac{|g_1^2 - h_1^2|}{2} + \left| o(g_1^2 - h_1^2) \right| \right).$$

En utilisant le fait que $g_1^2 - h_1^2 < \lambda(I)^2$, on obtient finalement, pour tout $\varepsilon > 0$,

$$\left| \frac{\mu_4(I)}{\lambda(I)} - f_4(0) \right| \leq \frac{1 + \varepsilon}{2(e - e^{-1})} \lambda(I).$$

1.3.2 Densité C^0

Densité lipschitzienne d'ordre α

Lemme 1.3.2 *Soit h une densité quelconque lipschitzienne en un point x de $\mathbb{R}$ telle que, pour tout t,*

$$|h(x) - h(t)| \leq C_x \left| x - t \right|^{\alpha},$$

et μ_h sa mesure associée. On a alors, pour tout intervalle I de mesure non nulle contenant x,

$$\left|\frac{\mu_h(I)}{\lambda(I)} - h(x)\right| \leq C_x \lambda(I)^\alpha.$$

Preuve :
En écrivant

$$\mu_h(I) = \int_I h(t)dt = \int_I (h(x) - h(t))\, dt + \lambda(I)h(x)$$

et en utilisant le fait que h soit lipschitzienne en x, on obtient

$$\left|\frac{\mu_h(I)}{\lambda(I)} - h(x)\right| \leq \frac{C_x}{\lambda(I)} \int_I |x-t|^\alpha\, dt$$

et $|x-t| \leq \lambda(I)$ nous donne le résultat final. □

Exemple 4 :
Reprenons la densité

$$f_0(x) = 1 - \frac{\sqrt{2}}{3} + \sqrt{|x|}$$

définie sur $\mathbb{R}$ pour $x \in [-1/2, 1/2]$ et μ_0 sa mesure de probabilité que nous avons déjà étudiée dans l'introduction de cette partie. En $x = 0$, la densité est continue mais non dérivable mais il est possible de démontrer qu'elle est $\frac{1}{2}$-lipschitzienne. Nous avons

$$F_0(x) = \begin{cases} 0 & \text{pour } x < -1/2, \\ 1/2 + \left(1 - \frac{\sqrt{2}}{3}\right) x + \frac{2}{3} x\sqrt{|x|} & \text{pour } -1/2 \leq x \leq 1/2, \\ 1 & \text{pour } x > 1/2. \end{cases}$$

Nous choisissons h_1 et g_1 deux entiers positifs avec $0 \in I = [-h_1, g_1]$ et $\lambda(I) < \lambda_0$. On a alors

$$\begin{aligned} \frac{\mu_0(I)}{\lambda(I)} &= 1 - \frac{\sqrt{2}}{3} + \frac{2}{3}\frac{g_1^{3/2} + h_1^{3/2}}{(g_1 + h_1)} \\ &= 1 - \frac{\sqrt{2}}{3} + \frac{2}{3}\frac{g_1^{3/2} + h_1^{3/2}}{(g_1 + h_1)^{3/2}} \lambda(I)^{1/2} \end{aligned}$$

d'où, comme $f_0(0) = 1 - \frac{\sqrt{2}}{3}$,

$$\left|\frac{\mu_0(I)}{\lambda(I)} - f_0(0)\right| \leq \frac{2}{3}\lambda(I)^{1/2}.$$

Exemple 5 :
Soit la densité

$$f_5(x) = 0.5 \exp(-|x|)$$

définie sur $\mathbb{R}$ pour $x \in [-1, 1]$ et μ_5 sa mesure de probabilité. En $x = 0$, la densité est continue mais non dérivable et elle est 1-lipschitzienne.
En utilisant le développement de la fonction exponentielle sous forme de série et le même type d'argument qu'à l'Exemple 1 sur les séries alternées, on arrive à

$$\left| \frac{\mu_5(I)}{\lambda(I)} - f_5(x) \right| \leq \frac{\lambda(I)}{4}.$$

Nous retrouvons donc les indices annoncés par le Lemme 1.3.2.

1.3.3 Discontinuité du second ordre

Nous pouvons revenir à la densité f_1 pour laquelle nous avons

$$F_1(x) = \begin{cases} 0 & \text{pour } x < -1, \\ 1/2 + (2x + x^2 \sin(1/x))/c & \text{pour } -1 \leq x \leq 1, \\ 1 & \text{pour } x > 1. \end{cases}$$

Nous choisissons h_1 et g_1 deux entiers positifs avec $0 \in I = [-h_1, g_1]$ et $\lambda(I) < \lambda_0$. On a alors

$$\frac{\mu_1(I)}{\lambda(I)} = \frac{2}{c} + \frac{g_1^2 \sin(1/g_1) + h_1^2 \sin(1/h_1)}{c\lambda(I)}$$

et en posant $f_1(0) = 2/c$ (qui est la limite du rapport des mesures de boules), on obtient

$$\left| \frac{\mu_1(I)}{\lambda(I)} - f_1(0) \right| \leq \frac{g_1^2 + h_1^2}{c\lambda(I)} \leq \frac{1}{c}\lambda(I).$$

Ces résultats nous poussent donc à envisager une nouvelle définition pour l'indice de régularité qui permettrait d'englober également les mesures comme la mesure μ_1.

1.3.4 Nouvelle définition

Les exemples ci-dessus nous amènent à considérer la nouvelle définition suivante. Soit μ une mesure de probabilité différentiable en un point de Lebesgue x de $\mathbb{R}$ de densité $f(x)$. On définit l'ensemble E_x par

$$\boxed{E_x = \Big\{\alpha \geq 0 \text{ tel que } \exists C > 0, \exists \lambda_0 > 0, \text{ tels que } \forall I \in \mathcal{I}_x \text{ vérifiant } \lambda(I) < \lambda_0 \text{ on ait } \left|\frac{\mu(I)}{\lambda(I)} - f(x)\right| \leq C\lambda(I)^\alpha \Big\}}$$

où $\mathcal{I}_x$ est l'ensemble des intervalles contenant x.
S'il existe un réel α_x vérifiant

$$\boxed{\alpha_x = \sup E_x} \tag{1.9}$$

alors α_x sera l'indice de régularité de la mesure μ au point x. Si $\sup E_x = +\infty$ alors nous aurons $\alpha_x = +\infty$. Nous donnons ci-dessous un exemple de ce cas particulier.

Soit une densité quelconque d constante sur un intervalle I de mesure non nulle et μ_d sa mesure associée. Alors, pour tout point $x \in I$, il est possible de vérifier que

$$\mu_d(I) = \int_I d(t)dt = d(x).\lambda(I)$$

ce qui nous donne

$$\left|\frac{\mu_d(I)}{\lambda(I)} - d(x)\right| = 0.$$

Nous nous trouvons alors dans le cas où $\alpha_x = +\infty$. L'indice de régularité augmentant avec la régularité de la mesure, il n'est pas étonnant que, dans le cas d'une densité constante sur un intervalle, on trouve $\alpha_x = +\infty$ comme ici.

A l'aide de cette nouvelle définition, nous pouvons définir un indice de régularité dans chacun des exemples vus précédemment notamment pour la mesure μ_1 ce qui n'est pas le cas avec la première définition. Cette nouvelle définition peut donc nous permettre d'élargir certains résultats à des mesures dont le développement ne permet pas d'obtenir un indice de régularité exact.

Elle peut également être utilisée pour des estimateurs de la densité n'utilisant pas des développements de boules centrées en x, le point d'estimation, comme par exemple l'histogramme.

1.4 Application à l'estimateur des k_n-plus proches voisins

En utilisant la nouvelle définition pour l'indice de régularité, nous énonçons un théorème sur la normalité asymptotique de l'estimateur des k_n-plus proches voisins.

Théorème 1.4.1 *Si x est un point de Lebesgue où (1.9) est vérifié avec $f(x) > 0$ et $\alpha_x \in E_x$, et sous les conditions (1.7), la condition supplémentaire*

$$\lim_{n\to\infty} \frac{k_n^{1+1/2\alpha_x}}{n} = 0$$

implique la convergence asymptotique de la variable aléatoire

$$T_n(x) = \sqrt{k_n}\frac{f_{k_n}(x) - f(x)}{f(x)}$$

vers une loi $\mathcal{N}(0,1)$.

Ce théorème nous permet d'élargir nos résultats précédents à des classes de mesures ρ-régulières mais n'admettant pas de développement exact du rapport des mesures de boules, comme par exemple la mesure μ_1.

Preuve :
Il est clair que la condition (1.9) avec $\alpha_x \in E_x$ implique la ρ-régularité en x avec $\rho(\delta) = C_x\delta^{\alpha_x}$. D'après le Théorème 1.1.1, la condition

$$\sqrt{k_n}\rho(R_n(x)) \xrightarrow{P} 0$$

implique la normalité asymptotique de $T_n(x)$. Or, nous avons

$$\sqrt{k_n}\rho(R_n(x)) = \frac{C_x k_n^{1/2+\alpha_x}}{2^{\alpha_x} n^{\alpha_x}} \left(\frac{\lambda(B_n(x))}{\mu(B_n(x))}\right)^{\alpha_x} \left(\frac{n\mu(B_n(x))}{k_n}\right)^{\alpha_x}.$$

Et, comme

$$a_n = \frac{n\mu(B_n(x))}{k_n} \xrightarrow{P} 1$$

et, pour $f(x) > 0$ et $R_n(x) \xrightarrow{P} 0$ (Loftsgaarden et Quesenberry [7]), alors

$$\frac{\lambda(B_n(x))}{\mu(B_n(x))} \xrightarrow{P} f(x)^{-1},$$

ce qui permet de conclure. □

1.5 L'histogramme

1.5.1 Construction de l'estimateur

L'histogramme est un estimateur de la densité très simple à mettre en oeuvre. Pour le définir, nous avons besoin de séparer notre espace de manière régulière et nous supposons donc qu'il existe une suite

$$P_n = \{\Pi_{nq};\ q \in \mathbb{N}\}$$

de partitions équilibrées de $\mathbb{R}$ en boréliens, c'est-à-dire vérifiant

$$\lim_n \sup_q \operatorname{diam}(\pi_{nq}) = 0 \quad \text{et} \quad \limsup(\beta_n/\gamma_n) < \infty$$

avec $\gamma_n = \inf_q \lambda(\Pi_{nq})$ et $\beta_n = \sup_q \lambda(\Pi_{nq})$.
Dans $\mathbb{R}$, on construit une suite d'intervalles

$$\Pi_{nq} = [(q-1)h_n, qh_n[, q \in \mathbb{Z},$$

où h_n est un nombre réel positif dépendant de n. L'histogramme s'écrit alors

$$f_h(x) = \frac{\nu_{nq}}{nh_n}$$

avec $x \in \Pi_{nq}$, ν_{nq} étant le nombre de points dans la $q$ème cellule et $q \in \mathbb{Z}$. La fonction f_h est donc constante sur chacun des éléments de la partition. Afin d'assurer la convergence de f_h vers f, il est donc nécessaire que les partitions deviennent de plus en plus fines. Cet estimateur est alors convergent en moyenne quadratique sous les conditions suivantes dans $\mathbb{R}$ (Bosq et Lecoutre [3])

$$\lim_{n\to\infty} h_n = 0 \quad \text{et} \quad \lim_{n\to\infty} nh_n = +\infty. \tag{1.10}$$

1.5.2 Résultat sur la Normalité Asymptotique

Théorème 1.5.1 *Si x est un point de Lebesgue où (1.9) est vérifié avec $f(x) > 0$ et $\alpha_x \in E_x$, et sous les conditions (1.10), la condition supplémentaire*

$$\lim_{n\to\infty} nh_n^{2\alpha_x+1} = 0$$

implique la convergence asymptotique de la variable aléatoire

$$H_n(x) = \sqrt{nh_n}\frac{f_h(x) - f(x)}{\sqrt{f(x)}}$$

vers une loi $\mathcal{N}(0,1)$.

Preuve :
Nous avons

$$H_n(x) = R_n(x)\sqrt{\frac{\mu(\Pi_{nq})}{h_n}}\frac{1}{\sqrt{f(x)}}$$

avec

$$R_n(x) = \sqrt{nh_n}\,(f_h(x) - f(x))\sqrt{\frac{h_n}{\mu(\Pi_{nq})}}.$$

Comme x est un point de Lebesgue où $f(x) > 0$, le Lemme 1.5.1 nous donne le résultat. □
Nous allons maintenant énoncer les lemmes permettant d'établir le théorème.

Lemme 1.5.1 *Nous avons, sous les hypothèses du Théorème 1.5.1,*

$$R_n(x) \xrightarrow{L} \mathcal{N}(0,1).$$

Preuve :
En posant

$$S_n(x) = \frac{\nu_{nq} - n\mu(\Pi_{nq})}{\sqrt{n\mu(\Pi_{nq})}}$$

et

$$P_n(x) = \sqrt{nh_n}\sqrt{\frac{h_n}{\mu(\Pi_{nq})}}\left(\frac{\mu(\Pi_{nq})}{h_n} - f(x)\right)$$

un développement rapide nous permet de vérifier que

$$R_n(x) = S_n(x) + P_n(x),$$

les Lemmes 1.5.2 et 1.5.3 permettent alors de conclure. □

Lemme 1.5.2 *Sous les hypothèses du Théorème 1.5.1, nous avons*

$$S_n(x) \xrightarrow{L} \mathcal{N}(0,1).$$

Preuve :
La variable aléatoire ν_{nq} suit une loi binomiale de paramètres n et $\mu(\Pi_{nq})$ et donc, par le Théorème Central Limite pour la loi binomiale, on obtient le résultat du lemme. □

Lemme 1.5.3 *Sous les hypothèses du Théorème 1.5.1, nous avons*

$$P_n(x) \to 0.$$

Preuve :
D'après la définition de $P_n(x)$ et la relation (1.9), nous avons, si $\alpha_x \in E_x$ et dès que $h_n < \lambda_0$,

$$|P_n(x)| \leq \frac{\sqrt{n} h_n^{\alpha_x+1} C_x}{\sqrt{\mu(\Pi_{nq})}}$$

et comme x est un point de Lebesgue nous obtenons

$$|P_n(x)| \leq \frac{\sqrt{n} h_n^{\alpha_x+1/2} C_x}{\sqrt{f(x)}}.$$

Or, comme $f(x) > 0$ et $\lim_{n\to\infty} n h_n^{2\alpha_x+1} = 0$, nous obtenons finalement le résultat. □

1.6 Simulations

1.6.1 Indice de régularité

Afin de calibrer le plus précisément possible les différents paramètres des estimateurs de la densité ci-dessus, il est tout d'abord nécessaire d'obtenir une bonne estimation de l'indice de régularité de la mesure μ étudiée. Pour cela, nous utilisons l'estimateur $\bar{\alpha}_{n,x}$ défini par Berlaint, Berlinet et Biau [1] qui utilise l'estimateur des k_n-plus proches voisins de la densité qui est étudié en détails dans la troisième partie de cette thèse et qui est défini quelque soit $\tau > 1$, par

$$\bar{\alpha}_{n,x} = \frac{1}{\log \tau} \log \frac{f_{\lfloor \tau^2 k_n \rfloor}(x) - f_{\lfloor \tau k_n \rfloor}(x)}{f_{\lfloor \tau k_n \rfloor}(x) - f_{\lfloor k_n \rfloor}(x)},$$

si $[f_{\lfloor \tau^2 k_n \rfloor}(x) - f_{\lfloor \tau k_n \rfloor}(x)]/[f_{\lfloor \tau k_n \rfloor}(x) - f_{\lfloor k_n \rfloor}(x)] > 1$ et $\bar{\alpha}_{n,x} = 0$ sinon. Nous ferons également appel aux estimateurs de l'indice de régularité utilisant les estimateurs de la fonction de répartition que nous étudions dans la dernière partie de cette thèse et qui sont définis, quelque soit $\tau > 1$, par

$$\alpha_{n,x} = \frac{1}{\log \tau} \log \frac{\varphi_{n,\tau^2\delta_n}(x) - \varphi_{n,\tau\delta_n}(x)}{\varphi_{n,\tau\delta_n}(x) - \varphi_{n,\delta_n}(x)},$$

où

$$\varphi_{n,\delta_n}(x) = \frac{\mu_n(B_{\delta_n}(x))}{\lambda(B_{\delta_n}(x))}$$

avec μ_n la mesure empirique, et

$$\hat{\alpha}_{n,x} = \frac{1}{\log \tau} \log \frac{\hat{\varphi}_{n,\tau^2\delta_n}(x) - \hat{\varphi}_{n,\tau\delta_n}(x)}{\hat{\varphi}_{n,\tau\delta_n}(x) - \hat{\varphi}_{n,\delta_n}(x)},$$

où

$$\hat{\varphi}_{n,\delta}(x) = \frac{\hat{\mu}_n(B_\delta(x))}{\lambda(B_\delta(x))}$$

avec $\hat{\mu}_n$ la mesure associée à l'estimateur à noyau de la densité. Comme pour l'estimateur de Beirlant, Berlinet et Biau, ces estimateurs prendront la valeur 0 si le quotient à l'intérieur du logarithme est inférieur à 1.

Notons que les conditions pour obtenir un estimateur convergent de l'indice de régularité sur l'estimateur des k_n-plus proches voisins sont

$$\lim_{n\to\infty} \frac{k_n}{n} = 0 \qquad \text{et} \qquad \lim_{n\to\infty} \frac{k_n^{\alpha_x+1/2}}{n^{\alpha_x}} = +\infty.$$

Nous pouvons remarquer que la deuxième condition est en contradiction avec les hypothèses des Théorèmes 1.2.1 et 1.4.1. Nous avons en effet besoin d'un nombre k_n de voisins plus importants lors de l'estimation de l'indice de régularité que lors du chapitre précédent. Cela s'explique par la nature même de cet estimateur de l'indice de régularité. Il joue en effet sur les différences entre plusieurs estimations f_{k_n} avec différents k_n. Plus ces différences seront marquées (avec un k_n "grand"), plus l'estimateur de l'indice de régularité s'approchera de sa valeur réelle.

L'indice de régularité α_x étant inconnu la deuxième condition n'est pas utilisable en l'état. On peut par contre montrer qu'elle est vérifiée si

$$\lim_{n\to\infty} \frac{k_n \log n}{n} = +\infty.$$

Cela nous amène donc à considérer des suites d'entiers k_n telles que, lorsque $n \to \infty$,

$$\frac{n}{\log n} << k_n << n.$$

Pour les estimateurs utilisant les estimations de la fonction de répartition, nous utilisons également différentes valeurs de δ_n vérifiant les conditions de convergence de l'estimateur (voir Proposition 1.2.1 dans la troisième partie) et prenons l'estimation médiane pour l'estimation de l'indice de régularité. Concernant la fenêtre de l'estimateur à noyau, nous choisissons $h_n = S_n n^{-1/5} \approx 0.09 n^{-1/5}$, pour des raisons qui seront détaillées dans la troisième partie.

Nous choisissons de tester nos différents estimateurs sur la densité f_0 définie sur $[-0.5, 0.5]$ par

$$f_0(x) = 1 - \frac{\sqrt{2}}{3} + \sqrt{|x|}$$

et dont l'indice de régularité vaut

$$\begin{cases} 1 & \text{pour } x \neq 0 \\ 1/2 & \text{pour } x = 0. \end{cases}$$

Pour chaque taille d'échantillon nous simulons une dizaine d'échantillons. Afin d'éliminer certaines valeurs extrêmes, nous retiendrons la médiane de la dizaine d'estimations de l'indice de régularité. Nous obtenons alors les résultats rassemblés dans les Tableaux 1.1 et 1.2 ci-dessous.

Même s'ils donnent une approximation relativement correcte, les différents estimateurs ont tendance à surestimer la valeur de l'indice de régularité. Néanmoins, les différentes estimations au point 0 sont inférieures aux autres. De plus, l'estimateur des plus proches voisins réussit à s'en approcher de manière très intéressante. Nous pouvons également noter que les estimations s'améliorent logiquement avec l'augmentation de n. Afin d'obtenir encore

plus de précisions, il serait intéressant d'aller au delà du million de points mais les temps de calcul seraient alors encore plus longs (plusieurs jours).

	Estimateurs ppv			Estimateurs de la f.d.r.	
x	$k_n = \frac{n}{(\log n)^{1/3}}$	$k_n = \frac{n}{\sqrt{\log n}}$	$k_n = \frac{n}{(\log n)^{3/4}}$	Empirique	Noyau
-0.25	1.93	1.99	2.08	1.95	1.92
-0.125	1.63	1.66	1.70	1.91	1.74
0	0.67	0.78	0.87	1.70	1.35
0.125	1.42	1.43	1.57	2.00	1.88
0.25	1.85	1.96	1.9	1.99	1.91

TABLEAU 1.1 – Estimations de α_x pour la densité f_0 pour $n = 500000$.

	Estimateurs ppv			Estimateurs de la f.d.r.	
x	$k_n = \frac{n}{(\log n)^{1/3}}$	$k_n = \frac{n}{\sqrt{\log n}}$	$k_n = \frac{n}{(\log n)^{3/4}}$	Empirique	Noyau
-0.25	1.84	1.86	1.80	2.05	1.90
-0.125	1.34	1.23	1.31	1.88	1.59
0	0.36	0.65	0.71	1.30	1.29
0.125	1.25	1.44	1.32	1.95	1.62
0.25	1.82	1.90	1.91	2.00	1.94

TABLEAU 1.2 – Estimations de α_x pour la densité f_0 pour $n = 1000000$.

L'estimateur des k_n-plus proches voisins s'avère donc meilleur. Par conséquent, nous choisissons la médiane des estimations obtenues pour cet estimateur, pour les différentes valeurs de k_n, comme estimation de l'indice de régularité. Ces valeurs sont contenues dans le Tableau 1.3 et seront celles que nous utilisons dans la section suivante.

	$x = -0.25$	$x = -0.125$	$x = 0$	$x = 0.125$	$x = 0.25$
$n = 500000$	1.99	1.70	0.87	1.57	1.9
$n = 1000000$	1.85	1.34	0.71	1.44	1.9

TABLEAU 1.3 – Estimations de α_x pour la densité f_0.

