

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche
Scientifique

Université A.Mira de Béjaïa
Faculté des Sciences Exactes
Département de Mathématiques



Mémoire

En vue de l'Obtention du Diplôme de

Master en Mathématiques

Option : Probabilités Statistique et Applications

Thème

*Estimation non paramétrique de la fonction de densité
dans le cadre quasi-associé*

Présenté par :
DRIZI Hanane

Devant le jury composé de :

M ^{me} Karima. TIMRIDJINE	Présidente	M.C.A	U. A. Mira. Béjaïa
M ^r Mohand. BOURAINE	Examineur	M.A.A	U. A. Mira. Béjaïa
M ^{me} Hadjila. TABTI	Promotrice	M.A.B	U. A. Mira. Béjaïa

Béjaïa, 24 juin 2018

Dédicaces

Je dédie ce modeste travail :

A mes très chers parents (Amirouche et Dahbia) pour leur foi en moi, leur encouragement et amour qui m'a poussé vers l'avant, des bougies qui s'enflamment pour m'éclaircir le chemin de la vie.

A mes très chers soeurs Leila, Ouardia, Hakima et Samer les personnes qui étaient toujours à coté de moi pour m'encourager et me soutenir, je vous adore.

A mon très cher frère Rabah, je te souhaite un avenir radieux et plein de réussite.

A mon cher cousin Hamid, je te souhaite une vie pleine de prospérités.

A tous mes chers neveux et nièces :

Yacine, Layane, Nessma, Zayneb .

A ceux qui m'ont créé un milieu d'ambiance et de travail, mes chers ami(e)s :

Aziza, Katia, Lynda, Djamila, Abla, Hanane, Nadia, Massi, Ramzi.

Je le dédié aussi à tous ceux qui m'ont donné le moindre coup de pouce pour le réalisé .

A tous merci.

Hanane

Remerciements

Avant tous, louange à Allah, dieu le tout puissant qui m'a accordé le courage, la patience et la volonté afin de parvenir à la finalité de ce modeste travail.

J'adresse tout d'abord mes remerciements les plus sincères à ma promotrice **M^{me}. TABTI Hadjila** pour son aide précieux, sa disponibilité et grâce ces connaissances et son expérience, j'ai pu mener à bien l'élaboration de ce travail.

Mes remerciements vont également aux membres de jury La présidente **M^{me}. TIMRIDJINE Karima** et l'examineur **M^r. BORAINÉ Mohand** pour avoir accepter de juger et évaluer mon travail.

Mes vifs remerciements pour ma famille, surtout mes parents qui m'ont épaulés, soutenus et suivis tout au long de ce projet.

Je remercie mes amies pour leurs aide et conseils qu'elles m'ont donné tout au long de mon travail.

Sans oublier tous ceux qui ont contribué de près ou de loin à l'élaboration de ce mémoire.

Table des matières		ii
Listes des abréviations		iii
Introduction générale		1
1 Variables aléatoires associées		4
1.1 Généralités sur l'association		4
1.2 Association positive et négative		5
1.2.1 Association positive		5
1.2.2 Association négative		7
1.3 Quasi association		9
2 Estimation de la fonction de densité des variables aléatoires quasi associées		10
2.1 L'estimation non paramétrique		10
2.2 Estimation non paramétrique de la fonction de densité f		11
2.3 L'estimation à noyaux		12
2.3.1 Le cas unidimensionnel		12
2.3.2 Le cas multivarié		12
2.3.3 Exemples de noyaux réels		13
2.4 La convergence presque complète		15
2.4.1 Inégalités exponentielles		16
2.5 Convergence presque complète de l'estimateur de f dans le cas indépendant		18
2.6 Estimation de la fonction densité dans le cas quasi- associée		23

TABLE DES MATIÈRES

2.7	convergence presque complète dans le cas quasi-associée	24
3	Simulation	28
3.1	La simulation	28
	Conclusion générale	34
	Bibliographie	35

Notations et abréviations

FKG : L'inégalité **F**ortuin, **K**asteleyn et **G**inibre.

$\mathbb{E}(X)$: L'espérance mathématique de la variable aléatoire X .

$Var(X)$: La variance de la variable aléatoire X .

$Cov(X, Y)$: La covariance entre les variables aléatoires X et Y .

$a \wedge b$ Le minimum de a et b .

$a \vee b$ Le maximum de a et b .

F_n La fonction de répartition.

$f(x)$ La fonction de densité.

$\hat{f}_n(x)$ La fonction de densité estimée par la méthode de noyau.

K Le noyau.

K_0 Noyau symétrique standard.

h paramètre de lissage ou largeur de fenêtre h_n .

p.co convergence presque complète.

INTRODUCTION GÉNÉRALE

La notion d'indépendance pour les collection de variables aléatoires est un concept qui possède des propriétés intéressantes, les démonstrations sous l'hypothèse d'indépendance deviennent moins difficiles en les comparant au cas de la dépendance. Cette notion ne trouve pas beaucoup d'application en pratique à vrai dire, si l'on regarde profondément les choses, on s'apercevra que dans la majorité des cas, on ne peut pas échapper à la dépendances. Ainsi les phénomènes étudiés dans la physique, la chimie, la biologie, l'économie et la fiabilité étaient des sources principales dont l'émergence des modèles stochastiques et les modèles de variables aléatoires dépendantes.

Dans la littérature, il convient de modéliser la dépendance entre les variables aléatoires, deux types de dépendance sont largement utilisés les mélanges et l'association. La notion de mélange réfère plus à la σ - algèbre qu'au variables aléatoires. Ainsi l'avantage principal de l'association est que la plupart de ses propriétés sont déterminés par les covariances.

Dans ce travail, on s'est intéressé à la notion d'association de variables aléatoires dans \mathbb{R}^n , cette notion de dépendance a été introduite dans les années 1960 par Harris pour des processus de percolation, puis par Lahmen (1966) [16] en introduisant la notion de dépendance positive par quadrants entre deux variables aléatoires, par suite par Esary et Prochan et Walkup [13] ont généralisés cette notion et ils ont introduit la notion d'association.

Au début, cette notion a reçu peu d'attention par la communauté probabiliste et statistique, mais ces dernières années l'intérêt a augmenté dû à leur applicabilités dans différentes sciences de l'ingénieur. Elle trouve beaucoup d'application en divers domaines scientifiques et industriels à savoir la mécanique statistique, la théorie de fiabilité, les mathématiques statistiques, la théorie de la percolation, etc.

Divers propriétés asymptotiques pour de sommes de variables aléatoires associées ont été étudiées par Newmann (1980,1984) et Birkel (1988) [5] [23] et plusieurs autres auteurs. Ils ont observé que dans toute propriété asymptotique des variables aléatoires la structure de covariance joue un rôle fondamentale. En 1991 Roussas a établi la convergence ponctuelle sous des conditions de régularité.

Esary, Proschan et Walkup (1967), et Frtuyn Kastelyn et Ginibre (FKG, 1971) [14] ont utilisé cette notion et leur objectif était de trouver des applications dans la fiabilité des systèmes, les modèles de ferromagnétisme. Rappelons que dans les modèles de ferromagnétisme une signification de l'inégalité est la suivante : deux électrons voisins ont une probabilité plus élevé d'être orienté de la même manière que dans des sens opposés, autrement dit l'interaction entre les électrons est attractive.

Aussi, Newman [18] a démontré le théorème central limite pour les champs aléatoires satisfait l'inégalité (FKG) d'une part. D'autre part, Cox et Grimmett [9] ont donné des exemples des champs aléatoires associées pris de théorèmes de la percolation. Des rappels sur l'association (positives et

négatives) et ces applications en statistique peuvent être trouvés dans Roussas (1999, 2000, 2001). Des résultats importants sur le théorème limite pour des données associées. Le livre de de Barlow et Prochan [3] indique clairement l'intérêt de l'association dans ce domaine. En particulier négativement associées ont été obtenus par Borogna et al.

Bulinski (1996) [7] ainsi Doukhan et Louhichi (1999) ont établi une inégalité pour des données négativement associées et gaussienne . [1]

En 2001 il a introduit avec Sequet un nouveau concept de dépendance appelé "quasi association" pour étudier certain champs , il permet d'étudier une classe de variables aléatoires, indépendantes par la non corrélation.

Un des problèmes les plus rencontrés en statistique est l'estimation non paramétrique sous la dépendance qui est le cadre classique à la statistique, elle connaît un grand essor chez de nombreux auteurs et dans différents domaines. En effet celle-ci possède un champs d'application très large permettant ainsi l'explication de certains phénomènes mal modélisés jusqu'à présent tel que les séries chronologiques.

Elle a reçu un intérêt croissant tant sur le plan théorique que pratique. Cette branche de la statistique ne se résume pas à l'estimation d'un nombre fini de paramètres réels associés à la loi de l'échantillon (comme le cas de la théorie de l'estimation paramétrique), elle consiste généralement à estimer à partir des observations une fonction inconnue, élément d'une certaine classe fonctionnelle.

Parmi ces problèmes de l'estimation non paramétrique sous les données dépendantes, on trouve l'estimation non paramétrique de la fonction densité dans le cas uni, multidimensionnelle et fonctionnelle. Plusieurs auteurs ont été intéressé à l'étude des propriétés de l'estimateur à noyau de la densité des variables associées citons par exemple Cai et Roussas (1998, 1999), Masery (2002). Les principaux résultats portent sur la convergence presque sûre uniforme de l'estimateur à noyau de la fonction de densité pour des variables quasi-associées à valeurs dans \mathbb{R}^d sous des conditions de décroissance arithmétique ou géométrique des covariances et sous certaines conditions sur la régularité de la fonction de la fonction de densité. Begai et Prakasa Rao (1995) [4] ont étudié l'estimation de la densité pour un processus associé stationnaire et plusieurs autres travaux sont largement utilisé citons quelques uns : Parakasa Rao (1983), Silverman (1986), Roussas (1990), Scott(1992), Bosq et Lecontre(1987). [10] Depuis les travaux de Rosenblatt (1956) et Parzen (1962) puis Nadaraya-Watson (1964) portant respectivement sur les estimateurs non paramétrique des fonctions de densités et régression, la méthode de noyaux est largement utilisé dont des nombreux travaux on citons Begai et Prakasa Rao (1995), ils ont étudié l'estiamtion de la densité pour un processus associé stationnaire, Silverman (1986), Scott (1992), Bosq (1987).