1.6.2 Estimation de la densité

Nous traitons dans cette partie des jeux de données de taille 1 million avec les estimations de l'indice de régularité du Tableau 1.3. Nous obtenons des intervalles de confiance à 99% pour les estimations de la densité f_0 à l'aide des théorèmes de normalité asymptotique obtenus précédemment. Chaque borne de l'intervalle de confiance sera la médiane des bornes obtenues pour une quinzaine d'échantillons de même taille.

Estimateur des k_n-plus proches voisins

Pour l'estimateur des plus proches voisins, nous avons obtenu que les conditions

$$\lim_{n\to\infty} k_n = \infty \quad , \quad \lim_{n\to\infty} \frac{k_n}{n} = 0 \quad \text{et} \quad \lim_{n\to\infty} \frac{k_n^{1+1/2\alpha_x}}{n} = 0$$

impliquent la convergence de Z_n vers une loi $\mathcal{N}(0,1)$.

Ceci nous permet donc de définir des intervalles de confiance pour l'estimation de la densité f_0. Les estimations de l'indice de régularité situées ci-dessus nous permettent de choisir des suites k_n vérifiant les conditions de convergence. Nous obtenons alors qu'au point 0 il nous faut $k_n << n^{0.63}$ alors qu'aux autres $k_n << n^{3/4}$ suffit. L'approximation donnée par l'estimateur de l'indice de régularité est relativement bonne car les vraies conditions sur les suites k_n sont respectivement $k_n << n^{1/2}$ en 0 et $k_n << n^{2/3}$ aux autres points. Ceci nous donne les résultats du Tableau 1.4 suivant les valeurs de k_n.

	$x=-0.25$	$x=-0.125$	$x=0$	$x=0.125$	$x=0.25$
$k_n = n^{1/3}$	[0.74 ;1.25]	[0.72 ;1.22]	[0.34 ;0.55]	[0.63,1.06]	[0.75 ;1.28]
$k_n = n^{2/5}$	[0.90 ;1.32]	[0.66 ;0.96]	[0.37 ;0.53]	[0.65,0.94]	[0.95 ;1.39]
$k_n = n^{0.64}$	[1.02 ;1.05]	[0.85 ;0.87]	[0.55 ;0.56]	[0.85 ;0.88]	[1.02 ;1.05]
$f_0(x)$	1.03	0.88	0.53	0.88	1.03

TABLEAU 1.4 – Intervalles de confiances pour $f_0(x)$ à partir de l'estimateur des plus proches voisins.

Toutes les estimations des intervalles de confiance contiennent $f_0(x)$ sauf une. Cette erreur est obtenue lorsque la suite k_n ne converge pas de manière

à vérifier les hypothèses de convergence, que ce soit avec la vraie valeur de l'indice de régularité ou avec son estimation. Nous avons donc ici un nouvel exemple de l'importance de l'indice de régularité dans le choix du nombre de voisins k_n.

Histogramme

Pour l'histogramme, les conditions d'obtention de la normalité asymptotique sont

$$\lim_{n\to\infty} h_n = 0 \quad , \quad \lim_{n\to\infty} nh_n = +\infty \quad \text{et} \quad \lim_{n\to\infty} nh_n^{2\alpha_x+1} = 0.$$

En respectant ces conditions, nous obtenons le Tableau 1.5.

	$x = -0.25$	$x = -0.125$	$x = 0$	$x = 0.125$	$x = 0.25$
$h_n = n^{-0.35}$	[1.02 ;1.08]	[0.83 ;0.88]	[0.48 ;0.51]	[0.81 ;0.85]	[0.99 ;1.05]
$h_n = n^{-0.4}$	[1.01 ;1.09]	[0.82 ;0.89]	[0.47 ;0.53]	[0.81 ;0.86]	[0.99 ;1.05]
$h_n = n^{-2/3}$	[0.76 ;1.16]	[0.71 ;1.15]	[0.25 ;0.55]	[0.61 ;1.04]	[0.75 ;1.16]
$f_0(x)$	1.03	0.88	0.53	0.88	1.03

TABLEAU 1.5 – Intervalles de confiances pour $f_0(x)$ à partir de l'histogramme

Pour des valeurs de h_n trop petites, l'estimation perd en précision et l'intervalle de confiance grandit. De la même manière, pour $h_n = n^{-0.35}$ les conditions de convergence ne sont pas vérifiées au point 0 et $f_0(0)$ n'est pas dans notre intervalle de confiance. L'erreur est cependant corrigée pour $h_n = n^{-0.4}$. Pourtant ceci vérifie les conditions de convergence avec l'estimation de l'indice de régularité mais pas avec sa vraie valeur.

Bibliographie

[1] J. Beirlant, A. Berlinet, and G. Biau. Higher order estimation at lebesgue points. *Annals of the Institute of Statistical Mathematics*, 60 :651–677, 2008.

[2] A. Berlinet and S. Levallois. Higher order analysis at lebesgue points. In M. Puri, editor, *G. G. Roussas Festschrift - Asymptotics in Statistics and Probability*, pages 1–16. 2000.

[3] D. Bosq and J.-P. Lecoutre. *Théorie de l'Estimation Fonctionnelle*. Economica, 1987.

[4] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*. Springer-Verlag, New-York, 1998.

[5] E. Fix and J. J. Hodges. Discriminatory analysis, nonparametric discrimination : consistency properties. *USAF School of Aviation Medicine*, 1951.

[6] J. Lelong-Ferrand and J. Arnaudiès. *Cours de Mathématiques - Analyse*, volume 2. Dunod, Paris, 1977.

[7] D. Loftsgaarden and C. Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, pages 1049–1051, 1965.

[8] D. Moore and J. Yackel. Large sample properties of nearest neighbour density function estimates. In S. Gupta and D. Moore, editors, *Statistical Decision Theory and Related Topics II*. Academic Press, New-York, 1977.

Deuxième partie

Estimation du mode pour des densités non continues

Estimation du mode pour des densités non continues

1.1 Introduction

L'estimation du mode d'une densité de probabilité inconnue a suscité récemment une littérature considérable. Ce phénomène est principalement du aux récents progrès en informatique. En effet, l'augmentation continue des puissances de calcul a rendu possible l'expérimentation de nouvelles méthodologies que nous abordons dans la suite de cette partie.

Considérons une mesure de probabilité μ dans $\mathbb{R}^d$ à partir de laquelle nous obtenons une densité f. Nous nous intéressons au problème de l'estimation du mode θ de f à partir d'un échantillon i.i.d. $S_n = \{X_1, \ldots, X_n\}$ distribué selon μ. Formellement, le mode d'une densité est l'argument où la densité est maximisée, c'est-à-dire la valeur pour laquelle on approche le plus (voire on atteint) la borne supérieure essentielle de f.
Soient K un noyau, c'est-à-dire une densité de probabilité, et h_n une fenêtre qui tend vers 0 lorsque n tend vers l'infini. L'estimateur à noyau de la densité (Rosenblatt [15], Parzen [12], Devroye [5]) est alors donné par

$$\hat{f}_n(x) = \frac{1}{nh_n^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \quad x \in \mathbb{R}^d.$$

A partir de cet estimateur $\hat{f}_n$ de f, un estimateur naturel du mode est alors

$$\hat{\theta}_n \in \arg\max_{\mathbb{R}^d} (\hat{f}_n).$$

Cet estimateur du mode fut introduit sur $\mathbb{R}$ par Parzen [12] qui démontra sa convergence en probabilité sous l'hypothèse de la continuité uniforme de

f. Ce résultat a été par la suite étendu à $\mathbb{R}^d$ par Yamato [19]. Nadaraya [11] sur $\mathbb{R}$ puis Van Ryzin [17] sur $\mathbb{R}^d$ établirent par la suite la convergence presque sûre de $\hat{\theta}_n$ toujours sous l'hypothèse de continuité uniforme de f. Parzen [12] a également établi la normalité asymptotique de $\hat{\theta}_n$ sur $\mathbb{R}$ sous la condition que f soit deux fois différentiable. Konakov [8] et Samanta [16] en ont ensuite donné une version multivariée. Eddy ([6] et [7]) a amélioré le résultat de Parzen en obtenant la loi limite de $(\hat{\theta}_n - \theta)$ sous des conditions moins restrictives sur le choix du noyau K mais sous l'hypothèse que f ait une quatrième dérivée bornée et absolument continue. Tous ces estimateurs font intervenir des conditions globales sur la régularité de la densité f. Romano [14] s'est affranchi de ce problème en obtenant la convergence presque sûre et la loi limite de $\hat{\theta}_n$ en supposant que f était continue uniquement dans un voisinage du mode θ. Plus récemment, Vieu [18] a montré, sous les hypothèses supplémentaires que f soit de classe C^k $(k \geq 2)$, que $\hat{\theta}_n$ convergeait vers θ à la vitesse

$$O\left\{\left(\frac{\log n}{n}\right)^{(k-1)/(2k+1)}\right\}.$$

Ce résultat a été récemment amélioré par Leclerc et Pierre-Loti-Viaud [9] qui ont obtenu la vitesse

$$O\left\{\left(\frac{\log\log n}{n}\right)^{(k-1)/(2k+1)}\right\}$$

et conjecturent que c'est la vitesse presque sûre exacte. Mokkadem et Pelletier [10] obtiennent également des résultats intéressants en modifiant les hypothèses sur la fenêtre h_n.

Néanmoins cet estimateur présente certains inconvénients. En effet, il faut tout d'abord calculer $\hat{f}_n$ puis son argument maximum sur l'ensemble du domaine de définition. En pratique, un tel calcul est impossible. Il faut alors choisir une grille de manière plus ou moins arbitraire. Ce substitut est généralement efficace mais souffre de sa dépendance au choix de la grille. Ainsi, une grille choisie trop éparse dans les zones de fortes densités peut conduire à une très mauvaise estimation du mode. De plus, la taille de la grille augmente avec la dimension d de la densité f et peut rendre rapidement les calculs inextricables.

Afin de remédier à ces différents problèmes, Devroye [4] a défini et étudié d'autres estimateurs du mode. Il définit tout d'abord un estimateur récursif à l'aide de $\hat{f}_n$ et obtient sa convergence presque sûre si f est de Lipschitz. Puis il considère l'estimateur θ_n suivant

$$\theta_n \in \arg\max_{S_n}(\hat{f}_n).$$

Comme les points de l'échantillon S_n sont naturellement concentrés dans les zones de forte densité, S_n peut être vu comme la grille naturelle la mieux adaptée à l'estimation du mode. De plus, contrairement à $\hat{\theta}_n$, il ne nécessite qu'un nombre fini d'opérations. Devroye obtient la convergence en probabilité de cet estimateur si f est uniformément continue. Par la suite, Abraham, Biau et Cadre [1] se sont affranchis de cette condition globale en démontrant la convergence presque sûre de θ_n en supposant la continuité de f uniquement autour du mode θ.

Malheureusement, un certain nombre de densités ne vérifie pas cette dernière condition. Il est possible d'en retrouver notamment dans des domaines tels que l'illumination, l'analyse de contours, l'analyse d'images ou la spectrométrie. Nous étudions de telles densités dans la partie Simulations. C'est pourquoi nous définissons et étudions dans ce chapitre un estimateur du mode sans conditions de continuité sur la densité f.

1.2 Convergence

1.2.1 Hypothèses et Notations

Nous munissons $\mathbb{R}^d$ de la norme euclidienne $||.||$. Dans toute la suite nous notons $\text{supp}(K)$ le support compact du noyau K et nous supposons qu'il existe a un réel positif tel que

$$\text{supp}(K) \subset B(0, a)$$

où $B(0, a)$ est la boule fermée de centre 0 et de rayon a. De plus, afin d'utiliser certains résultats de Pollard [13], nous prenons K de forme $M(||.||)$, où M

est une fonction non croissante sur $\mathbb{R}^+$.
Pour tout $\varepsilon > 0$, nous définissons l'ensemble de niveau ε de la fonction f par

$$A(\varepsilon) = \big\{x \in \mathbb{R}^d : f(x) > f(\theta) - \varepsilon\big\}.$$

De plus, nous notons diam $A(\varepsilon)$ le diamètre de $A(\varepsilon)$ défini par

$$\operatorname{diam} A(\varepsilon) = \sup\{||x - y|| : x \in A(\varepsilon), y \in A(\varepsilon)\}.$$

Soit à présent μ une mesure de probabilité sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, nous rappellons que x est un point de Lebesgue de la mesure μ si

$$\lim_{\delta \to 0^+} \frac{\mu(B(x,\delta))}{\lambda(B(x,\delta))} = f(x),$$

existe où λ est la mesure de Lebesgue et $B(x,\delta)$ la boule ouverte de centre x et de rayon δ. Nous avons vu qu'il est alors possible d'aller plus loin que cette définition. Si x est un point de Lebesgue, Berlinet et Levallois [2] le définissent comme un point ρ-régulier de la mesure μ s'il satisfait

$$\Big|\frac{\mu(B(x,\delta))}{\lambda(B(x,\delta))} - f(x)\Big| \leq \rho(\delta), \tag{1.11}$$

où ρ est une fonction mesurable telle que $\lim_{\delta \to 0} \rho(\delta) = 0$. Nous supposons en outre qu'il existe δ_0 suffisament petit pour assurer $\inf_{B(\theta,\delta_0)}(f) > 0$.
Pour finir, pour toute fonction g et tout ensemble $V \subset \mathbb{R}^d$, nous notons

$$||g||_\infty = \sup_{\mathbb{R}^d} ||g(x)|| \quad \text{et} \quad ||g||_V = \sup_V ||g(x)||.$$

On introduit les Hypothèses :

A1 diam $A(\varepsilon) \to 0$ quand $\varepsilon \to 0$;

A2 $\forall \varepsilon > 0$, $\mathrm{P}\{X \in A(\varepsilon)\} > 0$;

A3 $\exists V_0$ avec $\theta \in V_0$ tel que $||E(\hat{f}_n) - f||_{V_0} \to 0$

1.2.2 Convergence de θ_n

Théorème 1.2.1 *Si la densité f de la mesure de probabilité μ est telle que* **A1**, **A2** *et* **A3** *sont vérifiées, alors, si $h_n \to 0$ et $nh_n^d / \log n \to +\infty$, nous avons $\theta_n \to \theta$ p.s.*

Remarques :

1. Une rapide analyse du Lemme 1 d'Abraham, Biau et Cadre [1] nous montre que, pour toute fonction $g : \mathbb{R}^d \to \mathbb{R}$, nous avons

$$\mathbf{A1} \Leftrightarrow \forall \delta > 0, \sup_{B(\theta,\delta)^c} (g) < g(\theta).$$

2. f est continue sur un voisinage quelconque W de $\theta \Rightarrow$ **A2** et **A3**.
   Il est intéressant de noter que, par exemple sur $\mathbb{R}$, W peut être de type $[\theta, \delta]$ ou $[\delta, \theta]$.

3. Si θ est un point de Lebesgue alors la condition **A2** est vérifiée.

4. Si K est uniforme sur $B(0, a)$, s'il existe une fonction ρ et un voisinage V de θ tel que tout point de V soit un point de Lebesgue ρ-régulier alors la condition **A3** est vérifiée.

Nous avons donc étendu les résultats d'Abraham, Biau et Cadre [1] en affaiblissant l'hypothèse de continuité de f autour de son mode θ. De plus, le problème 1 de la page 94 de Bosq et Lecoutre [3] nous permet d'étendre la Remarque 4 à tous les noyaux de Parzen-Rosenblatt pair et à variation bornée.

1.2.3 Vitesse de convergence de θ_n

Supposons que pour un réel $\kappa > 0$, nous ayons l'inégalité suivante

$$0 < \liminf_{\varepsilon \to 0} \frac{\text{diam } A(\varepsilon)}{\varepsilon^\kappa} \leq \limsup_{\varepsilon \to 0} \frac{\text{diam } A(\varepsilon)}{\varepsilon^\kappa} < \infty. \tag{1.12}$$

Dans ce cas, κ est appelé *l'indice de pic* de f. Cet indice mesure le degré de "lissage" de la densité f. Plus concrètement, plus la densité f sera "pointue"

autour de son mode θ, plus l'indice de pic sera grand. Nous étudions dans la section suivante la relation existante entre l'indice de pic et l'indice de régularité.
Dans cette section, nous supposerons que f admet un indice de pic κ et que

$$\liminf_{\varepsilon \to 0} \frac{\lambda\{A(\varepsilon)\}}{\{\text{diam } A(\varepsilon)\}^d} > 0. \tag{1.13}$$

Sans perdre de généralités, nous pouvons également admettre qu'il existe ε_0 suffisament petit pour que nous ayons la propriété suivante :

P1 Il existe $L > 0$ tel que diam $A(\varepsilon) \leq L\varepsilon^\kappa$ pour $\varepsilon \leq \varepsilon_0$.

Nous pouvons remarquer rapidement que **P1** $\Rightarrow$ **A1**. Nous introduisons deux hypothèses supplémentaires afin de déterminer la vitesse de convergence de notre estimateur. Nous introduisons donc les Hypothèses :

Il existe des nombres réels $c > 0$ et $\beta > 0$ tels que

A4 $\kappa\beta \leq 1$

et

A5 $||E(\hat{f}_n) - f||_{B(\theta,\delta_0)} \leq ch_n^\beta$.

Théorème 1.2.2 *Supposons que (1.12), (1.13),* **A4** *et* **A5** *soient vérifiés avec*

$$h_n = \frac{(\log n)^{\frac{2}{2\beta+d}}}{n^{\frac{1}{2\beta+d}}}.$$

Alors, pour tout $p > 0$, on a

$$P\left[||\theta - \theta_n|| \geq \{a + (16c)^\kappa L\}\frac{(\log n)^{\frac{2}{2\beta+d}}}{n^{\frac{\kappa\beta}{2\beta+d}}}\right] = o\Big(\frac{1}{n^p}\Big),$$

où L est la constante définie dans **P1** .

Nous pouvons donc noter que la vitesse de convergence de θ_n vers θ augmente avec κ.

1.2.4 Lien entre l'indice de pic et l'indice de régularité

Nous admettons qu'une relation plus précise que (1.11) existe en x :

$$\frac{\mu(B(x,\delta))}{\lambda(B(x,\delta))} = f(x) + C_x\delta^{\alpha_x} + o(\delta^{\alpha_x}) \text{ as } \delta \downarrow 0, \tag{1.14}$$

où C_x est une constante différente de 0 et α_x un nombre réel strictement positif. α_x est alors appellé *l'indice de régularité* et contrôle le degré de régularité de la mesure symétrique de μ par rapport à la mesure de Lebesgue λ. En fait, le paramètre κ semble être dans la plupart des cas l'inverse de l'indice de régularité de la mesure associée à la densité inconnue f. Par exemple, dans le cas particulier où $d = 1$, nous avons

Si **A1** *est vérifiée, et si la densité f satisfait*

$$f(x) = f(\theta) + a_p|x-\theta|^p + o(|x-\theta|^p) \text{ quand } |x-\theta| \downarrow 0, \tag{1.15}$$

où $p > 0$ et $a_p < 0$, alors, au point θ, la densité f a un indice de pic κ, la mesure associée a un indice de régularité α_θ et nous obtenons finalement

$$\kappa = \frac{1}{\alpha_\theta} = \frac{1}{p}.$$

Afin de comprendre ça, il faut noter que si la mesure associée à f a un indice de régularité α_θ en θ alors $f(\theta+\delta) - f(\theta)$ est d'ordre δ^{α_θ} ce qui implique que $f(\theta + \varepsilon^{1/\alpha_\theta}) - f(\theta)$ est d'ordre ε. Par conséquent, $D(\varepsilon)$ est d'ordre $\varepsilon^{1/\alpha_\theta}$ et f a un indice de pic $1/\alpha_\theta$ en son mode θ. Donc, un estimateur convergent de α_θ est suffisant pour estimer l'indice de pic κ.