On s'intéresse dans ce travail à l'étude de l'estimation non paramétrique à noyau de la fonction

de densité en introduisant la convergence presque complète et quelques inégalités exponentielles qui permettent de contrôler le comportement limite des déviations des estimateurs par rapport à leur espérances. Ce mémoire comporte trois chapitres, une conclusion et une bibliographie.

Dans le premier chapitre, nous étudions les variables aléatoires associées. Nous rappelons la définition de l'association des variables aléatoires, en donnant quelques théorèmes, propriétés et des exemples.

Dans le deuxième chapitre, nous présentons la notions de l'estimation de la fonction densité par la méthode de noyau, et l'étude de la convergence presque complète.

Dans le dernier chapitre, nous proposons une étude numérique à l'aide de logiciel R pour valider les résultats théoriques obtenus.

Introduction

L'association et quelques autres notions de dépendance ont été introduites dans les années 1960. Ces dernières années le concept d'association trouve beaucoup d'application en divers domaines, les phénomènes étudiés dans la physique, la chimie, la biologie, l'économie et la fiabilité étaient des sources principales pour ce modèle. Ce qui a fait que le contrôle de la dépendance entre variables aléatoires a toujours été un sujet d'intérêt et de préoccupation pour les probabilités et les statisticiens. [25] l'objet de ce chapitre est de rappeler quelques notions de l'association, et donner quelques théorèmes et propriétés puis exhiber des exemples concernant ces variables dépendantes.

1.1 Généralités sur l'association

Le concept d'association positive ou d'association tout simplement à été introduit par Esary et al (1967) [8] et Fortelyn et Ginibre (FKG 1971) [22]. Leur objectif était de trouver des applications dans la fiabilité des systèmes et en Statistique en se basant sur l'inégalité FKG. La définition de l'association est une généralisation de la définition de la dépendance positive introduite par Lehmann (1966)[20] : Un vecteur (X, Y) de variables aléatoires réelles x et y est dit positivement dépendant si, pour tout réels x et y ,

$$P(X \geq x, Y \geq y) - P(X \geq x)P(Y \geq y) \geq 0$$

ou d'une manière équivalent si

$$\text{cov}(f(X), g(Y)) \geq 0$$

pour toutes fonctions croissantes f et g .

Des variables aléatoires X_1, X_2, \dots, X_n sont dites positivement associées ou associées si

$$\text{cov}(f(X_1, \dots, X_n), g(X_1, \dots, X_n)) \geq 0$$

Ces variables sont dites négativement associées si pour tous sous-ensembles disjoints A_1 et A_2 de $\{1, 2, \dots, n\}$

$$\text{cov}(f(X_i, i \in A_1), g(X_j, j \in A_2)) \leq 0$$

1.2 Association positive et négative

1.2.1 Association positive

Burton, Dabrowski et Dehling (1986) définissent une classe strictement plus large de variables aléatoires incluant l'association : l'association positive ou faible association . [15]

Définition 1.2.1

Soit I un ensemble fini, on dit que la suite de variables aléatoires $(X_i, i \in I)$ est positivement associée ou tout simplement associée si, pour toutes fonctions f et g croissantes définies sur $\mathbb{R}^{|I|}$

$$\text{cov}(f(X_i, i \in I), g(X_j, j \in I)) \geq 0$$

Lorsque cette covariance existe.

Dans le but d'affaiblir le concept d'association Burton Dabrowski et Dehling (BDD) (1986) ont introduit la notion d'association faible. Ils ont formulé ce concept pour des suites de variables aléatoires à valeurs dans \mathbb{R}^d . [12]

Définition 1.2.2

Soit I un ensemble fini, on dit que la suite de variables aléatoires $(X_i)_{i \in I}$ est faiblement associée si pour tous sous-ensembles disjoints A et B de I et toutes fonctions f et g croissantes définies sur $\mathbb{R}^{|A|}$ et $\mathbb{R}^{|B|}$, [12]

$$\text{cov}(f(X_i, i \in A), g(X_j, j \in B)) \geq 0$$

Lorsque cette covariance existe.

Dans le cas où l'ensemble I est infini, la suite $(X_i)_{i \in I}$ est dites associées (respectivement faiblement associées) si, pour tous sous-ensemble fini J de I $(X_i)_{i \in J}$ est associées (respectivement faiblement associées).

Caractéristiques et propriétés

On présente les propriétés caractéristiques les plus importantes présentées par Esary et al(1967) des variables associées, ce qui va nous permettre de construire des exemples de celles-ci. Pour la démonstration de ces propriétés, (voir [7] [8])

1. Tous sous ensemble d'un ensemble fini de variables aléatoires réelles associées est encore associé.
2. L'union de deux sous-ensembles indépendants de variables aléatoires associées est associées.
3. Tout singleton formé d'une variable aléatoire réelle X est associé.
4. si (X_1, \dots, X_n) est un vecteurs de variables associées et f_1, \dots, f_m des fonctions réelles croissantes définies sur \mathbb{R}^n alors le vecteur $(f_1(X), \dots, f_m(X))$ est associé.
5. Si pour tout $k \geq 1$, $(X_1^{(k)}, \dots, X_n^{(k)})$ est un vecteur associé et si $X_i^{(k)}$ converge en loi vers X_i pour tout $i = 1, \dots, n$ alors (X_1, \dots, X_n) est associé

Exemples de variables aléatoires associée [12] [11]

a- Statistique d'ordre

Si $X = (X_1, \dots, X_n)$ est un vecteur associé, alors le vecteur (X_{n1}, \dots, X_{nn}) de la statistique d'ordre engendré par X est aussi associé.

b- Processus linéaire

Soit $(\varepsilon_i)_{i \in \mathbb{Z}}$ une suite de variables aléatoires indépendantes ou associées et $(a_i)_{i \in \mathbb{Z}}$ une suite de réels. Pour tout $n \in \mathbb{Z}$ et $N \geq 1$, on pose $X_{n,N} = \sum_{|i| \leq N} a_i \varepsilon_{n-i}$. Supposons qu'il existe une variable aléatoire X_n telle que

$$\lim_{N \rightarrow \infty} X_{n,N} = X_n \quad \text{p.s.} \quad |X_n| < +\infty \quad \text{p.s.} \quad \forall n \in \mathbb{Z}$$

$(X_n)_{n \in \mathbb{Z}}$ est un processus linéaire défini pour tout $n \in \mathbb{Z}$ par

$$X_n = \sum_{i \in \mathbb{Z}} a_i \varepsilon_{n-i}$$

Si les termes de la suite $(a_i)_{i \in \mathbb{Z}}$ sont positifs, alors le processus linéaire $(X_n)_{n \in \mathbb{Z}}$ est associé. En effet, pour tout $N \geq 1$, la suite $(X_{n,N})_{n \in \mathbb{Z}}$ est associée. Par conséquent, l'association de la suite $(X_n)_{n \in \mathbb{Z}}$.

c- Processus autorégressif d'ordre p

Soient $f : \mathbb{R}^p \rightarrow \mathbb{R}$ et $(\varepsilon_n)_{n \in \mathbb{Z}}$ une suite de variables aléatoires indépendantes. On considère le processus autorégressif $(X_n)_{n \in \mathbb{Z}}$ défini pour tout $n \in \mathbb{Z}$ par

$$X_n = f(X_{n-1}, \dots, X_{n-p}) + \varepsilon_n.$$

On suppose que le vecteur (X_0, \dots, X_{1-p}) est associé et indépendant de la suite $(\varepsilon_n)_{n \in \mathbb{Z}}$. Si f est une fonction croissante sur \mathbb{R} , alors la suite $f(X_n)_{n \geq p}$ est associée.

En effet, $f(X_n)$ est une fonction croissante des variables aléatoires associées .

d- Variables aléatoires binaires

Un vecteur aléatoire $X = (X_1, X_2)$ de variables binaires (qui prennent les valeurs 0 ou 1) est associé si et seulement si sa covariance $Cov(X_1, X_2) \geq 0$.

e- Processus gaussien

Tout vecteur gaussien (X_1, \dots, X_n) est associé si et seulement si $Cov(X_i, X_j) \geq 0$.

1.2.2 Association négative

La dépendance négative a été introduite par Alam et Saxena (1981) et développée par Joag-Dev et Prochan (1980) [2], ils ont donné de nombreuses propriétés et proposé plusieurs applications en Statistique. Joag-Dev et Prochan montrent, en plus des propriétés fondamentales des variables négativement associées, que certaines distributions multivariées (les lois gaussiennes négativement corrélées, les lois multinomiales, les lois de Dirichlet,...) possèdent la propriété de NA. Ils ont introduit la notion d'association négative, cette notion a un avantage par rapport aux autres types connus de dépendance négative, basé sur le fait que toute suite de fonctions croissantes de variables aléatoires négativement associées est négativement associées [11] .

Définition 1.3.1

Soit I un ensemble fini, on dit que la suite de variables aléatoires $(X_i, i \in I)$ est négativement associée si, pour tous sous-ensembles disjoints A et B de I est toutes fonctions f et g croissantes définies sur \mathbb{R}^A et \mathbb{R}^B , [12]

$$\text{cov}(f(X_i, i \in A), g(X_j, j \in B)) \leq 0$$

Lorsque cette covariance existe.

Dans le cas où l'ensemble I est infini, la suite $(X_i, i \in I)$ est dite négativement associées si, pour tout sous-ensemble fini J de I , $(X_i, i \in J)$ est négativement associée.

Exemples de variables aléatoires négativement associées [15]

a- Distributions multivariées ayant la propriété d'association négative

- Si (X_1, \dots, X_n) est un vecteur de variables aléatoires suivant une loi multinomiale, une loi hypergéométrique multivariée ou une loi de Dirichlet, alors (X_1, \dots, X_n) est négativement associé.
- Soit $X = (X_1, \dots, X_k)$ une suite de k variables aléatoires réelles. La distribution jointe du vecteur (X_1, \dots, X_k) est appelée permutation si (X_1, \dots, X_k) , elle prend comme valeurs les permutations de X avec une probabilité $1/k!, k > 1$.

Si la distribution d'un vecteur aléatoire (X_1, \dots, X_k) est une permutation alors (X_1, \dots, X_k) est négativement associé.

b- Processus gaussien

Un vecteur gaussien (X_1, \dots, X_n) est négativement associé si et seulement si

$$\text{Cov}(X_i, X_j) \leq 0 \text{ pour tous } i, j \in [1, n].$$

1.3 Quasi association

C'est la plus récente des formes d'association, dûe à Bulinski et Sabanovitch (1998). EN (2001) ils ont introduit ce concept pour étudier certains champs aléatoires. Ce concept permet d'étudier une classe de variables aléatoires indépendantes par la non corrélation. Cette classe contient en plus des variables gaussiennes, des variables associées et négativement associées.[7]

Définition 1.4.1 [15]

Soit $X = \{X_n, n \in \mathbb{N}\}$ une famille de variables aléatoires associées à valeurs dans \mathbb{R}^d , soit I et J deux sous-ensembles finis disjoints de \mathbb{N} alors la famille X est dite quasi-associée si pour toutes fonctions lipschitziennes f et g définies sur $\mathbb{R}^{|I|d}$ et $\mathbb{R}^{|J|d}$, on a

$$|\text{cov}(g(X_i, i \in I), f(X_j, j \in J))| \leq \text{Lip}_i(g) \text{Lip}_j(f) \sum_{i \in I} \sum_{j \in J} \sum_{k=1}^d \sum_{l=1}^d |\text{cov}(X_i^k, X_j^l)|$$

Où la constante de Lipschitz $\text{Lip}_i(g)$ est telle que, pour tout $x = (x_i, i \in I)$, $y = (y_i, i \in I)$ dans $\mathbb{R}^{|I|d}$,

$$|g(x) - g(y)| \leq \sum_{i \in I} \text{Lip}_i(g) |x_i - y_i|$$

avec

$$\text{Lip}_i(g) = \sup_{x_i \neq y_i} \frac{|f(x), f(y)|}{\|x_i - y_i\|}$$

Le sup étant pris pour $x_1, x_2, \dots, x_{|I|}, y_i \in \mathbb{R}$ avec $\|x_1, \dots, x_u\| = |x_1| + \dots + |x_u|$.

CHAPITRE 2

ESTIMATION DE LA FONCTION DE DENSITÉ DES VARIABLES ALÉATOIRES QUASI ASSOCIÉES

Introduction

La théorie de l'estimation est une des préoccupations majeures des statisticiens. Ainsi l'estimation non paramétrique sous des données associées est largement étudiée dans la littérature dans le cas uni et multidimensionnel. Les premières études pour des données associées ont été faites au début des années soixante par Harris pour des processus de percolation, puis par Lahmen (1966) pour des données dépendantes [20]. Les estimations non paramétriques de la fonction de densité par la méthode du noyau ont été largement utilisées dans de nombreux travaux. Dans ce chapitre on étudiera l'estimation de la fonction de densité par la méthode de noyau et la convergence presque complète de cet estimateur dans les deux cas indépendant et dépendant [1].

2.1 L'estimation non paramétrique

L'estimation est un élément fondamental de la statistique. Elle permet de généraliser, autant que faire se peut. On y distingue l'approche paramétrique, qui considère que les modèles sont connus, avec des paramètres inconnus, et l'approche non paramétrique dont on s'intéresse dans notre étude. Cette approche, qui ne fait aucune hypothèse sur la loi, ni sur ses paramètres. Nos connaissances sur le modèle ne sont pas précises, ce qui est souvent le cas dans la pratique. Dans cette situation, il est naturel de vouloir estimer une des fonctions décrivant le modèle, soit généralement la fonction de répartition ou la densité [2].

En effet, depuis les travaux fondateurs de Rosenblatt (1956) et Parzen (1962) [1] qui ont, en généralisant la notion d'estimation par histogramme (estimateur naïf), donné naissance à la méthode du noyau .

2.2 Estimation non paramétrique de la fonction de densité f

Supposons que nous observons n variables aléatoires indépendantes identiquement distribuées. Soit une suite de variables aléatoires X_1, X_2, \dots, X_n de densité de probabilité inconnue f et soit $F(x)$ la fonction de répartition de la loi de X . La fonction de répartition empirique est estimée par : [6]

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{(X_i < x)} \quad (2.1)$$

La loi forte des grand nombre permet de donner l'estimateur de F telle que

$$\forall x \in \mathbb{R}, \hat{F}_n \longrightarrow F \quad \text{si } n \longrightarrow \infty$$

.

Rosenblatt (1956) est le premier qui a donné un exemple d'estimateur à partir de \hat{F}_n pour $h > 0$ est petit .

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{X_i \leq x\}}, \forall x \in \mathbb{R}. \quad (2.2)$$

Cet estimateur peut encore s'écrire sous la forme :

$$\begin{aligned} \hat{f}_n(x) &= \sum_{i=1}^n \frac{\mathbf{1}_{\{-h \leq X_i - x \leq h\}}}{2nh} \\ &= \frac{1}{2nh} \sum_{i=1}^n \mathbf{1}_{\{-1 < \frac{X_i - x}{h} \leq 1\}} \\ &= \frac{1}{2nh} \sum_{i=1}^n K_0\left(\frac{X_i - x}{h}\right) \end{aligned}$$

Avec

$$K_0(u) = \frac{1}{2} \mathbf{1}_{\{-1 < u < 1\}}.$$

2.3 L'estimation à noyaux

L'origine de la méthode des noyaux est due à Rosenblatt (1956) [25]. Celui-ci a proposé une sorte d'histogramme mobile où la fenêtre de comptage des observations se déplace avec la valeur de x . La densité en x est estimée par la fréquence relative des observations dans l'intervalle $[x - h, x + h]$, donc centré sur x , divisée naturellement par la largeur de l'intervalle. On appelle h la largeur de fenêtre. Pour des raisons qui apparaîtront plus loin nous écrivons l'estimation ainsi obtenue à partir des observations x_1, x_2, \dots, x_n . On distingue deux cas :

2.3.1 Le cas unidimensionnel

Définition 2.3.1

On appelle noyau $K : \mathbb{R} \rightarrow \mathbb{R}$ une fonction mesurable. K est une fonction bornée, intégrable appelé noyau. On appelle $h_n > 0$ la fenêtre ou largeur de fenêtre qui contrôle le lissage de la courbe estimé et \hat{f}_n l'estimateur à noyau de f , défini pour tout $x \in \mathbb{R}$ par

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{X_i - x}{h_n}\right). \quad (2.3)$$

L'estimation à noyau est alors une densité si K est une densité quelles que soient les valeurs des observations X_1, \dots, X_n . Un noyau est dit symétrique si, pour tout u dans son ensemble de définition, $K(u) = K(-u)$.

2.3.2 Le cas multivarié

Dans le cas multivarié on considère n variables aléatoires à valeurs X_1, X_2, \dots, X_n dans \mathbb{R}^d , le noyau donné précédemment peut être étendue facilement à cette situation. Pour cela, en considérant un noyau multivarié K^* qui sera une fonction dans \mathbb{R}^d dans \mathbb{R} . La première façon de le faire est de définir K^* comme un produit de d fonctions noyaux K_1, K_2, \dots, K_d réelles telles que

$$\forall u = (u_1, \dots, u_d)^t \in \mathbb{R}^d, K^*(u) = K_1(u_1) * K_2(u_2) * \dots * K_d(u_d)$$

. Une seconde façon consiste à combiner un noyau réelle de fonction f avec une norme dans \mathbb{R}^d définit comme suit :

$$\forall u \in \mathbb{R}^d : K(u) = K(\|u\|)$$

Dans ce cas, la construction de cet estimateur est présentée comme suit :

Considérons un ensemble de données de dimension d avec la taille de l'échantillon n , $X_i = (X_{i1}, \dots, X_{id})$, $i = 1, \dots, n$, alors l'estimation à noyau de la fonction de densité est :

$$\begin{aligned}\hat{f}_n(x) &= \frac{1}{n} \sum_{i=1}^n \frac{1}{h^d} K^* \left(\frac{x - X_i}{h} \right) \\ &= \frac{1}{n} \sum_{i=1}^n \frac{1}{h^d} K^* \left(\frac{x_1 - X_{i1}}{h}, \dots, \frac{x_d - X_{id}}{h} \right)\end{aligned}$$

où

K est une fonction de noyau multivariée avec d arguments et h est le même pour chaque composant. si h_i sont différents c'est à dire $h = (h_1, \dots, h_d)^t$ alors

$$\begin{aligned}\hat{f}_n(x) &= \frac{1}{n} \sum_{i=1}^n \frac{1}{h_1 \dots h_d} K \left(\frac{x_1 - X_{i1}}{h_1}, \dots, \frac{x_d - X_{id}}{h_d} \right) \\ &= \frac{1}{n} \sum_{i=1}^n \left\{ \prod_{j=1}^d \frac{1}{h_j} K \left(\frac{x_j - X_{ij}}{h_j} \right) \right\}\end{aligned}$$

2.3.3 Exemples de noyaux réels

- Noyau rectangulaire :

$$K_1(x) = \begin{cases} \frac{1}{2}, & \text{si } |x| \leq 1; \\ 0, & \text{si } |x| > 1. \end{cases}$$

- Noyau triangulaire :

$$K_2(x) = \begin{cases} 1 - |x|, & \text{si } |x| \leq 1; \\ 0, & \text{si } |x| > 1. \end{cases}$$

- Noyau d'Epanechnikov ou parabolique :

$$K_3(x) = \begin{cases} \frac{3}{4}(1 - x^2), & \text{si } x \in [-1, 1]; \\ 0, & \text{sinon.} \end{cases}$$

- Noyau quadratique :

$$K_4(x) = \begin{cases} \frac{15}{16}(1 - x^2)^2, & \text{si } x \in [-1, 1]; \\ 0, & \text{sinon.} \end{cases}$$

- Noyau cubique :

$$K_5(x) = \begin{cases} \frac{35}{32}(1-x^2)^3, & \text{si } x \in [-1, 1]; \\ 0, & \text{sinon.} \end{cases}$$

- Noyau gaussien :

$$K_6(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right), x \in \mathbb{R}.$$

- Noyau sinus :

$$K_7(x) = \begin{cases} \frac{1}{2\pi} \left(\frac{\sin(\frac{x}{2})}{\frac{x}{2}}\right)^2, & \text{si } x \neq 0; \\ \frac{1}{2\pi}, & \text{si } x = 0. \end{cases}$$

- Noyau cosinus :

$$K_8(x) = \begin{cases} \frac{\pi}{4} \cos\left(\frac{\pi x}{2}\right), & \text{si } -1 \leq x \leq 1; \\ 0, & \text{sinon.} \end{cases}$$

- Noyau de Silverman :

$$K_9(x) = \frac{1}{2} \exp(-|x|/\sqrt{2}) \sin(|x|/\sqrt{2} + \frac{\pi}{4}), x \in \mathbb{R}$$

Selon la définition 2.1.1, toute fonction K peut servir comme noyau d'estimation d'une densité f .