Remarque :

5. Si K est uniforme sur $B(0,a)$, s'il existe un voisinage W de θ tel que, $\forall x \in V$,

$$\frac{\mu(B(x,\delta))}{\lambda(B(x,\delta))} = f(x) + C\delta^{\alpha} + R(x,\delta) \text{ lorsque } \delta \downarrow 0,$$

avec

$$\alpha \leq \frac{1}{\kappa} \text{ et } \sup_{x\in V}|R(x,\delta)| = o(\delta^{\alpha}),$$

alors les conditions **A4** et **A5** sont vérifiées pour les réels

$$c = Ca^{\alpha} + 1 \text{ et } \beta = \alpha.$$

1.3 Preuves

Preuve du Théorème 1.2.1.
Soit $0 < \varepsilon < f(\theta)$, alors

$$\mathrm{P}\left\{f(\theta) - \max_{S_n \cap B(\theta,\delta)} f \geq \varepsilon\right\} \;=\; [1 - \mathrm{P}\,\{X \in A(\varepsilon) \cap B(\theta,\delta)\}]^n,$$

pour plus de détails nous renvoyons le lecteur à Abraham, Biau et Cadre [1] qui ont étudié le cas continu.
A l'aide du lemme de Borel-Cantelli et de **A2**, nous obtenons

$$\forall \delta > 0, \max_{S_n \cap V(\delta)} f \to f(\theta) \text{ lorsque } n \to \infty.$$

Avec ce résultat et la Remarque 1 nous pouvons maintenant prouver le Théorème 1 comme Abraham, Biau et Cadre [1], la condition **A3** remplaçant le recours au lemme de Bochner. □

Preuve de la Remarque 2.
Soit $\varepsilon > 0$, par continuité sur W, il existe $0 < h_0 \leq \delta_0$ tel que pour $h \in \mathbb{R}^d$ with $||h|| \leq h_0$, $f(\theta) - f(\theta + h) < \varepsilon$. Cela implique $W \subset A(\varepsilon)$ et, par conséquent, $\mathrm{P}\,\{X \in A(\varepsilon)\} > \mathrm{P}\,\{X \in W\}$. Nous pouvons voir que le terme de droite de cette inégalité est positif dès que $\lambda\{W\} > 0$, $\inf_W f > 0$ et f continue sur W. □

La condition **A3** se démontre quant à elle par une simple application du lemme de Bochner. □

Preuve de la Remarque 3.
Supposons que **A2** ne soit pas vérifiée, nous avons

$$\exists \varepsilon_1 > 0,\ \mathrm{P}\{X \notin A(\varepsilon_1)\} = 1,$$

qui nous donne

$$\exists \varepsilon_1 > 0,\ \forall x \in \mathbb{R}^d,\ f(x) \leq f(\theta) - \varepsilon_1 \quad \text{p.s.}$$

par la définition de $A(\varepsilon)$. Ceci nous mène à

$$\begin{aligned}\frac{\mu(B(x,\delta))}{\lambda(B(x,\delta))} &= \frac{\int_{\theta-\delta}^{\theta+\delta} f(t)d\lambda(t)}{2\delta} \\ &\leq \frac{\int_{\theta-\delta}^{\theta+\delta}(f(\theta)-\varepsilon_1)d\lambda(t)}{2\delta} \\ &\leq f(\theta)-\varepsilon_1 \quad \text{avec probabilité un.}\end{aligned}$$

Mais comme $0 < \varepsilon_1$,

$$\lim_{\delta\to 0^+} \frac{\mu(B(\theta,\delta))}{\lambda(B(\theta,\delta))} < f(\theta) \quad \text{p.s.}$$

et θ n'est pas un point de Lebesgue. □

Preuve de la Remarque 4.
Pour tout $x \in V$,

$$\begin{aligned}E\hat{f}_n(x) &= \frac{1}{\lambda(B(0,a))}\int_{B(0,a)} f(x-h_n t)dt \\ &= \frac{\mu(B(x,h_n a))}{\lambda(B(x,h_n a))}\end{aligned}$$

donc

$$\left|E\hat{f}_n(x) - f(x)\right|_V \leq \rho(h_n a)$$

puisque x est un point de Lebesgue ρ-régulier. Comme h_n tend vers 0 et que $\lim_{\delta\to 0^+}\rho(\delta) = 0$, alors la condition **A3** est satisfaite. □

Preuve de la Remarque 5.
La condition **A4** est bien vérifiée. D'autre part, pour tout $x \in V$, nous avons

$$\begin{aligned}E\hat{f}_n(x) &= \frac{\mu(B(x,h_n a))}{\lambda(B(x,h_n a))} \\ &= f(x) + C\,(ah_n)^\alpha + R(x,h_n a).\end{aligned}$$

Ceci nous donne, pour n assez grand,

$$\left|E\hat{f}_n(x) - f(x)\right|_V \leq (Ca^\alpha + 1)\,h_n^\alpha$$

ce qui achève la preuve. □

Preuve du Théorème 1.2.2.
Remplacer V par $B(\theta, \delta_0)$ dans la preuve d'Abraham, Biau et Cadre [1] nous donne le résultat avec la même démonstration. □

1.4 Simulations

Nous allons étudier sur un exemple les diverses performances de θ_n. Pour cela, nous choisissons une mesure (et la densité associée) qui vérifie nos hypothèses sans qu'il y ait continuité autour du mode. A savoir que le mode se trouve en un point de Lebesgue non continu.
Soit

$$k = 2/3 + \sin 1$$

et f définie sur $[0, 1]$ par $f(0) = 3/k$ et

$$f(x) = \frac{1}{k}\left(2 - 2\sqrt{x} - \cos\left(\frac{1}{x}\right) + 2x\sin\left(\frac{1}{x}\right)\right) \quad \text{si} \quad x \in [0, 1]$$

dont la représentation graphique est la Figure 1.4.
En 0 nous n'avons pas de limite à droite et une discontinuité du second ordre. Nous montrons cependant par la suite que 0 est un point de Lebesgue de la mesure μ_f associée à la densité f.

1.4.1 Étude de f

Nous avons

$$\int_0^1 f(x)dx = \frac{1}{k}\left[2x - \frac{4}{3}x^{3/2} + x^2\sin\left(\frac{1}{x}\right)\right]_0^1 = 1.$$

Il suffit maintenant d'étudier soigneusement la fonction

$$p(u) = \frac{2\sin(u)}{u} - \cos u$$

pour montrer que la fonction f est toujours positive ou nulle.
Par conséquent, la fonction f est bien une densité.

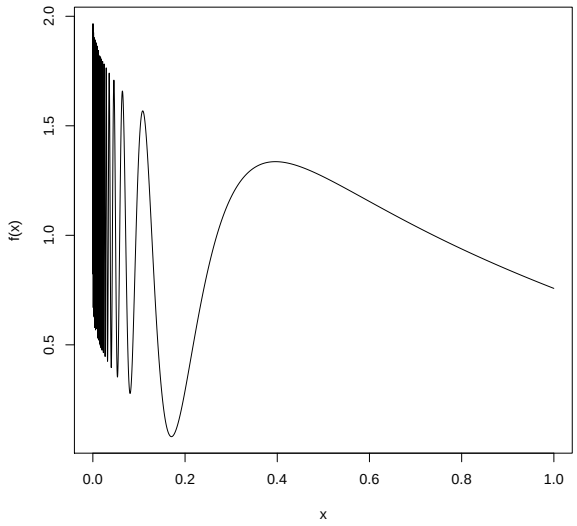


FIGURE 1.4 – Représentation graphique de la densité f.

1.4.2 Calcul du mode

Pour $q \in \mathbb{N}$, soit

$$t_q = \frac{1}{(2q+1)\pi}.$$

La suite (t_q) décroit strictement vers 0 et vérifie

$$\lim_{q\to+\infty} f(t_q) = \lim_{q\to+\infty} \frac{3}{k} - \frac{2}{\sqrt{(2q+1)\pi}k} = \frac{3}{k}.$$

Comme f est continue sur $]0,1]$, on en déduit que la borne supérieure essentielle de f est supérieure ou égale à $3/k$.
Pour montrer que la borne supérieure essentielle est $3/k$ il suffit maintenant de montrer que $3/k$ est un majorant de f. Or, nous avons

$$f(x) \leq \frac{2 - 2\sqrt{x} + 1 + 2x}{k}$$

en majorant simplement le cosinus et le sinus par 1 et

$$\frac{3 - 2\sqrt{x} + 2x}{k} \leq \frac{3}{k}$$

pour tout $x \in [0,1]$.
Par conséquent, $3/k$ est bien la borne supérieure essentielle de la densité f et elle est atteinte en 0.

1.4.3 Étude du point $x = 0$

Nous avons

$$\frac{\mu_f[0,\delta]}{\lambda[0,\delta]} = \frac{2}{k} - \frac{4}{3k}\delta^{1/2} + \frac{\delta}{k}\sin\left(\frac{1}{x}\right)$$

et 0 est donc bien un point de Lebesgue de la mesure μ_f avec $2/k$ comme valeur pour la dérivée à droite de la mesure. De plus, la relation (1.14) est vérifiée avec

$$\alpha_0 = 1/2 \quad \text{et} \quad C_0 = -\frac{4}{3k}.$$

1.4.4 Vérification des hypothèses

Il est possible de démontrer qu'en prenant ε_0 suffisamment petit on a, pour $0 < \varepsilon < \varepsilon_0$,

$$\operatorname{diam} A(\varepsilon) < \operatorname{diam} B(\varepsilon) = \operatorname{diam}\{x \in [0,1] : x < k^2\varepsilon^2\} < k^2\varepsilon^2.$$

Ceci nous permet de vérifier immédiatement la propriété **P1** avec $L = k^2$ et $\kappa = 2$. Nous pouvons ainsi remarquer que nous obtenons bien

$$\alpha_0 = \kappa^{-1} = 1/2.$$

Les hypothèses (1.12), (1.13), **A4** et **A5** sont par conséquent également vérifiées à l'aide des différentes remarques.

1.4.5 Résultats

Nous simulons des échantillons de taille n tirés à partir de la densité f et étudions les résultats obtenus pour notre estimateur θ_n défini plus haut. Cette densité a été choisie de telle sorte que l'extremum local atteint par f en $x \approx 0.4$ perturbe notre estimateur du mode. En effet, si notre estimation n'est pas assez précise, cette caractéristique de f a toutes les chances d'induire θ_n en erreur, c'est pourquoi nous choisissons $h_n = n^{-1/2}$ afin que $\hat{f}_n$

ne soit pas trop lissée. De plus, le fait que le mode se trouve sur un bord du domaine de définition rajoute une difficulté supplémentaire, les effets de bords dus à l'estimateur à noyau étant bien connus.

Pour chaque n, nous simulons une cinquantaine d'échantillons et calculons à chaque fois θ_n. La Figure 1.5 représente la médiane de ces θ_n et nous donnons, à la Figure 1.6, certains box-plots pour des valeurs de n.

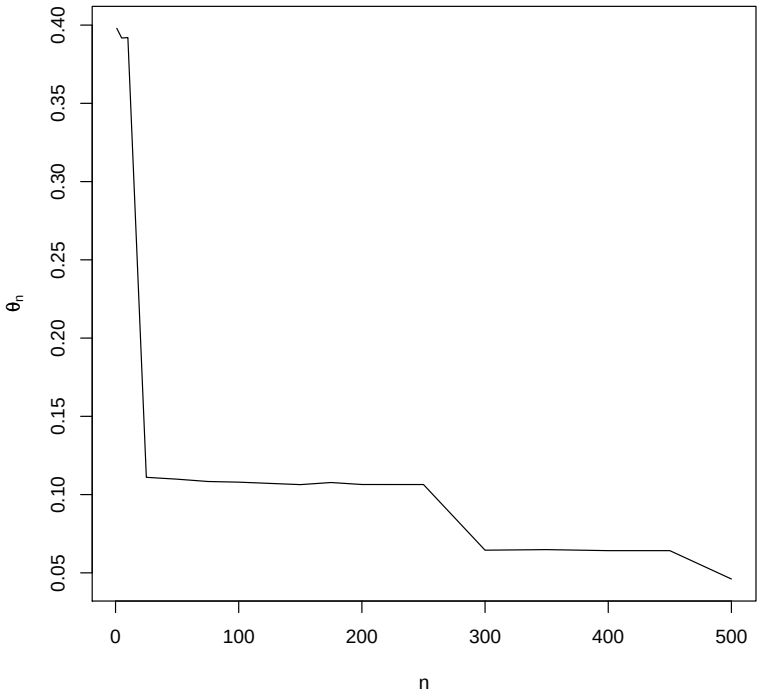


FIGURE 1.5 – θ_n en fonction de n pour la densité f avec $h_n = n^{-0.5}$.

Notons que notre estimateur θ_n converge vers $\theta = 0$. Cependant cette convergence est relativement lente et il faut $n = 500k$ points pour obtenir une médiane des θ_n valant environ 0.04. Ceci s'explique par la forme de la densité et par ses multiples oscillations lorsque x est proche de 0. Toutefois, notre estimateur converge clairement et nous pouvons supposer qu'en augmentant le nombre de points, l'estimation θ_n ne ferait que se rapprocher de θ.

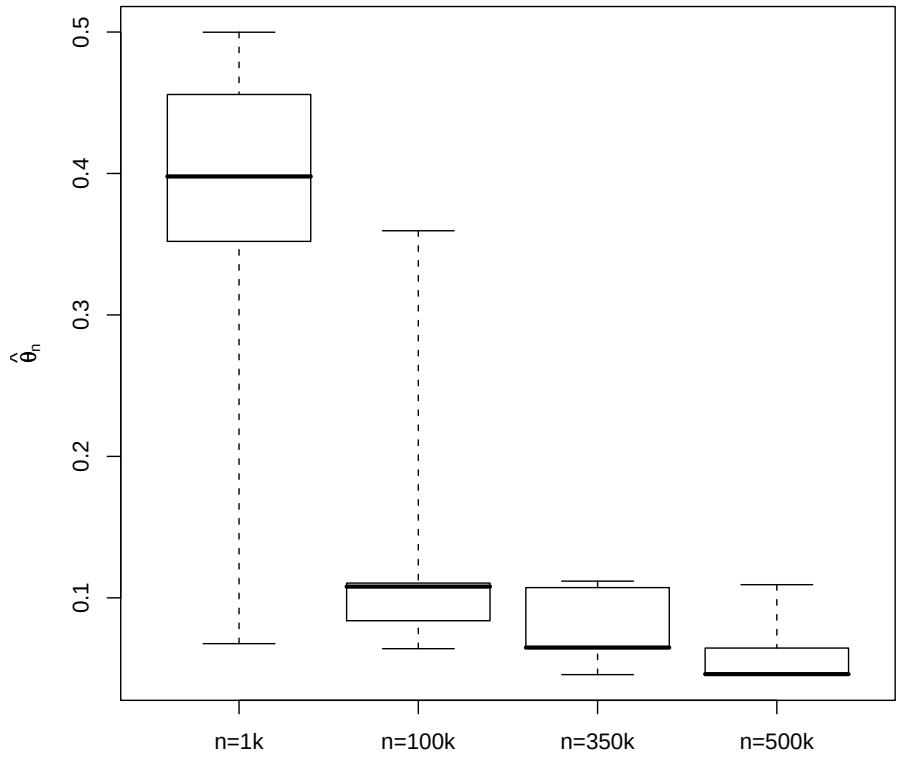


FIGURE 1.6 – Box-plots basés sur une cinquantaine d'estimations de θ_n en fonction de n pour la densité f avec $h_n = n^{-0.5}$.

Toutes les hypothèses étant vérifiées, il est également intéressant d'étudier l'application du Théorème 1.2.2 à cet exemple. Nous avons ici

$$\begin{aligned} a &= 1 \\ c &= \frac{-4+3k}{3k} \\ L &= k^2 \\ \kappa &= 2 \\ \beta &= 1/2. \end{aligned}$$

En prenant

$$h_n = \frac{(\log n)^{\frac{2}{2\beta+d}}}{n^{\frac{1}{2\beta+d}}}$$

comme dans le Théorème 1.2.2, l'estimateur à noyau est bien trop lisse et on obtient $\theta_n \approx 0.4$ même avec 500 000 points dans l'échantillon. Afin d'obtenir des intervalles de confiance asymptotiques exploitables pour cet exemple plus de travail serait donc nécessaire.

Bibliographie

[1] C. Abraham, G. Biau, and B. Cadre. Simple estimation of the mode of a multivariate density. *The Canadian Journal of Statistics*, pages 23–34, 2003.

[2] A. Berlinet and S. Levallois. Higher order analysis at lebesgue points. In M. Puri, editor, *G. G. Roussas Festschrift - Asymptotics in Statistics and Probability*, pages 1–16. 2000.

[3] D. Bosq and J.-P. Lecoutre. *Théorie de l'Estimation Fonctionnelle*. Economica, 1987.

[4] L. Devroye. Recursive estimation of the mode of a multivariate density. *The Canadian Journal of Statistics*, 7 :159–167, 1979.

[5] L. Devroye. *A Course in Density Estimation*. Birkhäuser, Boston, 1987.

[6] W. Eddy. Optimum kernel estimates of the mode. *The Annals of Statistics*, 8 :870–882, 1980.

[7] W. Eddy. The asymptotic distributions of kernel estimators of the mode. *Probability Theory and Related Fields*, 59 :279–290, 1982.

[8] V. Konakov. On asymptotic normality of the sample mode of multivariate distributions. *Theory of Probability and its Applications*, 18 :836–842, 1973.

[9] J. Leclerc and D. Pierre-Loti-Viaud. Vitesse de convergence presque sûre de l'estimateur à noyau du mode. *Comptes Rendus de l'Académie de Sciences de Paris*, 331 :637–640, 2000.

[10] A. Mokkadem and M. Pelletier. The law of the iterated logarithm for the multivariate kernel mode estimator. *ESAIM Probability and Statistics*, 7 :1–21, 2003.

[11] E. Nadaraya. On non-parametric estimates of density functions an regression curves. *Theory of Probability and its Application*, 10 :186–190, 1965.

[12] E. Parzen. On the estimation of a probability density and mode. *The Annals of Mathematical Statistics*, 33 :1065–1076, 1962.

[13] D. Pollard. *Convergence of Stochastic Processes*. Springer, New-York, 1984.

[14] J. Romano. On weak convergence and optimality of kernel density estimates of the mode. *The Annals of Statistics*, 16 :629–647, 1988.

[15] M. Rosenblatt. Remarks on some non-parametric estimates of a density function. *The Annals of Mathematical Statistics*, 27 :832–837, 1956.

[16] M. Samanta. Non-parametric estimation of the mode of a multivariate density. *South African Statistical Journal*, 7 :109–117, 1973.

[17] J. Van Ryzin. On strong consistency of density estimates. *The Annals of Mathematical Statistics*, 40 :1765–1772, 1969.

[18] P. Vieu. A note on density mode estimation. *Statistics & Probability Letters*, 26 :297–307, 1996.

[19] H. Yamato. Sequential estimation of a continuous probability density function and mode. *Bulletin on Mathematical Statistics*, 14 :1–12, 1971.

Troisième partie

Estimateurs de l'indice de régularité utilisant des estimateurs de la fonction de répartition

Chapitre 1

Estimateurs de l'indice de régularité

1.1 Introduction

Soient $\mathcal{B}(\mathbb{R}^d)$ le σ-champ borélien de $\mathbb{R}^d$, $d \geq 1$ et μ une mesure de probabilité sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Nous notons λ la mesure de Lebesgue sur $\mathbb{R}^d$ et nous munissons $\mathbb{R}^d$ d'une norme notée $||.||$. Pour x un point quelconque de $\mathbb{R}^d$, nous notons $B_\delta(x)$ la boule ouverte de centre x et de rayon δ un réel strictement positif. Afin d'étudier le comportement local de $\mu(B_\delta(x))$ par rapport à $\lambda(B_\delta(x))$ il est naturel de considérer le quotient de ces deux quantités. Si, pour x fixé, la limite suivante

$$f(x) = \lim_{\delta \to 0} \frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} \tag{1.1}$$

existe, alors x est appelé *point de Lebesgue* de la mesure μ. Si μ est absolument continue par rapport à λ, alors le théorème de Radon-Nykodim nous dit que μ et f coincident λ-presque partout. De plus, dans ce cas, il est possible de sélectionner parmi toutes les densités issues de μ une densité particulière, toujours notée f, satisfaisant (1.1) en tout point où la limite existe. Dans ce contexte, Berlinet and Levallois [3] définissent un point ρ-régulier de la mesure μ comme un point de Lebesgue x de μ vérifiant

$$\left| \frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} - f(x) \right| \leq \rho(\delta), \tag{1.2}$$

où ρ est une fonction mesurable telle que $\lim_{\delta \downarrow 0} \rho(\delta) = 0$.

Nous pouvons aisément remarquer que la fonction ρ dans (1.2) n'est pas unique et dépend notamment de la norme choisie sur $\mathbb{R}^d$. Néanmoins, il est possible de vouloir aller plus loin que cette relation en essayant notamment de caractériser la vitesse de convergence du rapport des mesures de boules vers la valeur de la dérivée de la mesure. Pour cela, nous supposons dans la suite qu'une relation plus précise que (1.2) intervient en x point de Lebesgue, à savoir

$$\frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} = f(x) + C_x\delta^{\alpha_x} + o(\delta^{\alpha_x}) \text{ quand } \delta \downarrow 0, \tag{1.3}$$

où C_x est une constante différente de 0 et α_x un réel strictement positif. L'indice α_x est un *indice de régularité* qui contrôle le degré de régularité de la dérivée symétrique de μ par rapport à λ. L'objectif de ce chapitre est d'obtenir des estimateurs convergents de l'indice α_x à partir d'un échantillon d'observations dans $\mathbb{R}^d$.

Un premier estimateur de l'indice de régularité a été donné par Beirlant, Berlinet et Biau [2]. Ce dernier fait intervenir l'estimateur des k_n-plus proches voisins de la densité que nous notons f_{k_n}. Ils définissent en effet $\bar{\alpha}_{n,x}$ quelque soit $\tau > 1$ par

$$\bar{\alpha}_x = \frac{d}{\log\tau}\log\frac{f_{\lfloor\tau^2 k_n\rfloor}(x) - f_{\lfloor\tau k_n\rfloor}(x)}{f_{\lfloor\tau k_n\rfloor}(x) - f_{\lfloor k_n\rfloor}(x)}, \tag{1.4}$$

si $[f_{\lfloor\tau^2 k_n\rfloor}(x) - f_{\lfloor\tau k_n\rfloor}(x)]/[f_{\lfloor\tau k_n\rfloor}(x) - f_{\lfloor k_n\rfloor}(x)] > 1$ et $\bar{\alpha}_{n,x} = 0$ sinon. Ils obtiennent la convergence en probabilité de cet estimateur ainsi que sa normalité asymptotique. Néanmoins, lors des simulations, il s'avère peu stable et améliorable.

Ils prouvent également le résultat suivant.

Proposition 1.1.1 *Soit $x \in \mathbb{R}^d$ un point de Lebesgue de μ satisfaisant la condition* (1.3)*. Alors, pour tout $\tau > 1$,*

$$\lim_{\delta\to 0}\frac{\varphi_{\tau^2\delta}(x) - \varphi_{\tau\delta}(x)}{\varphi_{\tau\delta}(x) - \varphi_\delta(x)} = \tau^{\alpha_x},$$

où nous notons

$$\varphi_\delta(x) = \frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))}.$$

Une idée naturelle pour déterminer un estimateur de $\mu(B_\delta(x))$ est alors d'utiliser des estimateurs connus de la fonction de répartition. Nous adaptons

donc, dans la suite de ce chapitre, cette proposition avec des estimateurs de la fonction de répartition dont nous avons réalisé une synthèse bibliographique dans Servien [8]. Cette synthèse est ici reproduite dans le deuxième chapitre de cette partie.

1.2 Résultats de convergence

Afin d'estimer l'indice de régularité α_x, nous considérons un échantillon $(X_i)_{1\leq i\leq n}$ de variables i.i.d. définies sur un espace probabiliste $(\Omega, \mathcal{A}, \mathbb{P})$ et tirées à partir d'une mesure de probabilité μ.