Théorème 2.3.1

Si K est positive et $\int_{\mathbb{R}} K(u)du = 1$, alors $\hat{f}_n(x)$ est une densité de probabilité. De plus, \hat{f}_n est continue si K est continue [21].

Démonstration du théorème 2.3.1

L'estimateur à noyau est positive et continue car la somme des fonctions positives et continues est elle-même une fonction positive et continue. Il faut donc vérifier que l'intégrale de $\hat{f}_n(x)$ vaut un.

En effet,

$$\begin{aligned} \int_{\mathbb{R}} \hat{f}_n(x) dx &= \int_{\mathbb{R}} \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) dx \\ &= \frac{1}{nh} \sum_{i=1}^n \int_{\mathbb{R}} K\left(\frac{X_i - x}{h}\right) dx, \quad \left(u = \frac{X_i - x}{h}\right) \\ &= \frac{1}{nh} \sum_{i=1}^n \int_{\mathbb{R}} K(u) h du \\ &= 1 \end{aligned}$$

On voit donc que l'estimateur à noyau est une densité de probabilité.

2.4 La convergence presque complète

Le concept de convergence presque complète a été introduit par Hsu et Robbins (1947). Elle implique la convergence presque sûre et se prête bien aux calculs faisant intervenir des sommes de variables aléatoires. Malgré cela, elle ne commence à devenir populaire dans la communauté statistique que dans les années 1980 après les travaux de Collomb. Elle est utilisée surtout en statistique non-paramétrique.

Définition 2.4.1

On dit que la suite de variables aléatoires réelles $(X_n)_{n \in \mathbb{N}}$ converge presque complètement vers une variable aléatoire X lorsque $n \rightarrow \infty$ (et on note $\lim_{n \rightarrow \infty} X_n = X$ p.co), si et seulement si :

$$\forall \epsilon > 0, \sum_{n \in \mathbb{N}} \mathbb{P}[|X_n - X| > \epsilon] < \infty \quad (2.4)$$

Définition 2.4.2

On dit que la vitesse de convergence presque complète de la suite de variables aléatoires réelles $(X_n)_{n \in \mathbb{N}}$ vers X est d'ordre (U_n) ((U_n) étant une suite numérique déterministe), et on note $X_n = O(U_n)$ p.co, si et seulement si :

$$\exists \epsilon_0 > 0, \sum_{n \in \mathbb{N}} \mathbb{P}[|X_n - X| > \epsilon_0 U_n] < \infty$$

Notons que la convergence presque complète entraîne à la fois la convergence presque sûre et la convergence en probabilité.

Proposition 2.4.1

Si $\lim_{n \rightarrow \infty} X_n = X$ p.co, alors X_n converge en probabilité et presque sûrement vers X .

Preuve de proposition 2.4.1

La convergence en probabilité se déduit facilement de la convergence de la série (2.1)

($\mathbb{P}[|X_n - X| > \epsilon_0 U_n] < \infty$) est le terme général d'une série convergente implique que [13]

$$\forall \epsilon > 0, \mathbb{P}\left(\limsup_{n \rightarrow \infty} |X_n - X| > \epsilon\right) = 0.$$

De plus, $\lim_{n \rightarrow \infty} X_n(w) \neq X(x)$ implique l'existence de $\epsilon > 0$, tel que

$$\lim_{n \rightarrow \infty} \sup |X_n(w) - X(w)| > \epsilon.$$

on a alors $\mathbb{P}(\lim_{n \rightarrow \infty} X_n = X) = 1$ c'est à dire $X_n \rightarrow X$ P.s

2.4.1 Inégalités exponentielles

Nous allons donner deux versions des inégalités exponentielles de type Bernstein qui nous seront utiles pour l'établissement des résultats que nous avons choisi de reprendre.

Nous supposons que X_1, X_2, \dots, X_n est une suite de variables aléatoires réelles, indépendantes et centrées. Pour démontrer la convergence presque complète, nous avons besoin de trouver des bornes supérieures pour certaines probabilités concernant des sommes de variables aléatoires.

Proposition 2.4.1[26]

On a l'inégalité de Crammer suivante :

Si

$$\forall m \geq 2, |\mathbb{E}(X_i^m)| \leq \left(\frac{m!}{2}\right) (a_i)^2 b^{m-2},$$

Alors,

$$\forall \epsilon \geq 0, \mathbb{P} \left[\sum_{i=1}^n |X_i| > \epsilon A_n \right] \leq 2 \exp \frac{-\epsilon^2}{2(1 + \frac{b\epsilon}{A_n})}$$

où

$(a_i)_{1 \leq i \leq n}$ sont des réels positifs, $b \in \mathbb{R}^+$ et $A_n^2 = a_1^2 + a_2^2 + \dots + a_n^2$.

Corollaire 2.4.1[26]

a) Si pour tout $m \geq 2$, il existe un réel C_m strictement positif et une constante a positive, tels que :

$$\mathbb{E}|X_1^m| \leq C_m a^{2(m-1)},$$

Alors on a

$$\forall \epsilon > 0, \mathbb{P} \left[\left| \sum_{i=1}^n X_i \right| > \epsilon n \right] \leq 2 \exp \frac{-\epsilon^2 n}{2a^2(1 + \epsilon)}.$$

b) Supposons que les $(X_i)_{1 \leq i \leq n}$ dépendent de $n(X_i = X_{i,n})$.

Si pour tout $m \geq 2$, il existe un réel C_m strictement positif et une suite (a_n) de réels positifs, tels que :

$$\mathbb{E}|X_1^m| \leq C_m a_n^{2(m-1)},$$

et si

$U_n = n^{-1}a_n^2 \log n$, vérifie $\lim_{n \rightarrow \infty} U_n = 0$, alors on a

$$\frac{1}{n} \sum_{i=1}^n X_i = O\left(\sqrt{U_n}\right) a.co$$

Tandis que ce résultat s'applique à des variables dont on a majoré les moments d'ordre m , le corollaire suivant est donné pour des variables identiquement distribuées et bornées.

Démonstration du Corollaire 2.4.1 [22]

- a) En remplaçant $b = a^2$ et $A_n = a\sqrt{n}$ dans la proposition précédente, on aboutit à a) .
- b) En posant $\epsilon = \epsilon_0\sqrt{U_n}$ dans a) et comme U_n tend vers zéro , pour une certaine constante C' on a :

$$\begin{aligned} \mathbb{P} \left[\frac{1}{n} \left| \sum_{i=1}^n X_i \right| > \epsilon_0 U_n \right] &\leq 2 \exp \frac{-\epsilon_0^2 \log n}{2(1 + \epsilon_0 \sqrt{U_n})} \\ &\leq 2n^{-C'} \epsilon_0^2. \end{aligned}$$

D'où, pour un choix convenable de ϵ_0 on déduit que

$$\frac{1}{n} \sum_{i=1}^n X_i = O\left(\sqrt{U_n}\right)$$

Corollaire 2.4.2

- a) S'il existe une constante positive $M < \infty$, telle que :

$$|X_1| \leq M;$$

Alors on a,

$$\forall \epsilon > 0, \mathbb{P} \left[\left| \sum_{i=1}^n X_i \right| > \epsilon n \right] \leq 2 \exp \frac{-\epsilon^2 n}{2\sigma^2(1 + \frac{M\epsilon}{\sigma^2})};$$

où

$$\sigma^2 = \mathbb{E}X_1^2.$$

- b) Supposons que les $(X_i)_{1 \leq i \leq n}$ dépendent de n et que $\sigma_n^2 = \mathbb{E}X_i^2$, s'il existe $M = M_n < \infty$ telle que :

$$|X_1| \leq M,$$

et

$$\frac{M}{\sigma_n^2} \leq C < \infty;$$

et si

$U_n = n^{-1}\sigma_n^2 \log n$, vérifie $\lim_{n \rightarrow \infty} U_n = 0$, Alors on a

$$\frac{1}{n} \sum_{i=1}^n X_i = O\left(\sqrt{U_n}\right).a.co$$

Démonstration du Corollaire 2.4.2

a) En appliquant la proposition (2.4.1) à $a_i^2 = \sigma^2$, $A_n^2 = n\sigma^2$ et $b = M$ on aboutit à a).

b) Comme $\frac{MU_n}{\sigma_n^2}$ tend vers zéro, il suffit de reprendre le résultat a) pour $\epsilon = \epsilon_0 \sqrt{U_n}$, on arrive donc à l'existence d'une constante C' telle que :

$$\begin{aligned} \mathbb{P} \left[\frac{1}{n} \left| \sum_{i=1}^n X_i \right| > \epsilon_0 U_n \right] &\leq 2 \exp \frac{-\epsilon_0^2 \log n}{2 \left(1 + \epsilon_0 \sqrt{\frac{MU_n}{\sigma_n^2}} \right)} \\ &\leq 2n^{-C'} \epsilon_0^2. \end{aligned}$$

2.5 Convergence presque complète de l'estimateur de f dans le cas indépendant

Soit X une variable aléatoire à valeurs dans \mathbb{R}^d , on notera f la fonction de densité définie dans \mathbb{R}^d et rappelons que l'estimateur à noyau de $f_n(x)$ et $\hat{f}_n(x)$ définie par

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h^d} K\left(\frac{x - X_i}{h}\right)$$

et la norme que nous utilisons dans \mathbb{R}^d est donné par

$$\|x\| = \sqrt{\sum_{i=1}^d (x_i^2)} \tag{2.5}$$

Par la suite, nous introduisons les hypothèses de base permettant de donner un théorème général sur la convergence presque complète de \hat{f}_n vers f .