1.2.1 L'estimateur empirique

L'estimateur le plus naturel de la fonction de répartition est l'estimateur empirique F_n défini par

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(X_i\leq x)}$$

où

$$I_A = \begin{cases} 1 \text{ si } x \in A \\ 0 \text{ sinon.} \end{cases}$$

Cet estimateur est extrêmement simple à mettre en oeuvre car il ne fait intervenir aucun paramètre extérieur. Il reste cependant discontinu en de nombreux points ce qui peut poser certains problèmes, notamment ici en terme de précision locale de l'estimateur. Pour cette raison, nous testerons également dans la partie Simulations de ce chapitre un estimateur de l'indice de régularité utilisant une version lissée (par un noyau) de l'estimateur empirique.

Nous renvoyons le lecteur au chapitre suivant pour de plus amples informations sur les différentes caractéristiques de ces estimateurs de la fonction de répartition.

Soit μ_n la mesure empirique associée à l'échantillon $(X_i)_{1\leq i\leq n}$. Nous avons alors la proposition suivante.

Proposition 1.2.1 *Soit $x \in \mathbb{R}^d$ un point de Lebesgue de μ qui satisfait la condition* (1.3)*. Alors, sous les conditions*

$$\lim_{n\to\infty} \delta_n = 0 \quad et \quad \lim_{n\to\infty} n\delta_n^{d+2\alpha_x} = +\infty$$

et, pour tout $\tau > 1$, nous avons

$$\lim_{n\to\infty} \frac{\varphi_{n,\tau^2\delta_n}(x) - \varphi_{n,\tau\delta_n}(x)}{\varphi_{n,\tau\delta_n}(x) - \varphi_{n,\delta_n}(x)} = \tau^{\alpha_x} \quad en\ probabilité\ ,$$

où

$$\varphi_{n,\delta_n}(x) = \frac{\mu_n(B_{\delta_n}(x))}{\lambda(B_{\delta_n}(x))}.$$

Beirlant, Berlinet et Biau [2] obtiennent cette forme de convergence avec leur estimateur des k_n-plus proches voisins sous la condition

$$\lim_{n\to\infty} \frac{k_n^{\alpha_x+d/2}}{n^{\alpha_x}} = +\infty$$

sur le nombre de voisins k_n. Ils utilisent l'estimateur de la densité $f_{\lfloor k_n \rfloor}(x)$ en lieu et place de $\varphi_{n,\delta_n}(x)$.
En modifiant les hypothèses sur la vitesse de convergence de δ_n vers 0, il est également possible de prouver la convergence presque complète de cet estimateur.

Proposition 1.2.2 *Soit $x \in \mathbb{R}^d$ un point de Lebesgue de la mesure μ satisfaisant la condition* (1.3)*. Alors, sous les conditions*

$$\lim_{n\to\infty} \delta_n = 0 \quad et \quad \lim_{n\to\infty} \frac{n\delta_n^{2(d+\alpha_x)}}{\log n} = \infty$$

et, pour tout $\tau > 1$, nous avons

$$\lim_{n\to\infty} \frac{\varphi_{n,\tau^2\delta_n}(x) - \varphi_{n,\tau\delta_n}(x)}{\varphi_{n,\tau\delta_n}(x) - \varphi_{n,\delta_n}(x)} = \tau^{\alpha_x} \quad p.co.$$

1.3 Preuves

Afin de prouver la Proposition 1.2.1 nous avons besoin du résultat suivant.

Lemme 1.3.1 *Sous* (1.1) *et pour tout* $\tau > 1$, *nous avons*

$$\left\{ \begin{array}{l} \delta_n \to 0 \\ n\delta_n^{d+2\alpha_x} \to +\infty \end{array} \right. \Rightarrow R_n(x) \xrightarrow{L^2} 1$$

avec

$$R_n(x) = \frac{\varphi_{n,\tau\delta_n}(x) - \varphi_{n,\delta_n}(x)}{\varphi_{\tau\delta_n}(x) - \varphi_{\delta_n}(x)}.$$

De plus, la condition

$$\left\{ \begin{array}{l} \delta_n \to 0 \\ n\delta_n^{d+2\alpha_x} \to +\infty \end{array} \right.$$

est nécessaire si $f(x) > 0$ et si $A_n(x) \xrightarrow{L^2} 0$ ou $B_n(x) \xrightarrow{L^2} 0$ où

$$A_n(x) = \frac{\varphi_{n,\tau\delta_n}(x) - \varphi_{\tau\delta_n}(x)}{\varphi_{\tau\delta_n}(x) - \varphi_{\delta_n}(x)} \text{ et } B_n(x) = \frac{\varphi_{n,\delta_n}(x) - \varphi_{\delta_n}(x)}{\varphi_{\tau\delta_n}(x) - \varphi_{\delta_n}(x)}.$$

Preuve du Lemme 1.3.1.
Tout d'abord, nous avons

$$\mu_n(B_{\delta_n}(x)) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \in B_{\delta_n}(x)\}}$$

et

$$n\mu_n(B_{\delta_n}(x)) \sim \mathcal{B}(n, \mu(B_{\delta_n}(x))),$$

$$\text{d'où} \qquad \left\{ \begin{array}{lcl} E(\mu_n(B_{\delta_n}(x))) & = & \mu(B_{\delta_n}(x)) \\ V(\mu_n(B_{\delta_n}(x))) & = & n^{-1}\mu(B_{\delta_n}(x))(1 - \mu(B_{\delta_n}(x))). \end{array} \right.$$

D'autre part, nous utilisons par la suite la décomposition suivante

$$R_n(x) = 1 + A_n(x) - B_n(x).$$

Or

$$E(B_n(x)^2) = \frac{\mu(B_{\delta_n}(x))}{\lambda(B_{\delta_n}(x))}\left(1 - \mu(B_{\delta_n}(x))\right) \frac{1}{n\lambda(B_{\delta_n}(x))\delta_n^{2\alpha_x}\left(C_x(\tau^{\alpha_x} - 1) + o(1)\right)^2}$$

nous donne $E(B_n(x)^2) \to 0$ sous les conditions du lemme.
En utilisant une démonstration similaire, nous obtenons $E(A_n(x)^2) \to 0$ sous les mêmes conditions ce qui achève la preuve. □

Preuve de la Proposition 1.2.1.
En remplaçant δ_n par $\tau\delta_n$ dans $R_n(x)$, nous obtenons

$$S_n(x) = \frac{\varphi_{n,\tau^2\delta_n}(x) - \varphi_{n,\tau\delta_n}(x)}{\varphi_{\tau^2\delta_n}(x) - \varphi_{\tau\delta_n}(x)} \xrightarrow{L^2} 1$$

sous les hypothèses du Lemme 1.3.1.
De plus, comme nous avons

$$\begin{aligned}\frac{\varphi_{n,\tau^2\delta_n}(x) - \varphi_{n,\tau\delta_n}(x)}{\varphi_{n,\tau\delta_n}(x) - \varphi_{n,\delta_n}(x)} &= \frac{\varphi_{\tau^2\delta_n}(x) - \varphi_{\tau\delta_n}(x)}{\varphi_{\tau\delta_n}(x) - \varphi_{\delta_n}(x)} \frac{\varphi_{n,\tau^2\delta_n}(x) - \varphi_{n,\tau\delta_n}(x)}{\varphi_{\tau^2\delta_n}(x) - \varphi_{\tau\delta_n}(x)} \\ &\quad * \left(\frac{\varphi_{n,\tau\delta_n}(x) - \varphi_{n,\delta_n}(x)}{\varphi_{\tau\delta_n}(x) - \varphi_{\delta_n}(x)} \right)^{-1} \\ &= \frac{\varphi_{\tau^2\delta_n}(x) - \varphi_{\tau\delta_n}(x)}{\varphi_{\tau\delta_n}(x) - \varphi_{\delta_n}(x)} . \frac{S_n(x)}{R_n(x)}\end{aligned}$$

il suffit d'utiliser le Lemme 1.3.1 et la Proposition 1.1.1 pour conclure. □

Preuve de la Proposition 1.2.2.
L'inégalité d'Hoeffding appliquée à une loi Binômiale nous donne, $\forall t > 0$,

$$\mathbb{P}\big(|\mu_n(B_{\delta_n}(x)) - \mu(B_{\delta_n}(x))| \geq t\big) \leq 2\exp\left(-2nt^2\right).$$

En prenant

$$t = \varepsilon\lambda(B_{\delta_n}(x))\,|\varphi_{\tau\delta_n}(x) - \varphi_{\delta_n}(x)| = \varepsilon\lambda(B_{\delta_n}(x))\,|C_x\delta_n^{\alpha_x}(\tau^{\alpha_x} - 1) + o(\delta_n^{\alpha_x})|,$$

nous obtenons, $\forall \varepsilon > 0$,

$$\mathbb{P}\left(|B_n(x)| \geq \varepsilon\right) \leq 2\exp\left(-2n\left[\varepsilon\lambda(B_{\delta_n}(x))(C_x\delta_n^{\alpha_x}(\tau^{\alpha_x} - 1) + o(\delta_n^{\alpha_x}))\right]^2\right).$$

Par le lemme de Borel-Cantelli, nous avons

$$B_n(x) \overset{p.co.}{\to} 0$$

si

$$\sum_{n=1}^{\infty} \exp\left(-2n\left[\varepsilon\lambda(B_{\delta_n}(x))(C_x\delta_n^{\alpha_x}(\tau^{\alpha_x} - 1) + o(\delta_n^{\alpha_x}))\right]^2\right) < \infty.$$

Or ceci est vrai s'il existe $a > 1$ tel que

$$\exp\left(-2n\left[\varepsilon\lambda(B_{\delta_n}(x))(C_x\delta_n^{\alpha_x}(\tau^{\alpha_x} - 1) + o(\delta_n^{\alpha_x}))\right]^2\right) < \frac{1}{n^a}.$$

Ceci revient à

$$-2n\varepsilon C_x^2 \delta_n^{2(d+\alpha_x)}[(\tau^{\alpha_x}-1)+o(\delta_n^{\alpha_x})]^2 + a\log n < 0$$

qui est impliqué par la condition

$$\log n / n\delta_n^{2(d+\alpha_x)} \to 0.$$

Finalement, avec le même développement que dans le Lemme 1.3.1 nous prouvons

$$R_n(x) = \frac{\varphi_{n,\tau\delta_n}(x) - \varphi_{n,\delta_n}(x)}{\varphi_{\tau\delta_n}(x) - \varphi_{\delta_n}(x)} \overset{p.co.}{\to} 1.$$

En remplaçant δ_n par $\tau\delta_n$ dans $R_n(x)$, nous obtenons

$$S_n(x) = \frac{\varphi_{n,\tau^2\delta_n}(x) - \varphi_{n,\tau\delta_n}(x)}{\varphi_{\tau^2\delta_n}(x) - \varphi_{\tau\delta_n}(x)} \overset{p.co.}{\to} 1$$

sous les mêmes hypothèses. De plus, comme nous avons

$$\frac{\varphi_{n,\tau^2\delta_n}(x) - \varphi_{n,\tau\delta_n}(x)}{\varphi_{n,\tau\delta_n}(x) - \varphi_{n,\delta_n}(x)} = \frac{\varphi_{\tau^2\delta_n}(x) - \varphi_{\tau\delta_n}(x)}{\varphi_{\tau\delta_n}(x) - \varphi_{\delta_n}(x)} . \frac{S_n(x)}{R_n(x)},$$

l'utilisation de la Proposition 1.1.1 suffit pour conclure. □

1.4 Simulations

1.4.1 Estimateur des k_n-plus proches voisins

Comme nous avons pu le voir précédemment, Beirlant, Berlinet et Biau [2] ont présenté un estimateur de l'indice de régularité α_x faisant intervenir l'estimateur des k_n-plus proches voisins de la densité à partir d'un échantillon de variables aléatoires réelles $X_1, ..., X_n$ indépendantes et identiquement distribuées dans $\mathbb{R}^d$ et de même loi μ.

A l'aide de cet estimateur ils définissent $\bar{\alpha}_{n,x}$, quelque soit $\tau > 1$, par

$$\bar{\alpha}_{n,x} = \frac{d}{\log\tau} \log \frac{f_{\lfloor\tau^2 k_n\rfloor}(x) - f_{\lfloor\tau k_n\rfloor}(x)}{f_{\lfloor\tau k_n\rfloor}(x) - f_{\lfloor k_n\rfloor}(x)} \tag{1.5}$$

si $[f_{\lfloor \tau^2 k_n \rfloor}(x) - f_{\lfloor \tau k_n \rfloor}(x)]/[f_{\lfloor \tau k_n \rfloor}(x) - f_{\lfloor k_n \rfloor}(x)] > 1$ et $\bar{\alpha}_{n,x} = 0$ sinon. Nous pouvons noter que l'on obtient une estimation effective de l'indice de régularité pour des valeurs de k_n comprises entre 1 et $\lfloor n/\tau^2 \rfloor$. Cet estimateur est convergent en probabilité sous les conditions suivantes

$$\lim_{n\to\infty} k_n = \infty, \ \lim_{n\to\infty} \frac{k_n}{n} = \infty \ \text{ et } \ \lim_{n\to\infty} \frac{k_n^{\alpha_x + d/2}}{n^{\alpha_x}} = \infty.$$

Nous avons remarqué dans l'introduction générale de cette thèse que l'estimation de l'indice de régularité peut permettre d'améliorer l'estimation de la densité. En effet, il intervient dans l'expression d'un estimateur modifié $f_{k_n}^{(a)}$ des k_n-plus proches voisins de la densité qui s'est avéré moins dépendant du choix du nombre de voisins k_n que l'estimateur usuel f_{k_n}.

Nous testons tout d'abord l'estimateur $\bar{\alpha}_{n,x}$ sur une densité de loi Normale $N(0,1)$. En $x = 0$, un rapide calcul permet de montrer que $\alpha_0 = 2$. Nous obtenons alors les graphiques regroupés dans les Figures 1.1 et 1.2 ci-dessous.

L'estimation de l'indice de régularité s'avère donc délicate même si un palier proche de la vraie valeur α_0 peut être observé pour des bonnes valeurs de k_n. Ces simulations mettent en lumière la nécessité d'avoir un échantillon de grande taille. D'autre part, le choix $\tau = \sqrt{2}$ semble être plus judicieux même si l'influence de ce paramètre sur le résultat final n'est pas prépondérante. Il est en effet possible de s'assurer sur des exemples que pour des valeurs de $\tau \leq 2$, les résultats restent sensiblement identiques.

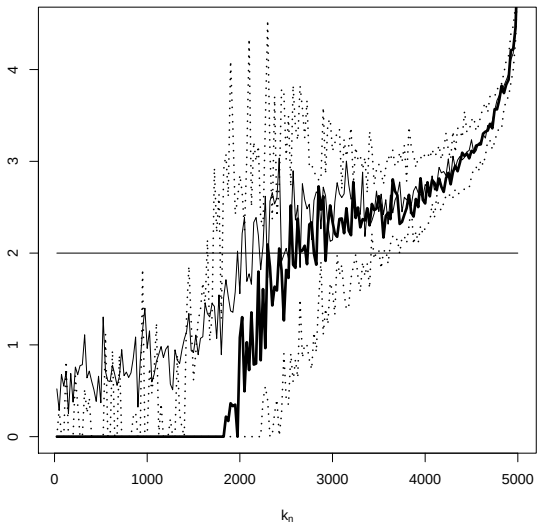
Nous nous intéressons maintenant à la densité f_1 définie pour $t \in \mathbb{R}$ par

$$f_1(t) = \frac{1}{2} \exp(-|t|).$$

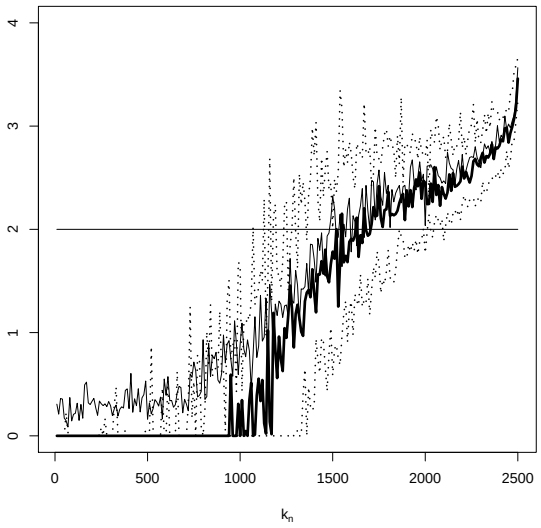
Cette densité est continue mais pas différentiable en 0 et on peut obtenir facilement

$$\frac{\mu(B_\delta(0))}{\lambda(B_\delta(0))} = \frac{1}{2} - \frac{\delta}{4} + o(\delta) \text{ lorsque } \delta \downarrow 0.$$

Par conséquent, nous avons $\alpha_0 = 1$. Nous obtenons alors les graphiques des Figures 1.3 et 1.4 ci-dessous, pour différentes valeurs de n et $\tau = \sqrt{2}$.

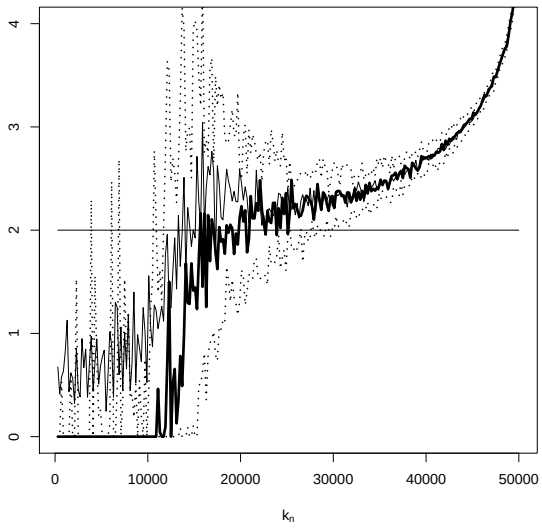


(a) $\tau = \sqrt{2}$

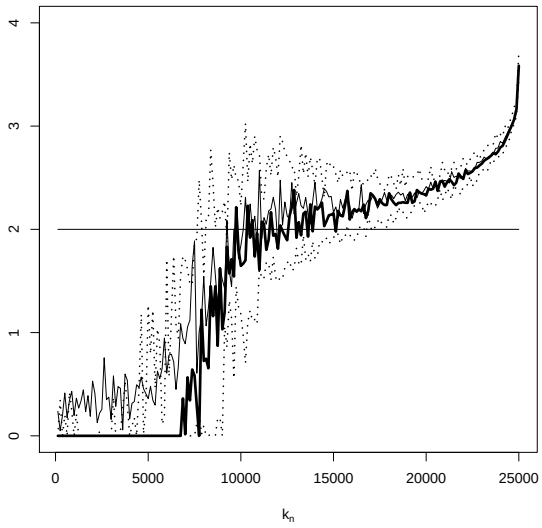


(b) $\tau = 2$

FIGURE 1.1 – Premier quartile (pointillés forts en bas), médiane (en gras), moyenne et troisième quartile (pointillés forts en haut) pour $\bar{\alpha}_{n,0}$ pour 100 échantillons de taille 10000 de loi $N(0,1)$ et différentes valeurs de τ. Le trait plein horizontal est la vraie valeur de α_0.

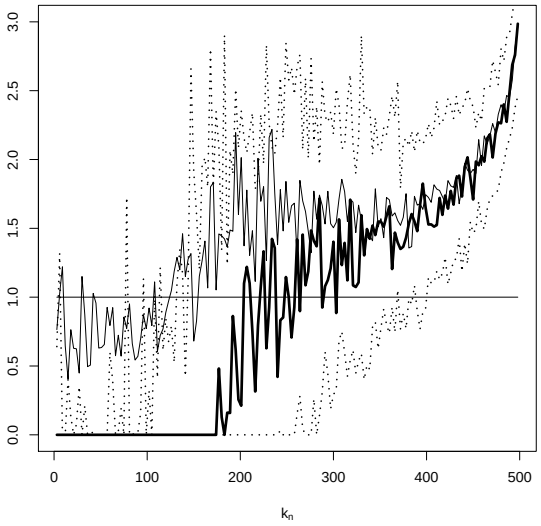


(a) $\tau = \sqrt{2}$

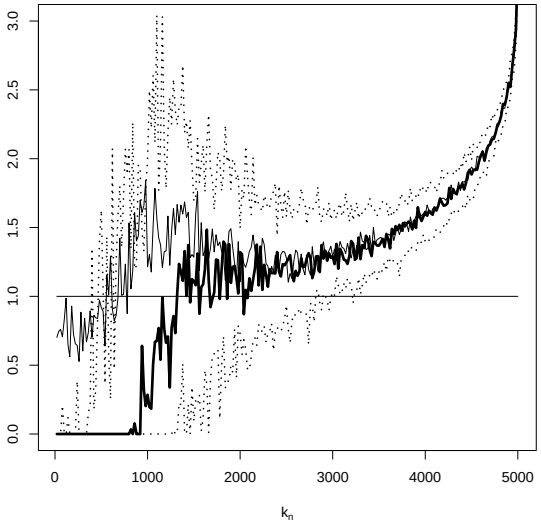


(b) $\tau = 2$

FIGURE 1.2 – Premier quartile (pointillés forts en bas), médiane (en gras), moyenne et troisième quartile (pointillés forts en haut) pour $\bar{\alpha}_{n,0}$ pour 100 échantillons de taille 100000 de loi $N(0,1)$ et différentes valeurs de τ. Le trait plein horizontal est la vraie valeur de α_0.