Hypothèses [16]

(H.1) f est continue au voisinage de x , x est un point fixé de \mathbb{R}^d

(H.2) Le paramètre de lissage h_n est tel que

$$\lim_{n \rightarrow \infty} h_n = 0 \quad \text{et} \quad \lim_{n \rightarrow \infty} \frac{\log n}{nh_n^d} = 0.$$

(H.3) Le noyau K est tel que

K est d'ordre k au sens de Gasser c'est à dire :

$$\int_{\mathbb{R}^d} t^j K(t) dt = 0 \quad \forall j = 1, 2, \dots, k-1 \quad \text{et} \quad 0 < \left| \int_{\mathbb{R}^d} t^k K(t) dt \right| < \infty$$

(H.4) K est borné, intégrable et à support compact.

Théorème 2.5.1 [16]

Si les conditions **(H.4)**, **(H.1)** et **(H.2)** sont vérifiées alors :

$$\lim_{n \rightarrow \infty} \hat{f}_n(x) = f(x).a.co \tag{2.6}$$

Démonstration du théorème 2.5.1 [16]

La démonstration de ce théorème est basée sur la décomposition suivante :

$$\hat{f}_n(x) - f(x) = \left(\hat{f}_n(x) - \mathbb{E} \left[\hat{f}_n(x) \right] \right) - \left(f(x) - \mathbb{E} \left[\hat{f}_n(x) \right] \right). \tag{2.7}$$

Le résultat du théorème découle alors des deux lemmes suivants.

Lemme 2.5.1

Si les conditions **(K.4)**, **(H.1)** et **(H.2)** sont vérifiées on a :

$$\mathbb{E} \left[\hat{f}_n(x) \right] = f(x). \tag{2.8}$$

Preuve de Lemme 2.5.1

Par équidistribution des X_i nous avons :

$$\mathbb{E}[\hat{f}_n(x)] = \frac{1}{h} \mathbb{E} X K \left(\frac{x-t}{h_n^d} \right)$$

En conditionnant par rapport à X on arrive à

$$\mathbb{E} \left[\hat{f}_n(x) \right] = \frac{1}{h_n^d} \int_{-\infty}^{+\infty} K \left(\frac{x-t}{h_n^d} \right) f(t) dt,$$

Le calcul de cette intégrale se fait en posons $z = \frac{x-t}{h_n}$ pour arriver à

$$\mathbb{E} [\hat{f}_n(x)] = \int_{\mathbb{R}} K(z) f(x - zh_n) dz.$$

La continuité uniforme de f sur le support compact de K entraîne

$$f(x - zh_n^d) \rightarrow f(x), \text{ uniformément en } z.$$

d'où

$$\lim_{n \rightarrow \infty} \mathbb{E} [\hat{f}_n(x)] = f(x).$$

Lemme 2.5.2

Sous les hypothèses **(H.4)**, **(H.1)** et **(H.2)** on a :

$$\hat{f}_n(x) - \mathbb{E} [\hat{f}_n(x)] = O \left(\sqrt{\frac{\log n}{nh_n^d}} \right) \tag{2.9}$$

Preuve de Lemme 2.5.2

Nous avons,

$$\begin{aligned} \hat{f}_n(x) - \mathbb{E} [\hat{f}_n(x)] &= \frac{1}{n} \sum_{i=1}^n h_n^{-1} \left[K \left(\frac{x - X_i}{h_n^d} \right) - \mathbb{E} K \left(\frac{x - X_i}{h_n^d} \right) \right] \\ &= \frac{1}{n} \sum_{i=1}^n \Gamma_i \end{aligned}$$

où

$$\Gamma_i = h_n^{-1} \left[K \left(\frac{x - X_i}{h_n^d} \right) - \mathbb{E} K \left(\frac{x - X_i}{h_n^d} \right) \right]$$

En utilisant l'hypothèse **(H.4)** on a :

$$|\Gamma_i| < \frac{c}{h_n^d}.$$

D'autre part le changement de variable $z = \frac{x-t}{h_n^d}$, nous donne

$$\begin{aligned} h^{-1}\mathbb{E}\left[h^{-1}K^2\left(\frac{X-x}{h_n^d}\right)\right] &= h^{-2}\int_{\mathbb{R}^d}K^2\left(\frac{x-t}{h_n^d}\right)f(t)dt. \\ &= h^{-1}\int_{\mathbb{R}^d}K^2(z)f(x-zh_n^d)dz. \end{aligned}$$

Comme K est bornée et f est continue sur le support compact de K , on a l'existence d'une constante C telle que :

$$\mathbb{E}\Gamma_i^2 < \frac{C}{h_n}.$$

On obtient alors, en appliquant le corollaire (1.1.2) de l'inégalité exponentielle de type Bernstein,

$$\hat{f}_n(x) - \mathbb{E}\hat{f}_n(x) = O\left(\sqrt{\frac{\log n}{nh_n^d}}\right). \quad (2.10)$$

En remplaçant l'hypothèse **(H.1)** par

(H.4) f est k fois continûment dérivable autour du point x .

On obtient une vitesse de convergence presque complète ponctuelle de l'estimateur à noyau.

Théorème 2.5.2 [12]

Sous les conditions **(K.4)**,**(H.2)** et **(H.4)** on a :

$$\hat{f}_n(x) - f(x) = O(h_n^k) + O\left(\sqrt{\frac{\log n}{nh_n^d}}\right) \quad (2.11)$$

Démonstration du Théorème 2.5.2

En reprenant la décomposition de la preuve précédente, le résultat du théorème sera établi par les lemmes suivant :

Lemme 2.5.2 [15]

Sous les conditions **(H.2)**,**(H.3)** et **(H.4)** on a :

$$\mathbb{E}\hat{f}_n(x) - f(x) = O(h_n^k). \quad (2.12)$$

Preuve du Lemme 2.5.2 [15]

On a :

$$\mathbb{E} [\hat{f}_n(x)] = \int_{\mathbb{R}^d} K \left(\frac{x-t}{h_n^d} \right) f(t) dt.$$

En posant $z = \frac{x-t}{h}$ on obtient

$$\mathbb{E} [\hat{f}_n(x) - f(x)] = \int_{\mathbb{R}^d} K(z)(f(x-zh) - f(x)) dz. \quad (2.13)$$

Il suffit alors de développer la fonction au voisinage de x , ce qui est possible au vu de la condition (H4) ceci s'écrit

$$g(x-zh) = \sum_{j=1}^k \frac{(-h)^j}{j!} \sum_{i_1+\dots+i_p=j} \left[T_k(i_1, \dots, i_p) \left(\frac{\partial^j g}{\partial x^{i_1} \dots \partial x^{i_p}} \right) \right]. \quad (2.14)$$

Puisque K est à support compact, les $o(h^j)$ intervenant dans l'expression ci-dessus sont uniformes $z = (z_1, \dots, z_p)$. Ainsi, on tire de condition (H.4),(2.8),(2.9) que

$$\mathbb{E} [\hat{f}_n(x)] - f(x) = \frac{(-h)^k}{k!} \sum_{j=1}^p \left(\frac{\partial^k g}{\partial x_j^k}(x) T_K(j) + o(h^k) \right) \quad (2.15)$$

Ceci achève la preuve.

Lemme 2.5.3

Sous les conditions (K.5),(H.1) et (H.2) on a

$$\mathbb{E} [\hat{f}_n(x)] - f(x) = O \left(\sqrt{\frac{\log n}{nh_n^d}} \right). \quad (2.16)$$

En prenant la décomposition de la preuve du lemme 2.5.2 précédent pour démontrer ce lemme .

2.6 Estimation de la fonction densité dans le cas quasi- associée

Dans cette partie, on suppose que les observations sont dépendantes, on donnant quelques notations et hypothèses sur celle ci .

Notations et hypothèses [12]

Soit une suite $(X_i)_{i \in \mathbb{N}}$ de variables aléatoires quasi associées à valeurs dans \mathbb{R}^d , $d > 1$. Nous supposons que ces variables ont une loi de probabilité P absolument continue par rapport à la mesure de Lebesgue λ_d sur \mathbb{R}^d et de densité f inconnue. On suppose en outre l'existence de la densité jointe $f(X_i, X_j)$ du couple (X_i, X_j) pour $i \neq j$ [9].

On considère l'estimateur à noyau de f défini par

$$f_n(x) = \frac{1}{nh_n^d} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i), x \in \mathbb{R}^d$$

où

$$K_h(u) = h_n^{-d} K(u/h_n), u = (u_1, \dots, u_d) \in \mathbb{R}^d$$

Le noyau K vérifier les conditions suivantes

- $K : \mathbb{R}^d \rightarrow \mathbb{R}$
- Pour tout $u = (u_1, \dots, u_d)$, $\lim_{\|u\| \rightarrow \infty} \|u\|^d K(u) = 0$
- $\|\cdot\|$ est une norme euclidienne sur \mathbb{R}^d et h_n une suite de nombre réels positifs telle que $h_n \rightarrow 0$ et $h_n^d \rightarrow \infty$

Notons θ_r le coefficient de covariance défini par

$$\theta_r = \sup_{s \geq r} \sum_{|i-j| \geq 1} \sum_{k=1}^d \sum_{l=1}^d |Cov(X_i^k, X_j^l)|, \forall r \geq 1$$

où

X_i^k est la $K^{\text{ème}}$ composante de vecteur X_i .

Dans le but d'établir les propriétés asymptotiques de f_n , nous introduisons les hypothèses suivantes sur le noyau K et la fonction de densité f .

Hypothèses sur le noyau K

- (K.1) K est lipschitzien borné ;
- (K.2) $K = K_1 - k_2$, avec $K = k_1$ et $K = k_2$ deux fonctions croissantes lipschitziennes bornées.
- (K.3) Il existe $R > 0$ tel que $K(t) = 0$ pour $\|t\| > R$.
- (K.4) Il existe $\rho > 0$ tel que pour $j = j_1 + \dots + j_d$ avec $(j_1, \dots, j_d) \in \mathbb{N}^d$.

$$\int x_1^{i_1} \dots x_d^{i_d} K(x_1, \dots, x_d) dx_1 \dots dx_d = \begin{cases} 1, & \text{si } j = 0; \\ 0, & \text{si } j \in \{1, \dots, [\rho] \dots 1\}; \\ \neq 0, & \text{si } j = [\rho]. \end{cases}$$

Hypothèse H1

- i) $\|f\|_\infty < \infty$ et $\sup_{|i-j| \geq 1} \|f_{(X_i, X_j)}\|_\infty < \infty$.
- ii) $\sup_{|i-j| \geq 1} \|g_{i,j}\|_\infty < \infty$, où $g_{i,j} = f_{(X_i, X_j)} - f \otimes f$

Nous avons aussi besoin des hypothèses suivantes sur la décroissance du coefficient de covariance θ_r

Hypothèse H2

- i) $\lim_{n \rightarrow \infty} \frac{nh_n^d}{\log^5(n)} = \infty$.
- ii) $\theta_r \leq a_0 e^{-ar}$, $a \geq 0$, $a_0 \geq 0$.
- iii) f est continument dérivable autour de point x .

2.7 convergence presque complète dans le cas quasi-associée

Pour démontré la convergence presque complète de l'estimateur dans ce cas, nous utilisons une inégalité exponentielle établie par Kallabis et Neumann (2006) suivante :

Inégalité exponentielle [19]

Soit X_1, \dots, X_n des variables aléatoires quasi associées telles que $\mathbb{E}X_i = 0$ et $\mathbb{P}(|X_i| \leq M) = 1$, pour tout $i = 1, \dots, n$, $M < \infty$. Soit $\sigma_n^2 = \text{Var}(X_1 + \dots + X_n)$.

Supposons qu'il existe $K < \infty$ et $\beta > 0$ telles que pour tous u -uplets (S_1, \dots, S_u) et v -uplets (t_1, \dots, t_v) avec $S_1 \leq \dots \leq S_u \leq S_u \leq t_1 \leq \dots \leq t_v$,

$$|Cov(X_{s_1}, \dots, X_{s_u}, X_{t_1}, \dots, X_{t_v})| \leq K^2 M^{u+v-2} v e^{-\beta(t_1 - s_u)}.$$

Alors,

$$\mathbb{P} \left(\sum_{i=1}^n X_i \geq t \right) \leq \exp \left(- \frac{t^2/2}{A_n + B_n^{\frac{1}{3}} t^{\frac{5}{3}}} \right);$$

Où $A_n \leq \sigma_n^2$ et $B_n = \left(\frac{16nK^2}{9A_n(1-e^{-\beta})} \vee 1 \right) \frac{2(K \vee M)}{1-e^{-\beta}}$.

Théorème 2.7.1

Supposons que les conditions **(K.1)**, **(H1.i)**, **(H2.iii)** et **(H2.i)** sont vérifiées alors

$$\hat{f}_n(x) - f(x) = O(h^k) + O \left(\sqrt{\frac{\log n}{nh_n^d}} \right) \text{ p.co.}$$

Démonstration de théorème 2.7.1

La démonstration de ce théorème découle de la décomposition suivant

$$\hat{f}_n(x) - f(x) = \left(\hat{f}_n(x) - \mathbb{E} \left[\hat{f}_n(x) \right] \right) - \left(f(x) - \mathbb{E} \left[\hat{f}_n(x) \right] \right). \quad (2.17)$$

Et les deux lemmes suivants

Lemme 2.7.1

Si les conditions **(K.1)** et **(H2.i)** sont vérifiées on a :

$$\mathbb{E} \left[\hat{f}_n(x) \right] - f(x) = O(h^k)$$

Preuve Lemme 2.7.1(voir preuve de Lemme(2.3.3) de cas indépendant)

Lemme 2.7.2

Supposons que les hypothèses **(k.1)**, **(H2.i)** et **(H2.ii)** sont vérifiées, alors, pour $c > 0$ et $b \geq 0$,

$$|\hat{f}_n(x) - \mathbb{E} \hat{f}_n(x)| = O \left(\sqrt{\frac{\log(n)}{nh_n^d}} \right) \text{ p.co} \quad (2.18)$$

Démonstration de Lemme 2.7.2

Soit

$$g(X_i) = K\left(\frac{X_i - x}{h_n}\right) - \mathbb{E}K\left(\frac{X_i - x}{h_n}\right).$$

et

$$\hat{f}_n(x) - \mathbb{E}\hat{f}_n(x) = \frac{1}{nh_n^d} \sum_{i=1}^n g(X_i).$$

La fonction g satisfait

$$\|g\| \leq 2\|K\| \text{ et } Lip(g) \leq 2\frac{Lip(K)}{h_n}$$

Soient (S_1, \dots, S_u) et (t_1, \dots, t_v) telles que $1 \leq S_1 \leq \dots \leq S_u \leq t_1 \leq \dots \leq t_v \leq n$. Si $r = t_1 - s_u > 0$, par quasi-association des variables X_1, \dots, X_n , on a

$$\begin{aligned} |cov(g(X_{s_1}) \dots g(X_{s_u}), g(X_{t_1}) \dots g(X_{t_v}))| &\leq \|g\|^{u+v-2} (Lip(g))^2 \sum_{i=1}^u \sum_{j=1}^v \sum_{k=1}^d \sum_{l=1}^d |cov(X_{s_i}^k, X_{t_j}^l)| \quad (2.19) \\ &\leq C^{u+v} h_n^{-2} (u \wedge v) \theta_r. \quad (2.20) \end{aligned}$$

Si f est continue et la condition **(H1.i)** est vérifiée, alors

$$|cov(g(X_{s_1}) \dots g(X_{s_u}), g(X_{t_1}) \dots g(X_{t_v}))| \leq C^{u+v} h_n^2. \quad (2.21)$$

En multipliant (2.19) par $\frac{1}{4}$ et (2.20) par $\frac{3}{4}$, nous obtenons

$$|cov(g(X_{s_1}) \dots g(X_{s_u}), g(X_{t_1}) \dots g(X_{t_v}))| \leq C^{u+v} h_n (u \wedge v) e^{-\frac{ad}{2d+2}r}. \quad (2.22)$$

Si $r = 0$,

$$|cov(g(X_{s_1}) \dots g(X_{s_u}), g(X_{t_1}) \dots g(X_{t_v}))| \leq C^{u+v} h_n^d;$$

après on a

$$Var\left(\sum_{i=1}^n g(X_i)\right) = (nh_n^d)^2 Var \hat{f}_n(x) = nh_n^d (f(x) \int_{\mathbb{R}} K^2(u) du + o(1)).$$

En appliquant l'inégalité exponentielle, sous la condition **(H2.i)**, on a

$$\begin{aligned} \mathbb{P} \left(|\hat{f}_n(x) - \mathbb{E}\hat{f}_n(x)| > \eta \sqrt{\frac{\log(n)}{nh_n^d}} \right) &= \mathbb{P} \left(\left| \sum_{i=1}^n g(X_i) \right| > \eta (nh_n^d \log(n))^{\frac{1}{2}} \right) \\ &\leq 2 \exp \left(- \frac{\eta^2 \log(n)}{4f(x) \int_{\mathbb{R}} K^2(u) du (1 + o(1))} \right). \end{aligned}$$

Pour $\eta > 2(f(x) \int_{\mathbb{R}} K^2(u) du)^{1/2}$, on a

$$\sum_{n=1}^{\infty} \mathbb{P} \left(\left(\frac{nh_n^d}{\log(n)} \right)^{1/2} |\hat{f}_n(x) - \mathbb{E}\hat{f}_n(x)| > \eta \right) < \infty \Leftrightarrow \hat{f}_n(x) - \mathbb{E}\hat{f}_n(x) = O \left(\sqrt{\frac{\log n}{nh_n^d}} \right) \quad (2.23)$$

Corollaire 2.7.1

Si les conditions **(K.1)**, **(K.3)** et **(k.4)** et les conditions **(H1.i)**, **(H2.iii)** sont vérifiées, et si $f \in C^\rho$, alors le choix de $h_n \sim \left(\frac{\log(n)}{nh_n^d} \right)^{\frac{\rho}{2\rho+d}}$ entraîne

$$|\hat{f}_n(x) - \mathbb{E}\hat{f}_n(x)| = O \left(\frac{\log(n)}{nh_n^d} \right)^{\frac{\rho}{2\rho+d}} \text{ p.co.}$$

CHAPITRE 3

SIMULATION

Introduction

Dans cette partie, nous proposons une étude numérique, à l'aide de logiciel R, pour illustrer nos résultats concernant l'estimation de la densité pour les variables quasi-associées.

3.1 La simulation

Soit $(X_t)_{t \in \mathbb{Z}}$ le modèle moyenne mobile d'ordre 1 défini par

$$X_t = \epsilon_t + 0.2\epsilon_{t-1}, t \in \mathbb{Z}$$

Où $(\epsilon_t)_{t \in \mathbb{Z}}$ est une suite de variables aléatoires indépendantes de même loi gaussienne centrée réduite ($\epsilon_t \sim N(0, 1), t \in \mathbb{Z}$) Le modèle $(X_t)_{t \in \mathbb{Z}}$ est aussi gaussien tel que $\mathbb{E}(X_t) = 0$ et $Var(X_t) = 1.4$.
Considérons le processus $(Y_t)_{t \in \mathbb{Z}}$ défini par

$$Y_t = (1 + W_t)X_t, t \in \mathbb{Z}$$

Où $(W_t)_{t \in \mathbb{Z}}$ est une suite de variables aléatoires indépendantes i.i.d, de même loi de Bernoulli de paramètre $\frac{1}{2}$ tel que
($P(W_t = 1) = P(W_t = 0) = 1/2$), et indépendante de la suite $(X_t)_{t \in \mathbb{Z}}$.