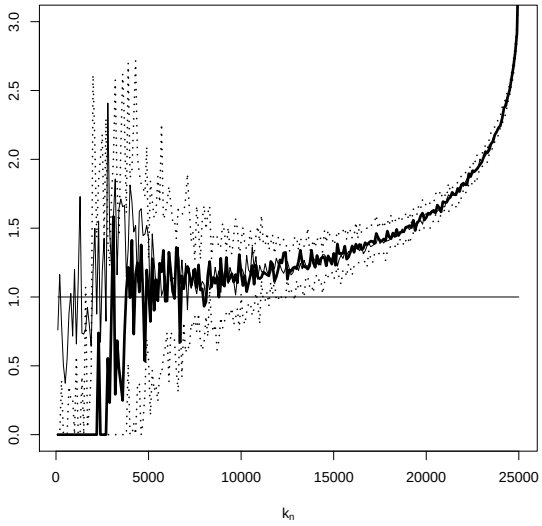


(a) $n = 1000$

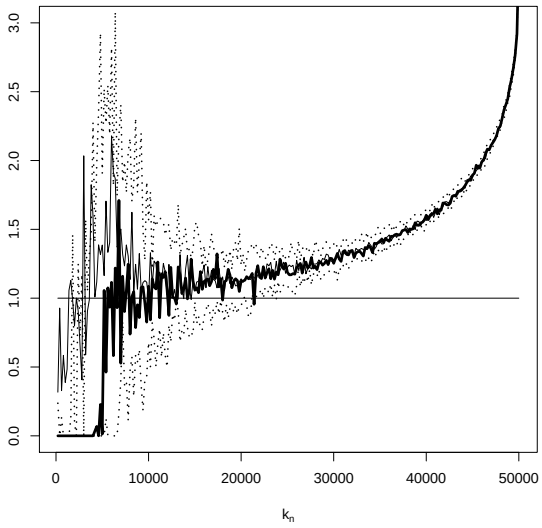


(b) $n = 10000$

FIGURE 1.3 – Premier quartile (pointillés forts en bas), médiane (en gras), moyenne et troisième quartile (pointillés forts en haut) pour $\bar{\alpha}_{n,0}$ pour 100 échantillons de densité f_1 pour différentes valeurs de n. Le trait plein horizontal est la vraie valeur de α_0.



(a) $n = 50000$



(b) $n = 100000$

FIGURE 1.4 – Premier quartile (pointillés forts en bas), médiane (en gras), moyenne et troisième quartile (pointillés forts en haut) pour $\bar{\alpha}_{n,0}$ pour 100 échantillons de densité f_1 pour différentes valeurs de n. Le trait plein horizontal est la vraie valeur de α_0.

Ces figures se rapprochent de celles obtenues précédemment. Elles mettent également en lumière l'importance d'un échantillon de grande taille et permettent d'approcher la valeur de α_0 pour de bonnes valeurs de k_n.

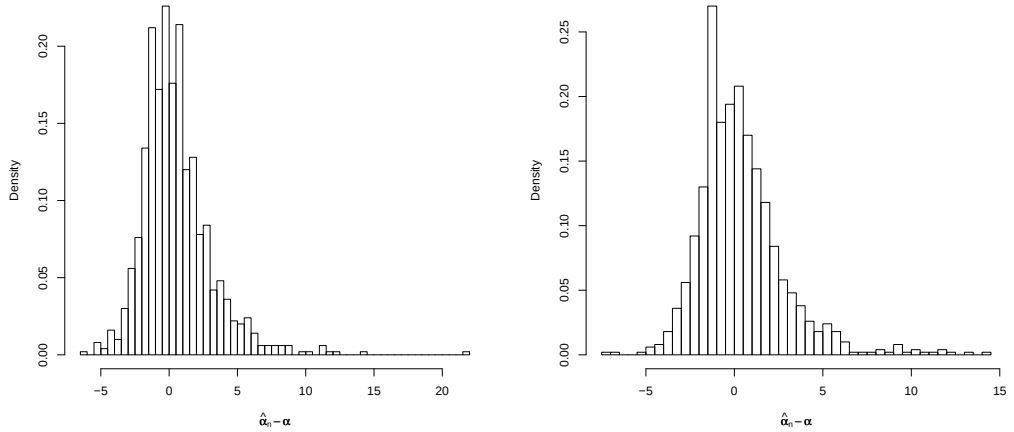
Comme nous l'avons remarqué précédemment, Beirlant, Berlinet et Biau [2] définissent également un nouvel estimateur $f_{k_n}^{(a)}$ faisant intervenir f_{k_n} et l'indice de régularité α_x. Ce nouvel estimateur s'avère moins dépendant du nombre de voisins k_n. Néanmoins, si nous essayons de remplacer f_{k_n} par $f_{k_n}^{(a)}$ dans l'équation (1.5) le nouvel estimateur de l'indice de régularité ainsi obtenu s'avère très peu performant. Ceci est du au fait que $\bar{\alpha}_{n,x}$ utilise les différentes erreurs d'estimation de f_{k_n} selon les valeurs de k_n alors que le terme correctif introduit dans $f_{k_n}^{(a)}$ permet de les corriger.

Dans cet article, ils démontrent également la normalité asymptotique de cet estimateur de l'indice de régularité. Nous pouvons vérifier cette normalité sur les graphiques de la Figure 1.5 qui sont réalisés en $x = 0$ pour la densité f_1. Nous observons bien une forte réduction de la variance pour des grands échantillons renforçant l'idée de la prépondérance de la taille de l'échantillon.

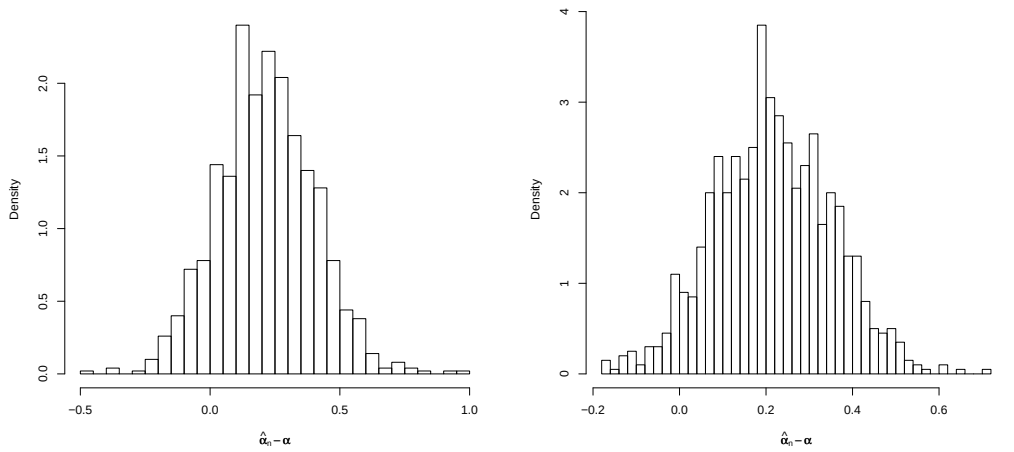
1.4.2 Comparaison des estimateurs de l'indice de régularité

Dans cette partie, nous notons respectivement α_n et $\hat{\alpha}_n$ les estimateurs de l'indice de régularité utilisant respectivement la fonction de répartition empirique et l'estimateur à noyau de la fonction de répartition. Nous prenons pour la fenêtre de l'estimateur à noyau $h_n = S_n n^{-1/5}$ (Bosq et Lecoutre [4]), de légères variations de h_n n'ayant qu'une influence minime sur les résultats. Nous comparons les différentes performances des estimateurs $\bar{\alpha}_n$, α_n et $\hat{\alpha}_n$ pour différentes densités et différentes tailles d'échantillon.

Nous commençons ici aussi par tester nos estimateurs sur la densité de la loi Normale centrée réduite. Nous rappelons qu'en $x = 0$, nous avons $\alpha_0 = 2$.



(a) $n = 1000$ (b) $n = 10000$



(c) $n = 100000$ (d) $n = 200000$

FIGURE 1.5 – Histogrammes de 100 valeurs de $\bar{\alpha}_n - \alpha$ pour $k_n = n/4$ et différentes tailles d'échantillon

Concernant le choix de δ_n, la Propositions 1.2.1 nous donne par exemple

$$n^{\frac{-1}{d+2\alpha_x}} << \delta_n << (\log n)^{\frac{-1}{d+2\alpha_x}}.$$

Il est intéressant de noter que plus α_x est petit, plus δ_n peut l'être. Nous obtenons alors les résultats des Figures 1.6 et 1.7.
Nous remarquons donc que $\hat{\alpha}_n$ produit un très bon estimateur de l'indice

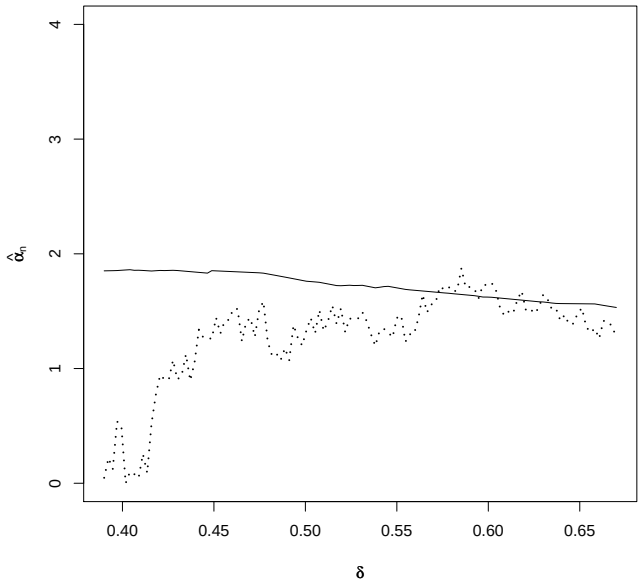
de régularité quelles que soient les valeurs de n. Pour α_n, en revanche, nous avons besoin de plus d'observations mais l'estimation est également très correcte. Il est possible qu'à cause de l'irrégularité de la fonction de répartition empirique, α_n nécessite une plus grande taille d'échantillon que $\hat{\alpha}_n$ pour arriver à une bonne estimation de l'indice de régularité.

Nous étudions maintenant les performances de ces estimateurs pour une mesure moins lisse. Nous choisissons la densité f_1 définie plus haut dans ce chapitre. En $x = 0$, nous avons $\alpha_0 = 1$ et nous obtenons les résultats des Figures 1.8 et 1.9.

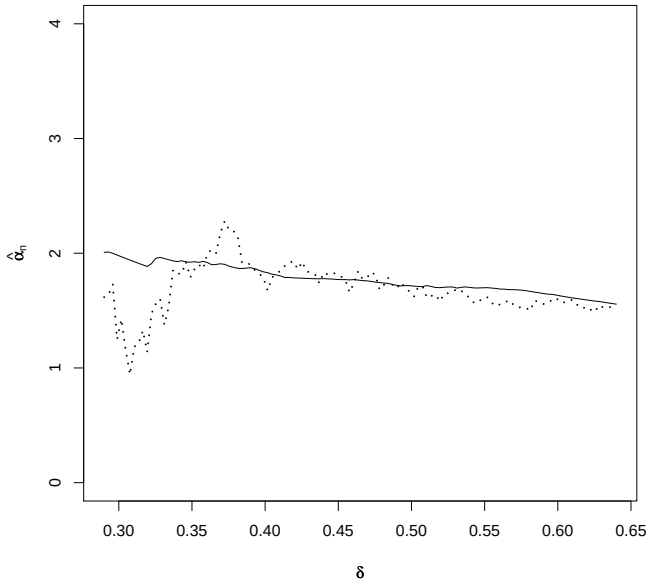
Nous avons une surévaluation de l'indice de régularité pour de petites valeurs de δ_n. Néanmoins, les estimations restent assez correctes bien que nécessitant une grande taille d'échantillon n. Il paraît donc logique, lorsque nous utiliserons (comme dans le chapitre Simulations de la première partie de cette thèse) ces estimateurs sur un seul jeu de données, d'en prendre la médiane obtenue pour différentes valeurs de δ_n.

Cependant, les simulations de la première partie de cette thèse, pour une mesure où l'indice de régularité vaut $1/2$ en $x = 0$ et 1 partout ailleurs, nous donnent des résultats relativement mauvais comparativement à l'estimateur $\bar{\alpha}_{n,x}$. Les estimateurs α_n et $\hat{\alpha}_n$ chutent eux aussi progressivement en se rapprochant de $x = 0$ mais les estimations obtenues restent assez éloignées de la vraie valeur.

Nos estimateurs nous donnent donc de bonnes approximations pour l'indice de régularité mais, au niveau des quelques simulations présentées, ils s'avèrent un peu moins performants que l'estimateur défini par Beirlant, Berlinet et Biau. Ils peuvent cependant constituer une alternative intéressante et nous permettre d'affiner l'estimation de l'indice de régularité. Un éventail de simulations plus large permettrait néanmoins d'affiner la comparaison.

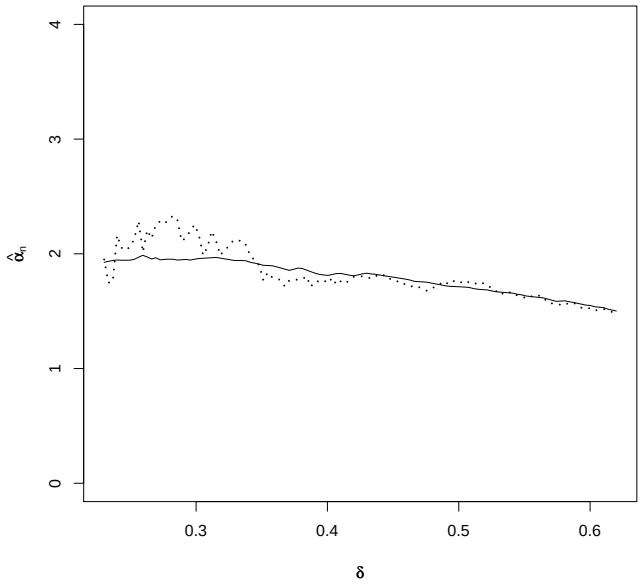


(a) $n = 1000$

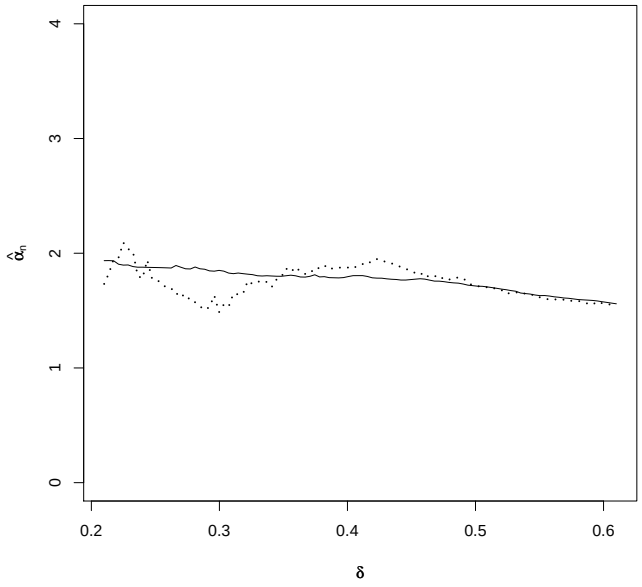


(b) $n = 10000$

FIGURE 1.6 – Médiane des estimateurs α_n (pointillés) et $\hat{\alpha}_n$ (trait plein) en fonction de δ, basée sur 100 simulations d'échantillons de taille n pour la loi Normale centrée réduite.

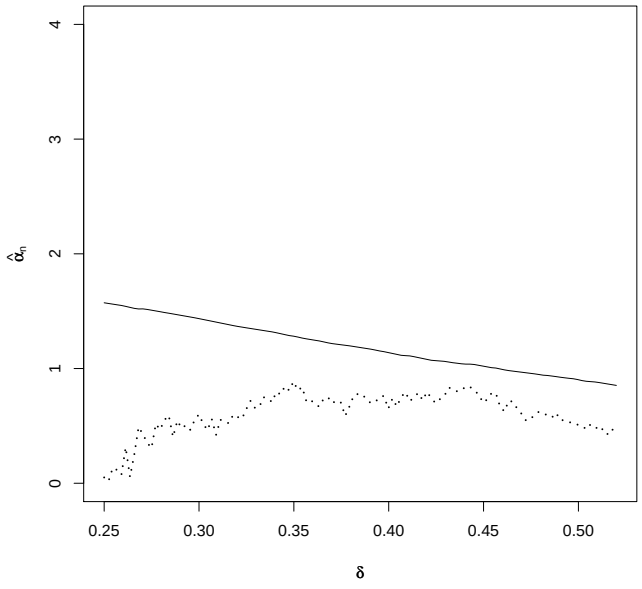


(a) $n = 50000$

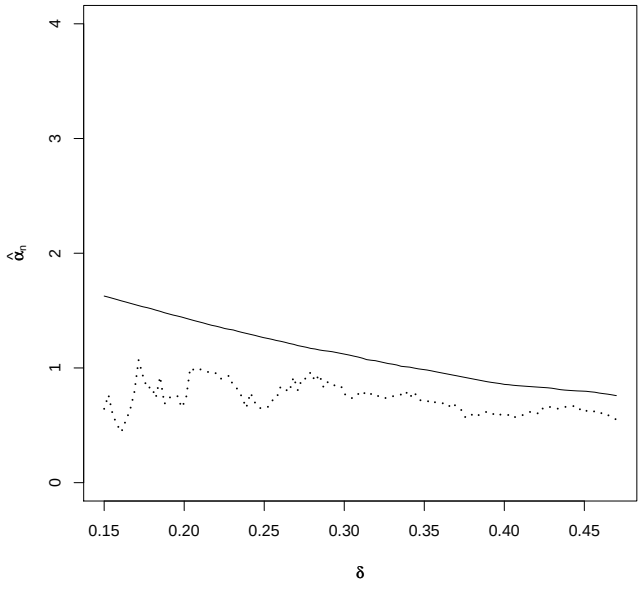


(b) $n = 100000$

FIGURE 1.7 – Médiane des estimateurs α_n (pointillés) et $\hat{\alpha}_n$ (trait plein) en fonction de δ, basée sur 100 simulations d'échantillons de taille n pour la loi Normale centrée réduite.

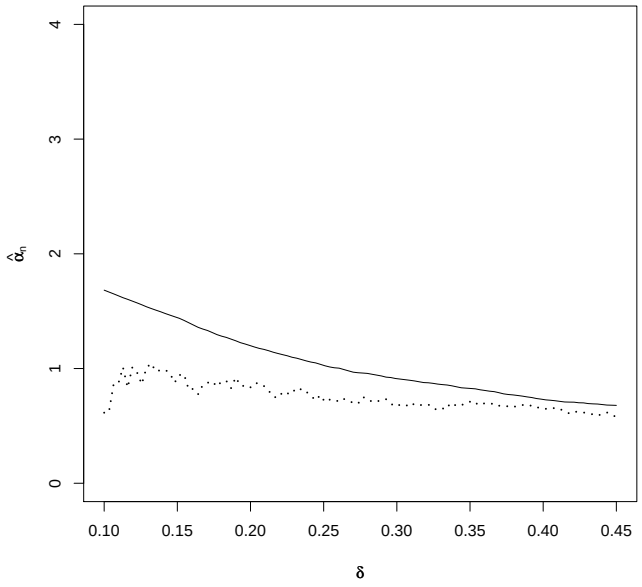


(a) $n = 1000$

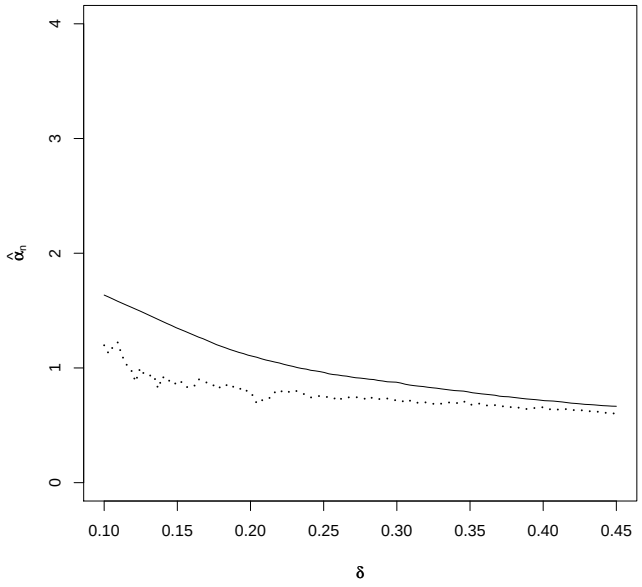


(b) $n = 10000$

FIGURE 1.8 – Médiane des estimateurs α_n (pointillés) et $\hat{\alpha}_n$ (trait plein) en fonction de δ, basée sur 100 simulations d'échantillons de taille n pour la mesure de densité f_1.



(a) $n = 50000$



(b) $n = 100000$

FIGURE 1.9 – Médiane des estimateurs α_n (pointillés) et $\hat{\alpha}_n$ (trait plein) en fonction de δ, basée sur 100 simulations d'échantillons de taille n pour la mesure de densité f_1.

Bibliographie

[1] H. Akaike. An approximation to the density function. *Annals of the Institute of Statistical Mathematics*, 6 :127–132, 1954.

[2] J. Beirlant, A. Berlinet, and G. Biau. Higher order estimation at lebesgue points. *Annals of the Institute of Statistical Mathematics*, 60 :651–677, 2008.

[3] A. Berlinet and S. Levallois. Higher order analysis at lebesgue points. In M. Puri, editor, *G. G. Roussas Festschrift - Asymptotics in Statistics and Probability*, pages 1–16. 2000.

[4] D. Bosq and J.-P. Lecoutre. *Théorie de l'Estimation Fonctionnelle*. Economica, 1987.

[5] G. Hardy and M. Riesz. *The General Theory of Dirichlet's Series*. Cambridge University Press, 1952.

[6] E. Parzen. On the estimation of a probability density and mode. *The Annals of Mathematical Statistics*, 33 :1065–1076, 1962.

[7] M. Rosenblatt. Remarks on some non-parametric estimates of a density function. *The Annals of Mathematical Statistics*, 27 :832–837, 1956.

[8] R. Servien. Estimation de la fonction de répartition : revue bibliographique. *Journal de la Société Française de Statistique*, 150 :84–104, 2009.