Le processus $(Y_t)_{t \in \mathbb{Z}}$ est associé (transformation croissante des suites associées $(X_t)_{t \in \mathbb{Z}}$ et $(W_t)_{t \in \mathbb{Z}}$)

.Sa loi est une mélange de deux lois gaussiennes de densité f définie, pour tout $x \in \mathbb{R}^d$.

$$f_Y(y) = \frac{1}{4\sqrt{2.8\pi}} \exp^{-\frac{y^2}{11.2}} + \frac{1}{2\sqrt{2.8\pi}} \exp^{-\frac{y^2}{2.8}}$$

En effet

on a $X \rightsquigarrow \mathcal{N}(0, 1.4)$ et $W \rightsquigarrow \mathcal{B}(\frac{1}{2})$, telle que X indépendant de W .

on cherche à trouver la loi de Y telle que $Y = (W + 1)X$.

On a

$$F_Y(y) = \mathbb{P}((W + 1)X \leq y)$$

On distingue deux cas

1) Le premier cas si $W = 0$

$$F_Y(y) = \frac{1}{2} \mathbb{P}(X \leq y)$$

2) Le deuxième cas si $W = 1$

$$F_Y(y) = \frac{1}{2} \mathbb{P}(2X \leq y)$$

d'où

$$\begin{aligned} F_Y(y) &= \frac{1}{2} \mathbb{P}(X \leq y) + \frac{1}{2} \mathbb{P}(2X \leq y) \\ &= \frac{1}{2} F_X(y) + \frac{1}{2} F_X\left(\frac{y}{2}\right) \\ &= \frac{1}{2} \left[\mathbb{P}\left(\frac{X}{\sqrt{1.4}} \leq \frac{y}{2\sqrt{1.4}}\right) + \mathbb{P}\left(\frac{X}{\sqrt{1.4}} \leq \frac{y}{\sqrt{1.4}}\right) \right] \end{aligned}$$

d'où On conclut que la loi de Y est une mélange de deux lois gaussiennes dont sa fonction de densité est la suivant

$$f_Y(y) = \frac{1}{2} \left(f_X(y) + \frac{1}{2} f_X\left(\frac{y}{2}\right) \right)$$

Avec

$$f_X(y) = \frac{1}{4\sqrt{2.8\pi}} \exp^{-\frac{y^2}{11.2}}$$

Et

$$f_X\left(\frac{y}{2}\right) = \frac{1}{2\sqrt{2.8\pi}} \exp^{-\frac{y^2}{2.8}}$$

Rappelons l'estimateur à noyau \hat{f}_n de f

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

Dans nos simulations, nous utilisons le noyau d'Epanachnikov, défini par

$$K(x) = \frac{3}{4}(1 - x^2)\mathbf{1}_{\{|x| \leq 1\}}$$

Et une fenêtre $h_n = 0.4$.

D'après la représentation graphique et les erreurs relatives calculer (ERM = 0.7670894, ERM = 0.7667207, ERM = 0.7602718) nous remarquons que la fonction de densité et son estimateur sont proches pour les tailles de l'échantillon $n = 500, 1000, 1500$.

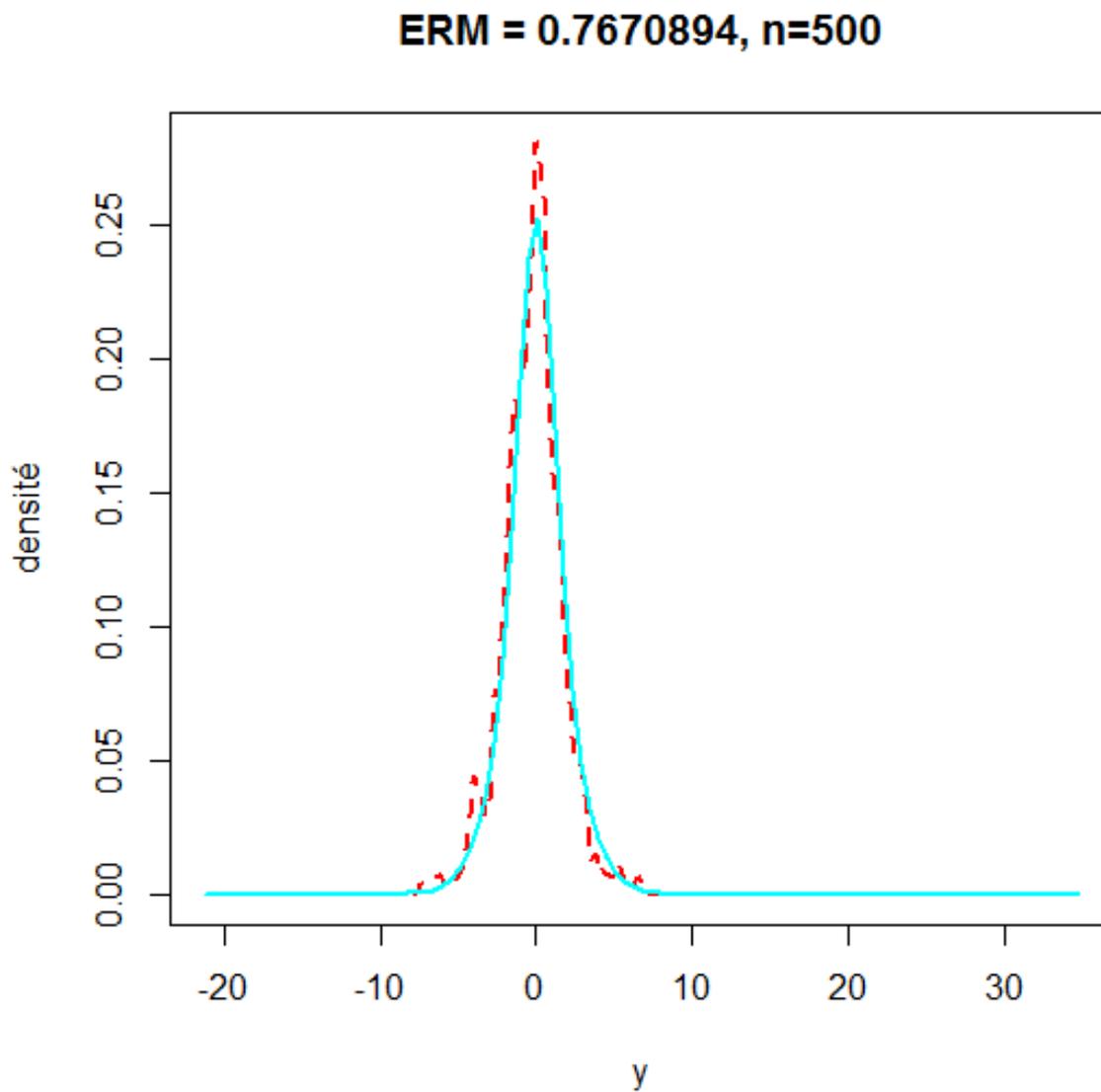


FIGURE 3.1 – Fonction de densité et sa fonction estimé par la méthode de noyau pour $n = 500$

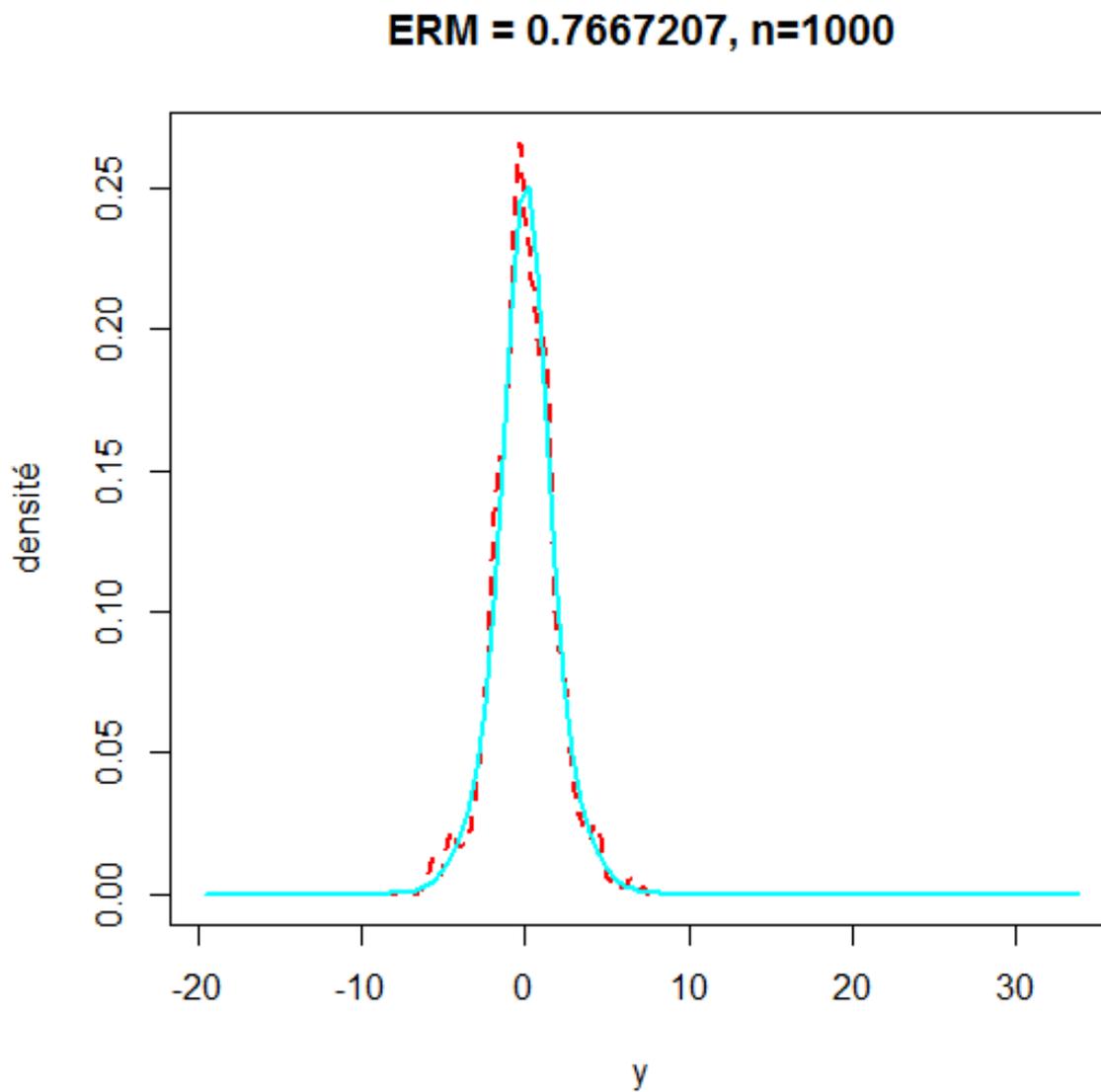


FIGURE 3.2 – Fonction de densité et sa fonction estimé par la méthode de noyau pour $n = 1000$

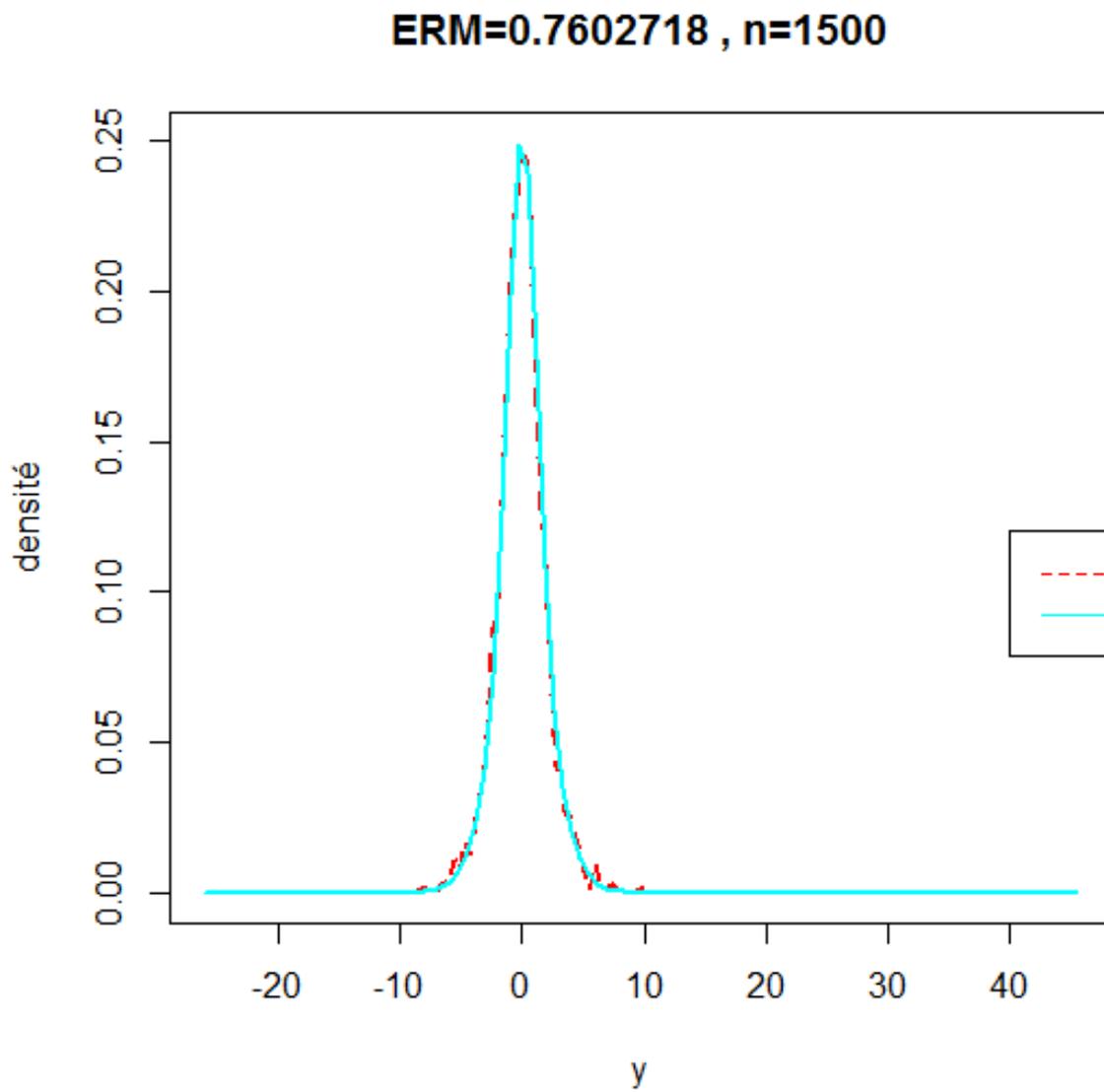


FIGURE 3.3 – Fonction de densité et sa fonction estimé par la méthode de noyau pour $n = 1500$

Conclusion

Dans ce mémoire, nous avons étudiés quelques notion de l'association en donnant quelques théorèmes, propriétés et exemples. Nous avons introduit l'estimation non paramétrique de la fonction de densité f avec la méthode du noyau dans le cas multivarié dans les deux cas indépendant et le quasi associé. Nous avons établi la convergence presque complète de l'estimateur ainsi que sa vitesse de convergence sous les hypothèses standards de la statistique semi paramétriques.

Enfin, pour valider les résultats théoriques obtenu une simulation a été réaliser a cet effet. Notre travail s'est limité à l'étude de la convergence presque complète de f dans le cas multidimensionnel. Il serait intéressant de faire le même travail dans le cas fonctionnel et de considérer d'autres paramètres comme la fonction de répartition, la fonction de hasard...etc.

BIBLIOGRAPHIE

- [1] Attaoui.S.(2012).Sue l'estimation semi paramétrique robuste en statistique fonctionnelle. Thèse de doctorat. Université Djillali (Algérie).
- [2] Arab.I.(2016).Procédures d'approximation stochastique à erreurs associées.Université Bejaia.
- [3] Barlow.R.E.Proschan.F.(1981).Statistical Theory of reliability and Life.
- [4] Begai,L.Prakasa Rao, L.S. (1995). Kernel-type density and failure rate estimation for associated sequences. Ann. Inst. Statist. Math. 47, 253-266.
- [5] Birkel, T. (1988). On the convergence rate in Central Limit Theorem for associated processus. Ann. Probab, 16,1685-1698.
- [6] Benchoulak.H.(2012).Bandes de confiance pour les fonction de densité et de régression. Thèse de doctorat.Université Constantine.
- [7] Bulinski,A.V.(1996). On the convergence rates in the CLT for positively and negatively random fields.Inprobability Theory and Mathématiqueal Statistics. Gordon and Breach Publishers, Amsterdam.
- [8] Bulinski.a.v. and Shanavich.E. (1998). Asymptotical behaviour for some functionals of positively and negatively dependand random fields.
- [9] Bulinski.A.V, and shashkin. A.(2007).Normal approximation for quasi-associated random fields.
- [10] Bochner.S (1955). Harmonic Analysis and the Theory of probability, University of chicago.
- [11] Cox.J.T. and Grimmett.G, (1984). Central limit theorems for associated random variables and the percolation model.
- [12] Douge.Lahcen.(2009).L'estimation fonctionnelle des processus associées et quasi-associés. Université PARIS.

- [13] Esary, J. Proschan, F. and Walkup, D. (1967). Association of random variables with applications. *Ann. Math. Statist.* 38, 1466-1476.
- [14] Fortuyn, C. Kastelyn, P. and Ginibre, J. (1971). Correlation inequalities in some partially ordered sets. *Comm. Math. Phys.* 74, 119-127.
- [15] Ferrani.yacine. (2016). Sur l'estimation non paramétrique de la densité et du mode dans les modèle de données incomplète et associées.Thèse de doctorat.Université du Littoral cote de d'Opale.
- [16] Frédéric.Ferraty et Philippe Vieu. Cours de DEU, module statistique fonctionnelle.
- [17] Gherair.Djemoi.(2017). Modélisation non paramétrique pour les variables aléatoires fonctionnelles cas de données indépendantes.Université Kasdi Merbah.Ouargla.
- [18] Hoeffding.W. (1940). Masstabinvariante Korrelations théorie. *schr.Math. university Bertin.*
- [19] Kallabis.R.S et Newman.M.H (2006). An exponential inequality under weak dependence.
- [20] Lehmann.E.L.(1966). Some concepts of dependence. *Ann.Math.Statist.* 37, 1137-1153.
- [21] Nadaraya.E.A.(1964). On estimating regression. *Theory.probability.*
- [22] Newman.C.M. (1980). Normal fluctuation and the FKG inequalities. *Comm. Math. Phys.* Vol 74, 2, 119-128.
- [23] Newman, C.M. (1984). Asymptotic independence and limit theorems for positively and negatively dependent random variables : in *Inequalities in Statistics and Probability, IMS Lect. Notes-Monographs Series*, 5, 127-140.
- [24] Pitt.L.D. (1982). Positively correlated normal variables are associated.
- [25] Rosenblatt.M. (1956). Remarks on some non parametric estimates of a density function.
- [26] Uspensky, J. (1935) *Introduction to mathematical probability.* McGraw-Hill, New York.

Résumé

Dans ce travail, nous introduisons l'estimation non paramétrique par la méthode du noyau de la fonction densité f dans le cas multivarié. Nous traitons les propriétés asymptotiques de cet estimateurs dans le cas indépendant et dépendant. La dépendance est modélisée via la corrélation quasi-associé. La vitesse de convergence ponctuelle presque complète de l'estimateur à noyau de la dite fonction a été établi. Afin de valider les résultats théoriques obtenus, une simulation a été réalisée à cet effet.

Mots clés

L'association, Quasi-association , L'estimation à noyau, Convergence presque complète , Simulation.

Abstract

In this work, we have introduced the non parametric estimation, with the kernel's method of the density function in the multivariate case. We have treat the asymptotic properties of this estimators in both independent and depended cases. The dependence is modeled via the quasi-correlated correlation. The almost convergence of the kernel estimator is established, the simulation is given to illustrate the theoretical results.

Keywords

Association, Quasi association, Kernel estimation, convergence almost complete, Simulation.