Chapitre 2

Estimation de la fonction de répartition : revue bibliographique

Ce chapitre a fait l'objet d'un article accepté dans le *Journal de la Société Française de Statistique*, Volume 150, numéro 2, pp.84-104, 2009.

2.1 Introduction

Un problème récurrent en statistique est celui de l'estimation d'une densité f ou d'une fonction de répartition F à partir d'un échantillon de variables aléatoires réelles $X_1, X_2, \ldots, X_n$ indépendantes et de même loi inconnue. Les fonctions f et F, tout comme la fonction caractéristique, décrivent complètement la loi de probabilité des observations et en connaître une estimation convenable permet de résoudre nombre de problèmes statistiques. Cette estimation tient donc naturellement une place importante dans l'étude de nombreux phénomènes de nature aléatoire. Elle peut être menée, sous des hypothèses restrictives, à l'aide de techniques paramétriques comme la méthode des moments ou celle du maximum de vraisemblance. Les approches non paramétriques que nous privilégions ici sont plus flexibles et constituent toujours un complément utile, même lorsque certains modèles paramétriques semblent s'imposer.

Même si les fonctions de répartition et de densité caractérisent toutes les deux la loi de probabilité d'une variable, la densité a un net avantage sur

le plan visuel. Elle permet d'avoir un aperçu très rapide des principales caractéristiques de la distribution (pics, creux, asymétries, ...), ce qui explique le volume important de littérature qui lui est consacré. La fonction de répartition contient bien sûr cette information mais de manière moins visible. Néanmoins c'est en terme de comportement local de la fonction de répartition que s'explique le plus facilement le comportement des estimateurs fonctionnels (vitesse de convergence, normalité asymptotique) et c'est finalement par un estimateur de la fonction de répartition que l'on passe pour estimer des probabilités d'ensembles : la probabilité qu'une variable se cantonne dans un intervalle donné ou qu'une observation au moins d'un nouvel échantillon dépasse un seuil fixé. Lorsqu'on veut donner une borne inférieure pour la probabilité qu'un paramètre θ inconnu appartienne à un intervalle de la forme $[\theta_n - \varepsilon, \theta_n + \varepsilon]$, où θ_n est un estimateur de θ, on a en fait besoin d'un estimateur de la fonction de répartition de θ_n.

Il est vrai que l'on peut souvent passer d'un estimateur de f à un estimateur de F par intégration et d'un estimateur de F à un estimateur de f par dérivation. Néanmoins une particularité est à souligner : c'est l'existence de la fonction de répartition empirique F_n, fonction de répartition de la loi empirique associée à l'échantillon. Il n'y a bien sûr pas d'équivalent pour la densité, ce qui différencie la nature de chacun des deux problèmes d'estimation. Il est important de noter que la fonction de répartition empirique ne fait appel à aucune structure algébrique ou topologique mais seulement à des notions ensemblistes et que les techniques d'estimation ne sont pas autre chose que des techniques de régularisation de cette référence empirique dont la donnée est équivalente à celle de l'échantillon.

Cette dernière remarque conduit donc à collecter en premier lieu les principaux résultats disponibles sur la fonction de répartition empirique. Ils constituent l'entrée en matière de la présente revue bibliographique. Puis nous passerons dans les sections suivantes à des estimateurs introduits plus spécifiquement pour la fonction de répartition. La section 2.3 sera consacrée au lissage local, dont un exemple bien connu est le lissage polynômial local. La méthode de lissage local décrite permet de donner un cadre général à l'estimation de fonctionnelles de la fonction de répartition et de leurs dérivées. Un cas particulier est la méthode du noyau dont l'application à la fonction

de répartition est abordée dans la section 2.4. Les estimateurs à noyau sont, avec les estimateurs splines décrits en section 2.5, les méthodes de lissage les plus communément utilisées dans les logiciels statistiques. Les sections suivantes abordent des approches plus récentes : les Support Vector Machines (S.V.M.) dans la section 2.6, le level-crossing dans la section 2.7 et les Systèmes de Fonctions Itérées (I.F.S.) dans la section 2.8. Nous mentionnons rapidement en section 2.9 quelques méthodes développées pour la densité et utilisables par intégration, comme celle des fonctions orthogonales, particulièrement celles des ondelettes. La section 2.10 traite du cas de la fonction de répartition d'une loi conditionnelle. Enfin, la prise en compte d'un biais éventuel sur les données est abordée en section 2.11.

2.2 Un estimateur naturel : la fonction de répartition empirique

La fonction de répartition empirique sera notée

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{]-\infty,x]}(X_i)$$

où

$$I_A(x) = \begin{cases} 1 \text{ si } x \in A \\ 0 \text{ sinon.} \end{cases}$$

Afin de passer en revue quelques résultats importants concernant cette fonction, nous notons $Z_n(x) = \sqrt{n}(F_n(x) - F(x))$ et définissons les statistiques suivantes

$$D_n^+ = \sup_{x\in\mathbb{R}} Z_n(x), \qquad D_n^- = \sup_{x\in\mathbb{R}}(-Z_n(x)) \quad \text{et} \quad D_n = \sup_{x\in\mathbb{R}} |Z_n(x)|.$$

Kolmogorov [63] et Smirnov [88] introduisent et étudient ces trois statistiques et démontrent que leur distribution ne dépend pas de F. En outre, à l'aide des théorèmes de Donsker [32] et Doob [33, 34], il est prouvé que les statistiques D_n^+ et D_n^- ont la même loi et que nous avons les résultats asymptotiques suivants :

$$\lim_{n\to\infty} P(D_n^- > \lambda) = \exp(-2\lambda^2),$$

$$\lim_{n\to\infty} P(D_n > \lambda) = 2\sum_{k=1}^{\infty}(-1)^{k+1}\exp(-2k^2\lambda^2).$$

Par la suite, Dvoretzky, Kiefer et Wolfowitz [35] déterminent une borne de la forme

$$P(D_n^- > \lambda) \leq C\exp(-2\lambda^2),$$

où C est une constante indéterminée. L'ensemble de cette démarche et de ces résultats sont analysés dans Hennequin et Tortrat [56]. Pour des propriétés liées aux statistiques d'ordre et de rang on pourra consulter le livre de Caperaa et Van Cutsem [20].
De nombreux auteurs essayent ensuite de trouver la meilleure constante C dans l'inégalité précédente. Devroye et Wise [30], Shorack et Wellner [86] ou Hu [58] font peu à peu diminuer cette constante. Finalement, Massart [68] démontre qu'il est possible de prendre $C = 1$ pourvu que

$$\lim_{n\to+\infty} P(D_n^- > \lambda) = \exp(-2\lambda^2) \leq 1/2.$$

Il montre aussi que, quel que soit λ,

$$P(D_n > \lambda) \leq 2\exp(-2\lambda^2),$$

et que ces bornes ne peuvent plus être améliorées, ces inéquations étant valides que F soit continue ou pas.
D'autre part, le théorème de Glivenko-Cantelli nous donne la convergence uniforme presque sûre de F_n vers F c'est-à-dire

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \to 0 \ \text{ p.s.}$$

Pour des variables dépendantes (plus précisément ϕ-mélangeantes), ce résultat est amélioré plus tard par Collomb, Hassani, Sarda et Vieu [24] en convergence presque complète sous certaines hypothèses, notamment l'uniforme continuité de f.
F_n est également l'estimateur non paramétrique du maximum de vraisemblance (Kiefer et Wolfowitz [62], Efron et Tibshirani [37]) et, en général, une transformée $t(F)$ a pour estimateur du maximum de vraisemblance $t(F_n)$. D'autre part, parmi les estimateurs sans biais de $F(x)$, $F_n(x)$ est également

l'unique estimateur de variance minimale (Yamato [102], Lehmann [66]). Cette dernière vaut par ailleurs $F(x)\,(1-F(x))\,/n$. Yamato [102] élargit les recherches à une famille de fonctions englobant F_n. Il étudie les fonctions du type

$$F_n^*(x) = \frac{1}{n}\sum_{j=1}^{n} W_n(x - X_j)$$

où W_n est une fonction de répartition connue. Il démontre que, en tout point de continuité x de F, $F_n^*(x)$ est asymptotiquement non biaisé et

$$P[\sup_{-\infty<x<\infty} |F_n^*(x) - F(x)| \to 0] = 1$$

si et seulement si $W_n \to e_0$ avec

$$e_0(x) = \begin{cases} 1 \text{ si } x \geq 0 \\ 0 \text{ sinon.} \end{cases}$$

Il obtient de plus la convergence vers une loi normale $\mathcal{N}(0,1)$ de la distribution de

$$\frac{\sqrt{n}[F_n(x) - EF_n(x)]}{\sqrt{F(x)[1-F(x)]}}.$$

Pour évaluer l'écart entre F et son estimateur, plusieurs autres critères peuvent être utilisés. Ainsi, Devroye et Györfi [28] choisissent comme critère la variation totale V qui se définit comme suit entre deux mesures de probabilité μ et ν

$$V(\mu,\nu) = \sup_A |\mu(A) - \nu(A)|,$$

avec le supremum calculé sur tous les boréliens A, et l'entropie relative I telle que

$$I(\mu,\nu) = \sup_{\{A_j\}} \sum_j \mu(A_j) \log \frac{\mu(A_j)}{\nu(A_j)}$$

où le supremum est pris sur l'ensemble des partitions finies en boréliens mesurables $\{A_j\}$ de $\mathbb{R}$. Au sens de ces critères, F_n est un mauvais estimateur. En effet, ils obtiennent $V(\mu,\mu_n) = 1$ et $I(\mu,\mu_n) = \infty$ pour des mesures non-atomiques. Mais ils démontrent ensuite la non-existence de bons estimateurs F_n^*, de mesure associée μ_n^*, tels que

$$V(\mu,\mu_n^*) \to 0 \text{ ou } I(\mu,\mu_n^*) \to 0,$$

qui rend ces critères non adaptés.

Aggarwal [2] introduit le problème consistant à déterminer le meilleur estimateur invariant de F, au sens des transformations monotones, avec la fonction de perte

$$L(F,a) = \int \{F(t) - a(t)\}^2 h(F(t)) dF(t).$$

L'éventuelle admissibilité, c'est-à-dire la minimisation de L, et minimaxité de l'estimateur sont également des questions importantes. Dans le cas particulier de fonction de perte de Cramér-von Mises, où $h(t) = t^{-1}(1-t)^{-1}$, F_n est le meilleur estimateur invariant. Aggarwal [2], Brown [18] et Yu [103] démontrent la non admissibilité de F_n pour $h(t) = t^{\alpha}(1-t)^{\beta}, \alpha, \beta \geq -1$ alors qu'il est admissible pour

$$L(F,a) = \int \{F(t) - a(t)\}^2 F(t)^{\alpha}(1-F(t))^{\beta} dW(t),$$

avec $-1 \leq \alpha, \beta \leq 1$ et W une mesure finie (Cohen et Kuo [23]). Dvoretzky, Kiefer et Wolfowitz [35] et Phadia [76] prouvent que F_n est asymptotiquement minimax pour une grande variété de fonctions de pertes. Néanmoins, pour celle de Kolmogorov-Smirnov, Friedman, Gelman et Phadia [49] déterminent le meilleur estimateur invariant de F. Celui-ci est une fonction en escalier différente de F_n.

D'autres fonctions en escalier ont également été étudiées. Ainsi, si F est continue sur $[0,1]$, nous savons qu'il existe un estimateur $\tilde{F}_n$ linéaire par morceaux et tel que la distribution de $\sqrt{n}[\tilde{F}_n(x) - F(x)]$ converge faiblement vers un pont brownien $B(x)$ connu (Billingsley [14]). Beran [7] s'appuie sur ce résultat pour définir un nouvel estimateur $\hat{F}_n$ tel que

$$\hat{F}_n = \begin{cases} H(x) & \text{si } \sup_x |\tilde{F}_n(x) - H(x)| \leq n^{-1/4} \\ \tilde{F}_n(x) & \text{sinon,} \end{cases}$$

où H est une fonction de répartition quelconque sur $[0,1]$. Il obtient alors que, hormis le cas trivial où la distribution de l'échantillon est H, $\sqrt{n}[\hat{F}_n(x) - F(x)]$ converge vers $B(x)$. De plus, il démontre que tout estimateur $\hat{F}_n$ régulier peut être représenté comme une convolution d'un pont brownien avec une autre fonction de répartition de $C[0,1]$ dépendant uniquement de la densité f.

La fonction de répartition empirique ne tient pas compte d'une éventuelle information que nous pouvons avoir sur la fonction à estimer. Modarres [69] utilise une possible symétrie pour bâtir un nouvel estimateur à partir de F_n :

$$\hat{F}^s(x) = \frac{1}{2}(F_n(x) + 1 - F_n(-x)) \quad \text{pour } x < 0$$

et démontre que cet estimateur est l'estimateur du maximum de vraisemblance sous une hypothèse de symétrie. Il bâtit ensuite un nouvel estimateur $\hat{F}^{aux}$ en utilisant de l'information amenée par une covariable Y. Un échantillon auxiliaire $(Y_i)_{1\leq i\leq m}$ avec $m >> n$ permet de connaître la fonction de répartition empirique de la variable Y. En collectant ensuite un échantillon $Z_i = (X_i, Y_i)_{1\leq i\leq n}$ et en maximisant sa vraisemblance, il obtient l'estimateur $\hat{F}^{aux}$. Cet estimateur est relativement simple et ne fait intervenir qu'un comptage des X_i dans certaines zones. En croisant les 2 modèles, il obtient l'estimateur $\hat{F}^h$ suivant

$$\hat{F}^h(x) = \frac{1}{2}(\hat{F}^{aux}(x) + 1 - \hat{F}^{aux}(-x)) \quad \text{pour } x < 0,$$

montre sa normalité asymptotique et étudie ses propriétés. Il faut noter que cet estimateur ne maximise pas la vraisemblance de ce modèle et ne s'avère pas robuste envers l'hypothèse de symétrie. Enfin, il montre sur des simulations une amélioration du biais et de l'efficacité relativement à la fonction de répartition empirique.

L'estimateur de Grenander [53] est également obtenu directement à partir de la fonction de répartition empirique. Il est défini comme le plus petit majorant concave de F_n. Pour une fonction F_n de support réel $[a, b]$, F_{Gr} est la plus petite fonction concave telle que

$$F_{Gr}(t) \geq F_n(t) \text{ et } F_{Gr}(a) = 0\,,\ F_{Gr}(b) = 1.$$

Mais cet estimateur, bien qu'étant défini à partir de la fonction de répartition et sans aucun paramètre supplémentaire, est beaucoup plus souvent dérivé pour estimer une densité f qu'utilisé directement pour déterminer F. Dans le cas où la densité f est monotone, f_{Gr} l'est également et maximise la vraisemblance. Pour une étude plus approfondie de f_{Gr} nous renvoyons le lecteur à Devroye [27].

2.3 Estimateurs par lissage local

Afin d'obtenir une estimation plus régulière de la fonction de répartition, Berlinet [8] puis Lejeune et Sarda [67] lissent la fonction de répartition empirique dans différents espaces. Lejeune et Sarda utilisent la régression polynômiale locale. La minimisation de la norme pondérée L^2 débouche sur des choix optimaux que l'on retrouve parmi les estimateurs à noyaux décrits en section 2.4. Ce résultat est ensuite élargi par Abdous, Berlinet et Hengartner [1] qui proposent d'estimer différentes fonctionnelles $\phi(x, F)$ de la fonction F au point x. Pour cela ils substituent $\phi(x, F_n)$ à $\phi(x, F)$ et, dans le cas où $\phi(x, F)$ a r dérivées continues, choisissent de minimiser le critère suivant

$$J(a_0, ..., a_r; x) = \int \frac{1}{h} K\left(\frac{z-x}{h}\right) \left\{\phi(z, F_n) - \sum_{k=0}^{r} \frac{a_k}{k!}(z-x)^k\right\}^2 dz.$$

Les dérivées successives de $\phi(x, F)$ sont estimées par les $\hat{a}_0(x), \hat{a}_1(x), ..., \hat{a}_r(x)$ minimisant le critère $J(a_0, ..., a_r; x)$ au point x. Une expression explicite est ensuite obtenue à l'aide d'une fonction K, une densité sur [-1,1], et $Q_0(z), ..., Q_n(z)$ une base orthonormale de $L^2(K)$ de l'espace P_r des polynômes de degré au moins r. On définit

$$K^{[m,r]}(u) = \left(\sum_{k=0}^{r} Q_k(u) \left.\frac{d^m}{dw^m} Q_k(w)\right|_{w=0}\right) K(u)$$

et on obtient alors le minimiseur $\hat{a}_0(x), \hat{a}_1(x), ..., \hat{a}_r(x)$ par

$$\hat{a}_m = \frac{1}{h^{m+1}} \int \phi_n(z) K^{[m,r]}\left(\frac{z-x}{h}\right) dz.$$

En considérant l'estimateur

$$\hat{\theta}^{(m)}_{n,h}(x) = \frac{1}{h^{m+1}} \int_{-\infty}^{\infty} \phi(z, F_n) K^{[m,r]}\left(\frac{z-x}{h}\right) dz$$

de $\theta^{(m)}(x) = \phi^{(m)}(x, F)$, ils démontrent ensuite sous certaines conditions que $\hat{\theta}^{(m)}_{n,h}(x)$ converge p.s. vers $\phi^{(m)}(x, F)$. Berlinet [9] et Berlinet et Thomas-Agnan [11] élargissent ce résultat à un espace V de Hilbert à noyau reproduisant à la place de P_r. L'estimation de la projection $\Pi_V(F)$ de F sur V est alors déterminée par les équations

$$F_x(hv) = \int \Pi_V(F(x+h.))(u)K(u,v)K_0(u)d\lambda(u)$$

et

$$h^m F_x^{(m)}(hv) = \int \Pi_V(F(x+h.))(u)\frac{d^m K(u,v)}{dv^m}K_0(u)d\lambda(u)$$

où K_0 est un noyau et K le noyau reproduisant de V.

2.4 Estimateur à noyau

L'estimateur à noyau de la densité

$$\tilde{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x-X_i}{h}\right),$$

avec un noyau k intégrable et d'intégrale 1 et une fenêtre $h > 0$, est un estimateur non paramétrique bien connu de $f(x)$ introduit par Akaike [3], Rosenblatt [80] et Parzen [74]. La littérature qu'il a suscitée est considérable. Dans le présent paragraphe nous nous limitons à ses applications orientées vers l'estimation de la fonction de répartition. On définit l'estimateur $\tilde{F}_n$ à noyau de F par

$$\tilde{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right),$$

où

$$K(x) = \int_{-\infty}^{x} k(y)dy.$$

Ses propriétés sont connues depuis longtemps, par exemple sa convergence uniforme vers F avec f continue (Nadaraya [72], Winter [98], Yamato [102]) puis sans conditions sur f (Singh, Gasser et Prasad, [87]) ou sa normalité asymptotique (Watson et Leadbetter, [97]). Winter [99] démontre aussi qu'il vérifie la propriété de Chung-Smirnov, c'est-à-dire

$$\limsup_{n\to\infty}\left\{\left(\frac{2n}{\log\log n}\right)^{(1/2)}\sup_{-\infty<t<\infty}|\tilde{F}_n(t)-F(t)|\right\} \leq 1$$

avec probabilité 1. Azzalini [4] trouve une expression asymptotique pour l'erreur quadratique moyenne ou M.S.E. $(E(\tilde{F}_n(x)-F(x))^2)$ et détermine la

fenêtre asymptotiquement optimale permettant d'avoir un M.S.E. plus faible que pour F_n. Reiss [77] prouve que l'inefficacité relative asymptotique de F_n par rapport à $\tilde{F}_n$ tend rapidement vers l'infini quand la taille de l'échantillon augmente avec un choix approprié de noyau, par exemple

$$k(x) = \frac{9}{8}\left(1 - \frac{5}{3}x^2\right) I_{[-1,1]}(x),$$

et certaines conditions vérifiées notamment lorsque le support de k est borné et

$$\int_{-\infty}^{\infty} tk(t)K(t)dt > 0. \tag{2.1}$$

Falk [39] donne ensuite une solution complète à ce problème en établissant la représentation de l'inefficacité relative de F_n par rapport à $\tilde{F}_n$ sous les conditions ci-dessus notamment lorsque le support de k est borné. Le nombre $\psi(k) = \int 2k(x)K(x)xdx$ est introduit par Falk [40] comme une mesure de la performance asymptotique du noyau k. Mais il démontre qu'aucun noyau de carré intégrable ne minimise ψ. Il utilise alors le nombre $\phi(k) = \int k(x)^2 dx$ défini par Epanechnikov [38] comme une mesure de la performance du noyau en estimation de la densité. Au sens de ϕ, le noyau d'Epanechnikov suivant

$$k(x) = (3/4)(1 - x^2)I_{(|x|\leq 1)}.$$

est le meilleur mais les noyaux gaussiens ou uniformes ont des performances très proches. En utilisant le critére ψ le noyau d'Epanechnikov est alors de loin le meilleur des trois.

Falk [40] montre ensuite que cette inefficacité relative s'applique aussi aux estimateurs des quantiles q_n de F_n par rapport aux quantiles $\tilde{q}_n$ de $\tilde{F}_n$. Enfin, Golubev et Levit [51] donnent les conditions permettant de trouver un estimateur minimax du second ordre pour la fonction de perte carrée $L(F, a) = \int (F(t) - a(t))^2 dF(t)$.
Au sens de l'erreur quadratique moyenne intégrée ou I.M.S.E., le meilleur noyau est le noyau uniforme bien que les performances d'autres noyaux (Epanechnikov, normal, triangulaire) ne soient, en pratique, que légèrement moins

bonnes (Jones [61]). Il est intéressant de noter que ce ne sera pas le meilleur noyau dans le cadre d'estimation de la densité.
L'expression asymptotique de l'I.M.S.E. est également étudiée par Swanepoel [93]. Pour une fonction f continue, il prouve que le meilleur noyau est le noyau uniforme $k(x) = (1/2\omega)I_{[-\omega,\omega]}(x)$ pour une constante arbitraire $\omega > 0$ (ce qui démontre que les critères de Falk pour définir un noyau optimal ne sont en fait pas adaptés à la fonction de répartition) alors que, pour f discontinue en un nombre fini de points, c'est le noyau exponentiel $k(x) = (c/2)\exp(-c|x|)$ pour une constante arbitraire $c > 0$. $\tilde{F}_n$ est ici aussi plus efficace que F_n pour $h_n = o(n^{-1/2})$ sous la condition (2.1).

Néanmoins, $\tilde{F}_n$ ne fournit pas toujours une meilleure estimation que F_n. En effet, dans le cas d'une fonction F uniformément lipschitzienne Fernholz [45] obtient que

$$\sqrt{n}||\tilde{F}_n - F_n||_\infty \to 0 \ p.s.$$

et que $\sqrt{n}||\tilde{F}_n - F||_\infty$ et $\sqrt{n}||\tilde{F}_n - F||_\infty$ ont la même distribution asymptotique. De plus, Shirahata et Chu [85] démontrent que sous certaines hypothèses sur F, l'erreur quadratique intégrée (I.S.E.) ($\int_{-\infty}^{\infty}(\tilde{F}_n(x)-F(x))^2dF(x)$) de $\tilde{F}_n$ est presque sûrement supérieure à celle de F_n.

2.5 Estimateurs splines

Les méthodes splines connaissent un large spectre d'applications de par la simplicité de leur mise en oeuvre, de la régularité des courbes obtenues et de la multiplicité des conditions que l'on peut imposer aux solutions. Néanmoins, les fonctions à estimer doivent obéir à certaines conditions de régularité et le nombre d'observations doit être suffisament grand pour éviter les phénomènes classiques de sur ou sous-lissage. Pour plus de précision sur le sujet on pourra se référer à Besse et Thomas-Agnan [13], Wahba [96] ou Green et Silverman [52].

Berlinet [8] utilise des splines cubiques pour lisser la fonction de répartition empirique et obtenir un estimateur uniformément asymptotiquement sans biais de la fonction de répartition. Les splines cubiques sont ici définies comme

un polynôme de degré au plus 3 interpolant la fonction de répartition empirique sur l'ensemble des points de l'échantillon S_n avec certaines conditions aux limites.
Une approche différente est étudiée par Restle [78] qui définit l'estimateur F_{sp} de F sous forme d'une spline naturelle cubique en définissant les valeurs de F_{sp} aux points $X_1, ..., X_n$ par

$$\mathbf{F}_{sp} = \left(\mathbf{I} + \alpha \mathbf{W}^{-1}\mathbf{K}\right)^{-1} \mathbf{F}_n$$

où α est le paramètre de lissage, la matrice $\mathbf{W}$ contient des poids et la matrice aléatoire $\mathbf{K}$ ne dépend que des écarts $H_i = T_{i+1} - T_i$ de l'échantillon. A cause des conditions de continuité imposées aux splines naturelles cubiques, les valeurs aux points $T_1, ..., T_n$ caractérisent la fonction F_{sp} complètement. Cet estimateur possède une erreur quadratique intégrée de l'ordre de $O_p(n^{-1})$ et le supremum de la différence absolue de F et F_{sp} est de l'ordre de $O_p(n^{-1/4})$. De plus, la probabilité que cet estimateur soit monotone tend vers 1.

2.6 Les Support Vector Machines

L'idée originale des Support Vector Machines (S.V.M.) est publiée par Vapnik [94] puis reprise dans Burges [19], Schölkopf, Burges et Smola [83], Schölkopf et Smola [84] et plus récemment dans Vapnik et Kotz [95]. Ces dernières années ont vu une explosion du nombre de travaux exploitant la méthode des S.V.M. dont le but premier est de résoudre certains problèmes de classification. Elle est basée sur l'utilisation de fonctions dites noyaux qui permettent une séparation optimale (sans problème d'optimum local) des points observés $X_1, ..., X_n$ en différentes catégories. Afin de remédier au problème de l'absence de séparateur linéaire, l'idée des S.V.M. est de reconsidérer le problème dans un espace de dimension supérieure. Dans ce nouvel espace, il existe un séparateur linéaire qui permet de classer au mieux les points observés dans les groupes qui conviennent. On peut ensuite projeter le séparateur linéaire dans l'espace d'origine pour visualiser le résultat de la classification. Le changement d'espace se fait au moyen d'une fonction symétrique $k(.,.)$ répondant au critère de Mercer, c'est à dire telle que, pour $1 \leq i, j \leq n$, $(k(X_i, X_j))_{i,j}$

est une matrice définie positive. Ce critère autorise un changement "dans les deux sens" ce qui permet à partir de l'expression de l'hyperplan dans l'espace de dimension plus élevé de classer les éléments dans l'espace de description initial.
Afin d'utiliser les S.V.M., le problème d'estimation de la densité est vu comme le problème suivant : on choisit tout d'abord un ensemble de densités $f(x, \alpha)$ où α est l'ensemble des paramètres à déterminer. La résolution du problème

$$\int_{-\infty}^{x} f(t, \alpha)dt = F(x)$$

revient donc à résoudre le problème inverse

$$Af(., \alpha) = F(., \alpha)$$

où A est une application linéaire de l'espace de Hilbert des fonctions $f(., \alpha)$ dans celui des fonctions $F(., \alpha)$. L'estimation de la densité est donc maintenant traduite en un problème inverse qui se résout en utilisant F_n pour estimer F. On peut alors utiliser les S.V.M. pour estimer la densité de différentes manières (Schölkopf, Burges et Smola [83]). Mohamed et Farag [71] choisissent tout d'abord de coupler l'approche S.V.M. et la théorie de champ moyen (qui permet des approximations efficaces). Puis ils incorporent l'algorithme EM à l'approche précédente (Mohamed, El-Baz et Farag [70]). Les estimateurs de $f(x)$ et $F(x)$ sont alors

$$\bar{f}_n^*(x) = \sum_{i=1}^{n} \omega_i k(x, X_i) \text{ et } \bar{F}_n^*(x) = \sum_{i=1}^{n} \omega_i K(x, X_i)$$

où k est un noyau vérifiant certaines propriétés (Vapnik [94]) dérivé de K et ω_i obtenus dans Mohamed, El-Baz et Farag [70]. L'algorithme EM est ensuite utilisé pour trouver les paramètres optimaux pour le noyau, ici la matrice de covariance pour un noyau gaussien, et permet un gain substantiel (notamment en terme de distance de Levy) par rapport à la première approche où on utilise les paramètres ad-hoc.

2.7 Le level-crossing

Le level-crossing, passage à niveau en français, est défini dans Leadbetter [65]. Si nous notons $X_{(1)}, X_{(2)}, ..., X_{(n)}$ les statistiques d'ordre de l'échantillon

$X_1, X_2, ..., X_n$ et prenons $x \in \mathbb{R}$ alors on dit que l'intervalle $[X_{(j)}, X_{(k)})$ "croise" le niveau x si et seulement si $x \in [X_{(j)}, X_{(k)})$, $1 \leq j \leq k < n$. Il est ensuite possible de définir une fonction $r_n(x)$ comptant le nombre de "croisements" se produisant au niveau x à partir d'un échantillon de départ $\{X_i,\ i = 1, ..., n\}$.
L'espérance de cette fonction $r_n(x)$ s'exprime en fonction de $F(x)$. Huang et Brill [59] appliquent le level-crossing à l'estimation de la fonction de répartition. L'estimateur obtenu diffère de la fonction de répartition empirique qui donne un poids équivalent de $1/n$ à chaque valeur de l'échantillon. Ils essayent d'améliorer l'efficacité en faisant varier les poids des observations. Des poids moins importants sont attribués aux données situées dans les queues de l'échantillon. Le nouvel estimateur est alors défini par

$$\hat{F}_n^*(x) = \sum_{i=1}^{n} I_{(-\infty,x]}(X_{(i)}) p_{n,i},\ n \geq 2,$$

avec

$$p_{n,i} = \begin{cases} \frac{1}{2}\Big[1 - \frac{n-2}{\sqrt{n(n-1)}}\Big], & \text{pour } i = 1, n, \\ \frac{1}{\sqrt{n(n-1)}}, & \text{pour } i = 2, .., n-1. \end{cases}$$

Nous remarquons que les valeurs $X_{(1)}$ et $X_{(n)}$ ont moins de poids que les valeurs centrales de l'échantillon. Ils obtiennent tout d'abord

$$P\Big\{ \lim_{n\to\infty} \sup_{-\infty<x<\infty} |\hat{F}_n^*(x) - F(x)| = 0 \Big\} = 1$$

et démontrent ensuite que pour toute fonction F continue, il existe un nombre réel $z_n \in (0, \frac{1}{2})$, tel que l'efficacité de leur estimateur par rapport à F_n est plus grande que 1 sur les intervalles ouverts $(0, z_n)$ et $(1-z_n, 1)$. Ces résultats théoriques sont ensuite complétés par des simulations sur lesquelles le level-crossing donne de meilleures estimations que la fonction de répartition empirique.

2.8 Les Systèmes de Fonctions Itérées

Les systèmes de fonctions itérées (I.F.S.), introduits dans Barnsley et Demko [6], sont principalement utilisés pour la modélisation de fractales.

2.9 Les Systèmes de Fonctions Itérées

Cette théorie est entièrement basée sur la propriété d'invariance par changement d'échelle. Ainsi une application à l'approximation de fonction est récemment proposée dans Forte et Vrscay [48]. Le principe est le suivant : on suppose que ω est une fonction contractante d'un espace métrique M dans lui-même. Le théorème du point fixe nous donne ensuite l'existence et l'unicité d'un point fixe x' dans M tel que $\omega(x') = x'$. De plus, pour tout point de départ x_0, la suite de points (x_n) définie par $x_{n+1} = \omega(x_n)$ converge vers x' lorsque n tend vers l'infini. C'est pourquoi x' est appelé attracteur de l'I.F.S.

On peut maintenant supposer qu'il existe n fonctions contractantes ω_i, $i = 1, \ldots, n$, dans M. Chacune de ces fonctions ω_i a son propre point fixe x'_i et si on applique une de ces fonctions, par exemple ω_l, alors la suite (x_n) converge vers le point fixe x'_l.
Ce sont ces résultats que Iacus et La Torre [60] utilisent en interprétant l'estimation de F comme la recherche du point fixe d'une famille de fonctions contractantes sur un espace métrique complet. Le problème se définit alors à l'aide de N fonctions affines ω et d'un vecteur de paramètres p. Pour une famille donnée W, le résultat dépend uniquement du choix de p. Il faut donc maintenant déterminer p en résolvant un problème d'optimisation quadratique sous contraintes. Cette approche est appelée *approche inverse*. Ils démontrent qu'une fois la famille de fonctions choisie le problème inverse revient à résoudre un problème quadratique dépendant uniquement des moments non centrés de l'échantillon.

Le choix $p_i = 1/N$ nous permet d'obtenir une version lissée de la fonction de répartition empirique et, par conséquent, de bonnes propriétés asymptotiques et cela même si le support de la loi n'est pas compact. Différentes simulations sur des fonctions bêta sont testées et l'I.F.S. apparaît comme plus efficace que les fonctions de répartition empiriques ou à noyaux en terme d'erreur quadratique moyenne ou de norme infinie. Cette méthode est notamment très performante dans les cas de données manquantes ou de petits échantillons mais nécessite une fonction à support compact ou au moins une connaissance du support.

2.9 D'autres estimateurs

Comme il a été dit précédemment tout estimateur de la densité fournit par intégration un estimateur de la fonction de répartition. Nous évoquons donc dans le présent paragraphe quelques familles importantes de tels estimateurs qui n'ont pas été mentionnées auparavant. Nous resterons volontairement succincts, le nombre de références sur le sujet concernant l'estimation de la densité étant considérable. Les premiers travaux traitant de manière unifiée des estimateurs de la densité et des conditions de leur convergence sont ceux de Foldes et Revesz [47] et Bleuez et Bosq [15, 16]. Ces auteurs ont considéré la famille des estimateurs de la densité de la forme

$$f_n(t) = \frac{1}{n}\sum_{j=1}^{n} K_{r(n)}(X_j, t),$$

où $(K_r(x,t), r \in I)$ est une famille de fonctions réelles mesurables et I une partie non bornée de $\mathbb{R}^+$. Cette famille contient les estimateurs à noyaux vus précédemment et ceux utilisant la méthode des fonctions orthogonales qui correspond au choix

$$K_r(x,t) = \sum_{i=0}^{r} e_i(x)e_i(t),$$

$r \in \mathbb{N}$ et où les e_i forment une base de hilbertienne de $L^2(\mathbb{R})$. Elle englobe également les histogrammes, de nombreux estimateurs basés sur des partitions (Bosq et Lecoutre [17], Berlinet et Biau [10]) et les estimateurs par ondelettes.

La théorie des ondelettes s'est développée au milieu des années 80 et est très utilisée dans de nombreux domaines (analyse harmonique, traitement du signal, compression d'images, statistique fonctionnelle...). Son succès est du à son adaptativité aux données et à sa facilité d'implémentation. Une ondelette est une onde "localisée". Lorsqu'on décompose une fonction en séries de Fourier, on la décompose en fait en fréquence. La décomposition en ondelettes ajoute une dimension, la décomposition est double : fréquentielle et spatiale. On décompose en effet une fonction comme une somme d'oscillations de fréquences précises se produisant à un endroit précis. La construction d'une base d'ondelettes dans un espace convenable permet d'estimer une fonction

si on sait estimer ses coordonnées dans la base choisie. On peut citer comme références Donoho, Johnstone, Kerkyacharian et Picard [31], Härdle, Kerkyacharian, Picard et Tsybakov [55] ou Herrick, Nason et Silverman [57]. Mais la plupart sont appliquées à l'estimation de la densité et il n'existe pas, à notre connaissance, d'application visant à obtenir des propriétés spécifiques d'estimateurs de la fonction de répartition.

Nous pouvons aussi évoquer brièvement une approche statistique différente : l'approche bayésienne. Dans la perspective bayésienne un état de connaissance initial est traduit par une loi de probabilité a priori sur les paramètres du modèle choisi. Le théorème de Bayes permet de passer ensuite à une loi dite a posteriori sur les paramètres conditionnellement aux données observées. Cette approche est notamment développée en détails dans Robert [79] et Bernardo et Smith [12]. Son application à l'estimation de la fonction de répartition, sous différents modèles, est évoquée dans Korwar et Hollander [64], Susarla et Van Ryzin [92] et Pantazopoulos, Pappis, Fifis, Costopoulos, Vaughan et Gasparini [73]. En statistique non paramétrique la construction des estimateurs bayésiens est souvent difficile. On pourra consulter Bosq et Lecoutre [17] au sujet de l'utilisation d'un processus de Dirichlet comme loi a priori. L'estimateur résultant a des propriétés asymptotiques analogues à celles de la fonction de répartition empirique.

Enfin, les méthodes combinatoires, développées par Devroye et Lugosi [29], apportent un regard nouveau sur le choix efficace d'estimateurs non paramétriques et sont applicables aux estimateurs de la fonction de répartition.

2.10 Fonction de répartition conditionnelle

Lorsque les données sont disponibles sous la forme de couples $(X_i, Y_i)_{1\leq i\leq n}$, on peut chercher à estimer la fonction de répartition conditionnelle $\pi(y|x) = P(Y_i \leq y|X_i = x)$. Roussas [81] définit son estimateur basé sur l'estimateur à noyau par

$$\pi_G(y|x) = \int_{-\infty}^{y} q_n(x, y')/p_n(x)$$

où

$$p_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \text{ et } q_n(x, y') = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h^{1/2}}\right) K\left(\frac{y' - Y_i}{h^{1/2}}\right).$$

Sous certaines hypothèses fortes, il démontre sa convergence en probabilité ainsi que celle de ses quantiles (voir aussi Samanta [82]). Stone [89] définit lui un estimateur de type

$$\hat{\pi}_n(A|X) = \sum_{i=1}^{n} W_{ni}(X) I_A(Y_i)$$

où W_{ni} est une fonction de poids propre à chaque X_i. Il obtient la consistance de l'estimateur sous certaines conditions sur les fonctions de poids. En s'appuyant sur ce travail, Cleveland [22] définit l'estimation polynômiale locale. L'idée sous-jacente est en fait un simple développement de Taylor : soit x et y fixés et supposons que la fonction qui à z associe $\pi(y|z)$ est suffisament régulière dans un voisinage de x. Alors, par une approximation de Taylor,

$$\pi(y|z) \approx \pi(y|x) + (z - x)\pi^{(1)}(y|x) + ... + (z - x)^p \frac{\pi^{(p)}(y|x)}{p!}$$

où $\pi^{(v)}(y|x)$ est la $v^{\text{ème}}$ dérivée de $\pi(y|z)$ par rapport à z, évaluée en x. Nous approximons donc localement $\pi(y|z)$ par un polynôme de degré p en $(z - x)$. Fan [41, 42] et Fan et Gijbels [43, 44] montrent que cet estimateur présente plusieurs avantages sur les autres méthodes de régression (comme celle du noyau de Gasser et Müller [50]) : elle contourne la principale difficulté des estimateurs à noyaux classiques en corrigeant automatiquement les effets de bord tout en conservant les propriétés d'optimalité théorique. Il est à noter que Ferrigno et Ducharme [46] construisent un test permettant de rejeter (ou de valider) un modèle en fonction de sa "distance" à l'estimateur polynômial local.

Stute [90, 91] définit lui un estimateur de type plus-proche-voisins

$$\pi_n(y|x) = \frac{1}{nh}\sum_{i=1}^{n} Y_i K\left(\frac{F_n(x) - F_n(X_i)}{h}\right)$$

où K est un noyau. Sous certaines conditions, il obtient sa normalité asymptotique et celle de ses quantiles.

Hall, Wolff et Yao [54] proposent deux méthodes différentes pour estimer cette fonction. Ils définissent ainsi un estimateur à noyau adapté (qui sera équivalent à un estimateur localement linéaire) et un estimateur faisant intervenir un modèle logistique. L'estimateur à noyau se définit par

$$\tilde{\pi}(y|x) = \frac{\sum_{i=1}^{n} I_{(Y_i \leq y)} p_i(x) K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} p_i(x) K\left(\frac{X_i - x}{h}\right)},$$

où K est un noyau de fenêtre h et p_i, pour $1 \leq i \leq n$, des poids uniquement définis par plusieurs équations.
Le modèle logistique général pour $P(x) = \pi(y|x)$ admettant $r-1$ dérivées continues est de la forme

$$L(x, \mathbf{b}) = A(x, \mathbf{b}) / \{1 + A(x, \mathbf{b})\},$$

où $A(., \mathbf{b})$ est une fonction non-négative dépendant du vecteur de paramètres $\mathbf{b} = (b_1, ..., b_r)$ représentant $P(x)$ et ses $r-1$ dérivées. On peut noter qu'il est possible de trouver des fonctions A relativement simples. Ce modèle amène à l'estimateur $\hat{\pi}(y|x) = L(0, \hat{\mathbf{b}}_{xy})$ où $\hat{\mathbf{b}}_{xy}$ minimise la fonction

$$R(\mathbf{b}; x, y) = \sum_{i=1}^{n} \{I_{(Y_i \leq y)} - L(X_i - x, \mathbf{b})\}^2 K\left(\frac{X_i - x}{h}\right).$$

Ces deux estimateurs font intervenir un noyau K. Ils seront consistants selon le choix toujours délicat de la fenêtre h.
Il est aussi possible d'estimer F en utilisant F_n et les techniques classiques de régression dans le modèle

$$F_n(X_i) = F(X_i) + \varepsilon_i,\ i = 1, ..., n.$$

En utilisant un estimateur localement linéaire de même type que celui décrit ci-dessus, Cheng et Peng [21]) déterminent un estimateur ayant une moyenne d'erreur quadratique intégrée plus petite que l'estimateur à noyau classique.

2.11 Données biaisées

En pratique, il peut arriver qu'observer directement la variable X de densité f soit impossible mais qu'on puisse, par contre, observer une variable Y à

travers un échantillon $Y_1, ..., Y_n$ et possédant la densité

$$g(y) = w(y)f(y)/\mu(f)$$

où $w(x)$ va être la fonction de biais et $\mu(f) = E_f\{w(X)\} = 1/E_g\{w^{-1}(Y)\}$. On rencontre également cette situation dans le cas de données manquantes (Berlinet et Thomas-Agnan [11]). Dans ce qui suit on suppose que la fonction $w(x)$ est connue, intégrable et que

$$\forall x,\ 0 < c_1 < w(x) < c_2 < \infty.$$

Patil, Rao et Zelen [75] référencent les différentes propriétés de ces fonctions biaisées pour laquelle Cox [25] propose la distribution suivante :

$$\tilde{F}_{co}(x) = \hat{\mu} n^{-1} \sum_{l=1}^{n} w^{-1}(Y_l) I_{(Y_l \leq x)},$$

avec

$$\hat{\mu} = \frac{1}{n^{-1} \sum_{l=1}^{n} w^{-1}(Y_l)}.$$

Cet estimateur est celui du maximum de vraisemblance dans le cadre non-paramétrique, il est asymptotiquement efficace et est un estimateur minimax du premier ordre. Efromovitch [36] modifie l'estimateur de Cox et propose le suivant :

$$\tilde{F}_e(x) = x + \sqrt{2} \sum_{j=1}^{J} \hat{\theta}_j (\pi j)^{-1} \sin(\pi j x),$$

où

$$\hat{\theta}_j = n^{-1} \sqrt{2} \hat{\mu} \sum_{l=1}^{n} w^{-1}(Y_l) \cos(\pi j Y_l),$$

ce qui est l'estimateur moyen de Cox pour θ_j, et J une constante calculable, fonction de n. Il obtient

$$|E\{\tilde{F}_e(x)\} - F(x)| \leq \frac{o(1)}{\sqrt{n} \log(n)}$$

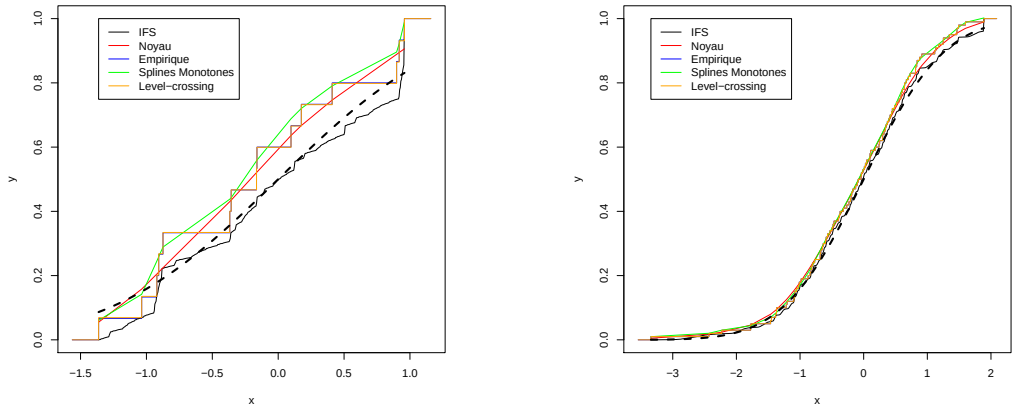
et la normalité asymptotique de son estimateur qui est de plus un estimateur minimax du second ordre en x.

2.12 Conclusion

L'estimateur le plus simple, la fonction de répartition empirique a de bonnes propriétés de convergence mais possède certains inconvénients comme celui de ne pas prendre en compte une éventuelle information supplémentaire ou bien le fait d'être une fonction en escalier. Dès que l'on restreint quelque peu le modèle envisagé pour les données il existe des estimateurs qui sont préférables à la fonction de répartition empirique. Néanmoins l'existence de ce premier estimateur, au contraire de ce qui se passe pour la densité, donne des facilités quant à l'utilisation de méthodes de lissage. Les méthodes d'estimation que nous avons passées en revue fournissent en général une fonction qui possède les propriétés caractérisant une fonction de répartition sauf parfois en ce qui concerne la masse totale ($\lim_{x\to+\infty} F(x) = 1$), ce qui est facilement corrigé, ou bien en ce qui concerne la croissance. On peut alors penser à appliquer certaines méthodes d'estimation de fonctions monotones (Delecroix, Simioni et Thomas-Agnan [26]), la régression ou les splines isotoniques (Barlow, Bartholomew, Brewner et Brunk [5] et Wegman et Wright [101]) afin de garantir la croissance de l'estimateur.

2.13 Simulations

Nous testons certaines des méthodes présentées ci-dessus sur des échantillons de tailles différentes. Les différents estimateurs sont explicités dans la bibliographie citée ci-dessus sauf l'estimateur par splines monotones qui provient d'un package développé par Stephan Ellner à partir de l'article de Wood [100]. Nous pouvons remarquer assez logiquement l'importance de la taille de l'échantillon mais aussi, comme annoncé par les auteurs de cette méthode, la meilleure estimation de l'estimateur I.F.S. dans le cas de petits échantillons (efficacité confirmée dans le cas d'échantillons avec des valeurs manquantes). Nous n'utiliserons cependant pas ou très peu cette caractéristique du fait du grand nombre d'observations dont nous avons besoin pour estimer l'indice de régularité en règle générale.



(a) $n = 15$ (b) $n = 100$

(c) $n = 1000$ (d) $n = 10000$

FIGURE 2.1 – Estimations de la fonction de répartition d'une N(0,1) en pointillés pour différentes méthodes et différentes tailles d'échantillon n

Bibliographie

[1] B. Abdous, A. Berlinet, and N. Hengartner. A general theory for kernel estimation of smooth functionals of the distribution function and their derivatives. *Revue Roumaine de Mathématiques Pures et Appliquées*, 48 :217–232, 2003.

[2] O. Aggarwal. Some minimax invariant procedures for estimating a cumulative distribution function. *The Annals of Mathematical Statistics*, 26 :450–462, 1955.

[3] H. Akaike. An approximation to the density function. *Annals of the Institute of Statistical Mathematics*, 6 :127–132, 1954.

[4] A. Azzalini. A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, 68 :326–328, 1981.

[5] R. Barlow, D. Bartholomew, J. Brewner, and H. Brunk. *Statistical inferences under order restrictions*. Wiley, New-York, 1972.

[6] M. Barnsley and S. Demko. Iterated function systems and the global construction of fractals. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 399 :243–275, 1985.

[7] R. Beran. Estimating a distribution function. *The Annals of Statistics*, 5 :400–404, 1977.

[8] A. Berlinet. Convergence des estimateurs splines de la densité. *Publications de l'Institut de Statistique de l'Université de Paris*, 26 :1–16, 1981.

[9] A. Berlinet. Reproducing kernels and finite order kernels. In G. Roussas, editor, *Nonparametric Functional Estimation and Related Topics*. Klüwer Academic Publishers, Dordrecht, 1991.

[10] A. Berlinet and G. Biau. Estimation de densité et prise de décision. In R. Lengellé, editor, *Décision et Reconnaissance de Formes en Signal*. Hermes, Paris, 2002.

[11] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Klüwer, Boston, 2004.

[12] J. Bernardo and A. Smith. *Bayesian Theory*. Wiley, New-York, 1994.

[13] P. Besse and C. Thomas-Agnan. Le lissage par fonctions splines en statistique : revue bibliographique. *Statistique et Analyse des données*, 14 :55–83, 1989.

[14] P. Billingsley. *Convergence of Probability Measures*. Wiley, New-York, 1968.

[15] J. Bleuez and D. Bosq. Conditions nécessaires et suffisantes de convergence pour une classe d'estimateurs de la densité. *Comptes rendus de l'Académie des Sciences de Paris Série A*, 282 :63–66, 1976.

[16] J. Bleuez and D. Bosq. Conditions nécessaires et suffisantes de convergence pour une classe d'estimateurs de la densité par la méthode des fonctions orthogonales. *Comptes rendus de l'Académie des Sciences de Paris Série A*, 282 :1023–1026, 1976.

[17] D. Bosq and J.-P. Lecoutre. *Théorie de l'Estimation Fonctionnelle*. Economica, 1987.

[18] L. Brown. Admissibility in discrete and continuous invariant nonparametric estimation problems, and in their multinomial analogs. *The Annals of Statistics*, 16 :1567–1593, 1988.

[19] C. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2 :1–47, 1998.

[20] P. Caperaa and B. Van Cutsem. *Méthodes et modèles en statistique non paramétrique*. Bordas, Paris, 1988.

[21] M.-Y. Cheng and L. Peng. Regression modelling for nonparametric estimation of distribution and quantiles functions. *Statistica Sinica*, 12 :1043–1060, 2002.

[22] W. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74 :829–836, 1979.

[23] M. Cohen and L. Kuo. The admissibility of the empirical distribution function. *The Annals of Statistics*, 13 :262–271, 1985.

[24] G. Collomb, S. Hassani, P. Sarda, and P. Vieu. Estimation non paramétrique de la fonction de hasard pour des observations dépendantes. *Statistique et Analyse des Données*, 10 :42–49, 1985.

[25] D. Cox. Some sampling problems in technology. In N. L. Johnson and H. Smith, editors, *New Developments in Survey Sampling*. Wiley, New-York, 1969.

[26] M. Delecroix, M. Simioni, and C. Thomas-Agnan. Functional estimation under shape constraints. *Journal of Nonparametrics Statistics*, 6 :69–89, 1996.

[27] L. Devroye. *A Course in Density Estimation*. Birkhäuser, Boston, 1987.

[28] L. Devroye and L. Györfi. Distribution and density estimation. *CISM Courses and Lectures*, 434 :221–270, 2002.

[29] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, New-York, 2001.

[30] L. Devroye and G. Wise. On the recovery of discrete probability densities from imperfect measurements. *Journal of the Franklin Institute*, 307 :1–20, 1979.

[31] D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *The Annals of Statistics*, 24 :508–539, 1996.

[32] M. Donsker. Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *The Annals of Mathematical Statistics*, 23 :277–281, 1952.

[33] J. Doob. Heuristic approach to the kolmogorov-smirnov theorm. *The Annals of Mathematical Statistics*, 20 :393–403, 1949.

[34] J. Doob. *Stochastic processes*. Wiley, New-York, 1953.

[35] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 33 :642–669, 1956.

[36] S. Efromovich. Distribution estimation for biased data. *Journal of Statistical Planning and Inference*, 124 :1–43, 2004.

[37] B. Efron and R. Tibshirani. *An introduction to the Bootstrap*. Chapman & Hall, London, 1993.

[38] A. Epanechnikov. Nonparametric estimation of a multivariate probability density. *Theory of Probability and its Applications*, 14 :153–158, 1969.

[39] M. Falk. Relative efficiency and deficiency of kernel type estimators of smooth distribution functions. *Statistica Neerlandica*, 37 :73–83, 1983.

[40] M. Falk. Relative deficiency of kernel type estimators of quantiles. *The Annals of Statistics*, 12 :261–268, 1984.

[41] J. Fan. Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, 87 :998–1004, 1992.

[42] J. Fan. Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, 21 :196–216, 1993.

[43] J. Fan and I. Gijbels. Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, 20 :2008–2036, 1992.

[44] J. Fan and I. Gijbels. *Local polynomial modelling and its applications*. Chapman & Hall, London, 1996.

[45] L. Fernholz. Almost sure convergence of smoothed empirical distribution functions. *Scandinavian Journal of Statistics*, 18 :255–262, 1991.

[46] S. Ferrigno and G. Ducharme. A global test of goodness-of-fit for the conditional distribution function. *Comptes Rendus Mathématiques de l'Académie des Sciences de Paris*, 341 :313–316, 2005.

[47] A. Foldes and P. Revesz. A general method for density estimation. *Studia Scientiarum Mathematicarum*, 9 :81–92, 1974.

[48] B. Forte and E. Vrscay. Solving the inverse problem for function/image approximation using iterated function systems. *Fractals*, 2 :325–334, 1995.

[49] Y. Friedman, A. Gelman, and E. Phadia. Best invariant estimation of a distribution function under the Kolmogorov-Smirnov loss function. *The Annals of Statistics*, 16 :1254–1261, 1988.

[50] T. Gasser and H. Müller. Estimating regression function and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, 3 :171–185, 1984.

[51] G. Golubev and B. Levit. Distribution function estimation : adaptive smoothing. *Mathematical Methods of Statistic*, 5 :383–403, 1996.

[52] P. Green and B. Silverman. *Nonparametric Regression and Generalized Linear Models. A Roughness Penalty Approach*. Chapman & Hall, London, 1994.

[53] U. Grenander. On the theory of mortality measurement part II. *Skandinavisk Aktuarietidskrift*, 39 :125–153, 1956.

[54] P. Hall, R. Wolff, and Q. Yao. Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94 :154–163, 1999.

[55] W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov. Wavelets, approximation and statisticals applications. *Lecture Notes in Statististics*, 129, 1999.

[56] P. Hennequin and A. Tortrat. *Théorie des Probabilités et Quelques applications*. Masson, Paris, 1965.

[57] D. Herrick, G. Nason, and B. Silverman. Some new methods for wavelet density estimation. *Sankhya*, A63 :394–411, 2001.

[58] I. Hu. A uniform bound for the tail probability of Kolmogorov-Smirnov statistics. *The Annals of Statistics*, 13 :811–826, 1985.

[59] M. Huang and P. Brill. A distribution estimation method based on level crossings. *Journal of Statistical Planning and Inference*, 124 :45–62, 2004.

[60] S. Iacus and D. La Torre. A comparative simulation study on the IFS distribution function estimator. *Nonlinear Analysis : Real World Applications*, 6 :858–873, 2005.

[61] M. Jones. The performance of kernel density functions in kernel distribution function estimation. *Statistics and Probability Letters*, 9 :129–132, 1990.

[62] J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *The Annals of Mathematical Statistics*, 27 :887–906, 1956.

[63] A. Kolmogorov. Sulla determinazione empirica di una legge de distribuzione. *Giornale dell'Instituto Italiano degli Attuari*, 4 :83–91, 1933.

[64] R. Korwar and M. Hollander. Empirical Bayes estimation of a distribution function. *The Annals of Statistics*, 4 :581–588, 1976.

[65] M. Leadbetter. Point processes generated by level crossings. In P. Lewis, editor, *Stochastic Point processes : Statistical Analysis, Theory and Applications*. Wiley-Interscience, New-York, 1972.

[66] E. Lehmann. *Theory of Point Estimation*. Wiley, New-York, 1983.

[67] M. Lejeune and P. Sarda. Smooth estimators of distribution and density functions. *Computational Statistics and Data Analysis*, 14 :457–471, 1992.

[68] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18 :1269–1283, 1990.

[69] R. Modarres. Efficient nonparametric estimation of a distribution function. *Computational Statistics and Data Analysis*, 39 :75–95, 2002.

[70] R. Mohamed, A. El-Baz, and A. Farag. Probability density estimation using advanced support vector machines and the EM algorithm. *International Journal of Signal Processing*, 1 :260–264, 2004.

[71] R. Mohamed and A. Farag. Mean field theory for density estimation using support vector machines. *Seventh International Conference on Information Fusion, Stockholm*, pages 495–501, 2004.

[72] E. Nadaraya. Some new estimates for distribution function. *Theory of Probability and its Application*, 9 :497–500, 1964.

[73] S. Pantazopoulos, C. Pappis, T. Fifis, C. Costopoulos, J. Vaughan, and M. Gasparini. Nonparametric Bayes estimation of a distribution function with truncated data. *Journal of Statistical Planning and Inference*, 55 :361–369, 1996.

[74] E. Parzen. On the estimation of a probability density and mode. *The Annals of Mathematical Statistics*, 33 :1065–1076, 1962.

[75] G. Patil, C. Rao, and M. Zelen. Weighted distribution. In N. L. Johnson and S. Kotz, editors, *Encyclopedia of Statistical Sciences*. Wiley, New-York, 1988.

[76] E. Phadia. Minimax estimation of a cumulative distribution function. *The Annals of Statistics*, 1 :1149–1157, 1973.

[77] R.-D. Reiss. Nonparametric estimation of smooth distribution functions. *Scandinavian Journal of Statistics*, 8 :116–119, 1981.

[78] E. Restle. *Estimating cumulative distributions by spline smoothing*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 2001.

[79] C. Robert. *L'Analyse Statistique Bayésienne*. Economica, Paris, 1992.

[80] M. Rosenblatt. Remarks on some non-parametric estimates of a density function. *The Annals of Mathematical Statistics*, 27 :832–837, 1956.

[81] G. Roussas. Nonparametric estimation of the transition distribution function of a Markov process. *The Annals of Mathematical Statistics*, 40 :1386–1400, 1969.

[82] M. Samanta. Non-parametric estimation of conditional quantiles. *Statistics and Probability Letters*, 7 :407–412, 1989.

[83] B. Schölkopf, C. Burges, and A. Smola. *Advances in Kernel methods. Support vector learning*. MIT Press, 1999.

[84] B. Schölkopf and A. Smola. *Learning With Kernels : Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002.

[85] S. Shirahata and I.-S. Chu. Integrated squared error of kernel-type estimator of distribution function. *Annals of the Institute of Statistical Mathematics*, 44 :579–591, 1992.

[86] G. Shorack and J. Wellner. *Empirical Processes with Applications to Statistics*. Wiley, New-York, 1986.

[87] R. S. Singh, T. Gasser, and B. Prasad. Nonparametric estimates of distributions functions. *Communication in Statistics - Theory and Methods*, 12 :2095–2108, 1983.

[88] N. Smirnov. Approximate laws of distribution of random variables from empirical data. *Uspekhi Matematicheskikh Nauk*, 10 :179–206, 1944.

[89] C. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5 :595–645, 1977.

[90] W. Stute. Asymptotic normality of nearest neighbor regression function estimates. *The Annals of Statistics*, 12 :917–926, 1984.

[91] W. Stute. Conditionnal empirical processes. *The Annals of Statistics*, 14 :638–647, 1986.

[92] V. Susarla and J. Van Ryzin. Empirical Bayes estimation of a distribution (survival) function from right censored observations. *The Annals of Statistics*, 6 :740–754, 1978.

[93] J. Swanepoel. Mean integrated squared error properties and optimal kernels when estimating a distribution function. *Communication in Statistics - Theory and Methods*, 17 :3785–379, 1988.

[94] V. Vapnik. *The nature of statistical learning theory*. Springer Verlag, New-York, 1995.

[95] V. Vapnik and S. Kotz. *Estimation on dependences based on empirical data*. Springer Verlag, New-York, 2006.

[96] G. Wahba. *Spline models for observational data*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1990.

[97] G. Watson and M. Leadbetter. Hazard analysis II. *Sankhya*, 26 :101–116, 1964.

[98] B. Winter. Strong uniform consistency of integrals of density estimators. *The Canadian Journal of Statistics*, 1 :247–253, 1973.

[99] B. Winter. Convergence rate of perturbed empirical distribution functions. *Journal of Applied Probability*, 16 :163–173, 1979.

[100] S. Wood. Monotonic smoothing splines fitted by cross validation. *SIAM Journal on Scientific Computing*, 15 :1126–1133, 1994.

[101] I. Wright and E. Wegman. Isotonic, convex and related splines. *The Annals of Statistics*, 8 :1023–1035, 1980.

[102] H. Yamato. Uniform convergence of an estimator of a distribution function. *Bulletin on Mathematical Statistics*, 15 :69–78, 1973.

[103] Q. Yu. Inadmissibility of the empirical distribution function in continuous invariant problems. *The Annals of Statistics*, 17 :1347–1359, 1989.

Conclusion et perspectives

Dans ce travail de thèse, nous nous sommes attachés à étudier l'indice de régularité d'une mesure de probabilité. Nous avons pu voir que cet indice intervient dans différents problèmes d'estimation fonctionnelle. Mais sa définition étant relativement récente, ses propriétés sont encore méconnues et il est probable qu'il apparaisse dans d'autres problèmes. Il se pourrait par exemple qu'il puisse expliquer certains problèmes liés à la vitesse de convergence de différents estimateurs.

Dans la première partie de ce travail, nous avons déterminé des conditions nécessaires et suffisantes pour l'obtention d'une loi asymptotique pour l'estimateur des k_n-plus proches voisins de la densité. Cependant, beaucoup d'estimateurs de la densité requièrent un développement pour un rapport de mesures d'ensembles non centrés au point d'estimation x et n'étant pas forcément des boules. Nous avons donc adapté la définition de l'indice de régularité. A l'aide de cette nouvelle définition, nous avons obtenu des conditions suffisantes pour la normalité asymptotique de l'estimateur des k_n-plus proches voisins ou pour l'histogramme. S'agissant de la recherche future, il serait intéressant de déterminer la loi asymptotique d'autres estimateurs de la densité, comme par exemple l'estimateur B.S.E. ou celui de Barron, en utilisant la nouvelle définition de l'indice de régularité.

Dans la deuxième partie, nous affaiblissons les hypothèses impliquant la convergence de l'estimateur du mode d'une densité quelconque f défini par Abraham, Biau et Cadre (2003) notamment l'hypothèse sur la continuité de la densité f autour du mode θ. Nous déterminons également un intervalle de confiance asymptotique dépendant de différentes constantes. Une pro-

chaine étape importante sera de définir des estimateurs convergents de ces constantes sous les hypothèses de notre théorème, plus explicitement sans utiliser la continuité autour du mode. Il serait également utile de réaliser une analyse de sensibilité de notre estimateur par rapport à la fenêtre de l'estimateur à noyau. De plus, il serait intéressant de l'appliquer à des données réelles présentant des discontinuités, par exemple dans le domaine de l'illumination ou de la spectrométrie.

Enfin, la dernière partie de cette thèse a permi de définir un nouvel estimateur convergent de l'indice de régularité utilisant l'estimateur empirique de la fonction de répartition. Cependant, les simulations sur cet estimateur, ainsi que sur un autre utilisant l'estimateur à noyau, ne donnent pas de meilleurs résultats que l'estimateur défini par Beirlant, Berlinet et Biau (2008). Nous avons également réalisé une revue bibliographique sur les différents estimateurs de la fonction de répartition ce qui ouvre un certain nombre de perspectives intéressantes. En particulier, un travail futur serait de déterminer une condition sur un estimateur de la fonction de répartition quelconque donnant un estimateur de l'indice de régularité convergent. Cela nous donnerait alors une large classe d'estimateurs de l'indice de régularité. Un certain nombre des estimateurs de la fonction de répartition ayant été implémenté, un package pourrait alors être créé et mis à la disposition de la communauté scientifique.

Résumé

L'objectif de cette thèse est d'étudier le comportement local d'une mesure de probabilité, notamment au travers d'un indice de régularité locale. Dans la première partie, nous établissons la normalité asymptotique de l'estimateur des k_n plus proches voisins de la densité et de l'histogramme. Dans la deuxième, nous définissons un estimateur du mode sous des hypothèses affaiblies. Nous montrons que l'indice de régularité intervient dans ces deux problèmes. Enfin, nous construisons dans une troisième partie différents estimateurs pour l'indice de régularité à partir d'estimateurs de la fonction de répartition, dont nous réalisons une revue bibliographique.

Mots-clefs : Indice de régularité locale, Mesure de probabilité, Estimation non paramétrique, Estimation du mode, Estimateurs de la fonction de répartition, Normalité asymptotique, Estimateur des k_n plus proches voisins de la densité.

Abstract

The goal of this thesis is to study the local behavior of a probability measure, using a local regularity index. In first part, we establish the asymptotic normality of the nearest neighbor density estimate and of the histogram. In the second one, we define a mode estimator under weakened hypothesis. We show that the regularity index interferes in this two problems. Finally, we construct in third part various estimators of the regularity index. The thesis ends with a review on distribution function estimators.

Keywords : Local regularity index, Probability measure, Nonparametric estimation, Mode estimators, Distribution function estimators, Asymptotic normality, Nearest neighbor estimate.