République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université d'ABEDRRAHMANE Mira de BEJAIA

Faculté des Sciences Exactes

Département de Mathématiques

Option : Statistique et Analyse Décisionnelle.

Mémoire de fin de cycle en vue de l'obtention du diplôme de Master en Mathématiques

Théme

Estimation non paramétrique de la densité de probabilité et de la fonction de répartition par des séries orthogonales

Présenté par :

-M^{elle} ZERNOUN Kahina.

Devant le jury compose de :

Président	M.BOURAINE.M	C.A.A	Université A-Mira de Béjaia.
Examinatrice	M^{me} LAGHA.K	M.C.B	Université A-Mira de Béjaia.
Rapporteur	M^{me} SAADI.N	M.C.A	Université A-Mira de Béjaia.

Année universitaire 2016-2017

Remerciements

Tout d'abord, nous tenons à remercier le "BON DIEU" le tout puissant de nous avoir accordé patience, courage et volonté afin de réaliser mener à terme ce modeste travail.

J'ai l'honneur et le plaisir d'exprimer mes profonde Gratitude à M^{elle} SAADI promotrice pour avoir accepté de m' encadrer, pour ses remarques, ses conseils et ses Orientations.

Les membres de jury d'avoir accepté d'évaluer et d'examiner ce travail.

kahina

Dédicaces

A ceux qui m'ont tout donné sans rien en retour A ceux qui m'ont encouragée et soutenue dans les moments les plus difficiles A vous mes chers parents Le plus beau cadeau que Dieu puissent faire à un enfant, pour leur amour et leur support continu. Que ce travail soit le témoignage sincère et affectueux de ma profonde reconnaissance pour tout ce que vous avez fait pour moi.

A mes chers grands pères et grands mère.

A mes chères frères hamza et nanou.

A mes chères Sœurs bahya et samou.

A mon mari houes et sa famille.

A toute la promo SAD.

J'adresse aussi mes dédicaces à mes amies avec qui j'ai passé des moments agréables, en particulier à : Madouche, sakou, Sabo, rozina, chacha et khadou.

Table des matières

Introduction générales

Un des plus vieux problèmes de la statistique non paramétrique consiste à estimer la densité de probabilité $f(.)$ et la fonction de répartition $F(.)$ à partir d'un échantillon de variables aléatoires indépendantes et identiquement distribuées $X_1, X_2, ..., X_n$. Il s'agit d'un problème fondamental qui a connu, durant ces dernières années, des développements théoriques et pratiques à la fois rapides et nombreux. Le problème de l'estimation de la densité de probabilité et la fonction de répartition est important pour plusieurs raisons :

- La densité de probabilité permet d'avoir un aperçu très rapide des principales caractéristiques de la distribution (pics, creux, symétries,...), ce qui explique le volume important de littérature qui lui est consacré.
- 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 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 $[\theta_n - \epsilon, \theta_n + \epsilon]$, où θ_n est un estimateur de θ, on a en fait besoin d'un estimateur de la fonction de répartition de θ_n. Les fonctions f et F comme la fonction caractéristique décrivent complètement la loi de probabilité des observations et en connaitre une estimation convenable permet de résoudre un nombre de problèmes statistiques. Cette estimation tient donc naturellement une place importante dans l'etude 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.

Pour ces raisons nous nous intéressons dans la première et la deuxième partie de ce mémoire à des problèmes d'estimation non paramétrique, de la densité de probabilité. Nous regarderons brièvement en quoi consiste la méthode du noyau et la méthode d'estimation par histogramme et en détail la méthode des séries orthogonales vu sa souplesse d'utilisation et ses propriétés de convergences.

L'outil d'estimation non paramétrique de la densité de probabilité nous est fourni par l'histogramme : une fois les données regroupées en classes de valeurs, les fréquences empiriques sont représentées par des aires rectangulaires dont les bases correspondent aux classes elles mêmes. L'histogramme convient bien pour des analyses relativement grossières. Nénéanmoins, ses discontinuités n'apparaissent pas très naturelles. Pour des densités raisonnablement lisses, l'histogramme apparîât donc comme un estimateur sévèrement limité. Il existe d'autres méthodes non paramétriques plus robustes que la méthode par histogramme : la méthode d'estimation par des séries orthogonales et la méthode du noyau. C'est Rosenblatt (1956), suivi de Parzen (1962), qui ont proposé une classe d'estimation à noyau d'une densité univariée. Les estimateurs à noyau sont fonction de deux paramètres K, appelé noyau, et h dit paramètre de lissage (largeur de fenêtre). Les propriétés de convergence de l'estimateur à noyau ont été établies par Parzen (1962), Silverman et Nadaraya (1964). Les théorèmes relatifs à l'erreur quadratique moyenne et l'erreur quadratique moyenne intégrée ont été obtenus par Parzen (1962) .

Estimation de la densité de probabilité par la méthode des fonctions orthogonales a été introduite par Cencov (1962). Cette méthode est très utilisée dans de nombreux domaines (analyse harmonique, traitement du signal, compression d'images, statistique fonctionnelle...).

Son succès est du à son adaptabilité aux données et à sa facilité d'implémentation. La différence entre l'estimateur de la fonction densité par des fonctions orthogonales et les estimateurs à noyau réside surtout dans leur utilisation pratique. En effet, l'emploi de

l'estimateur de la fonction densité par des fonctions orthogonales ne nécessite, une fois choisie les fonctions orthogonales que la détermination préalable du nombre de termes appelé aussi paramètre de lissage . Par contre l'emploi des estimateurs à noyau ne peut se faire qu'après avoir choisi auparavant le noyau optimal. Or, en l'absence de toute information sur $f(x)$, ce choix ne peut être qu'arbitraire. L'estimateur à noyau obtenu est fonction d'un paramètre appelé paramètre de lissage ou fenêtre, actuellement il n'existe pas de choix optimal pour ce paramètre, le choix optimal qui minimise l'erreur quadratique moyenne intégrée dépend de la dérivée seconde de la densité inconnue.

A la suite de Cencov(1962), de nombreux auteurs ont contribué à améliorer cette méthode. Les principales études ont concerné le choix de la base orthogonale et celui du paramètre de lissage (nombre de termes de la série orthogonale). Sur le premier point, Efromovich (1997) s'est basé sur les travaux de Kolmogorov (1955) concernant les systèmes orthonormaux. On trouvera de nombreux autres exemples de bases possibles dans les travaux de Sanson (1959) ou Kolmogorov et Fomin (1957) . Le choix du paramètre de lissage fut souvent choisi de manière à minimiser un estimateur de l'espérance quadratique intégrée Hart (1985), Diggle et Hall (1986) ou Tarter et Lock (1993). Préalablement, certains auteurs avaient essayé de contourner l'obstacle d'une somme avec une infinité de termes dans l'espérance quadratique intégrée. Ils s'étaient basés sur la contribution individuelle de chaque composante ajoutée sur l'estimation de l'erreur quadratique intégrée Kronmal et Tarter (1968), 1976). Pour compléter cette bibliographie, on peut se référer aux ouvrages de Prakasa (1983) et Bosq (2002) Saadi et Adjabi (2008).

Sam Efromovich (2010) a introduit un nouvel estimateur de la densité de probabilité basé sur un système trigonométrique sur $[0, 1]$. L'auteur a donné ses propriétés statistiques (biais , la variance, l'erreur quadratique moyenne , l'erreur quadratique moyenne intégrée). D'autres résultats furent également obtenu par les auteurs Lagha et Adjabi (2016) qui ont introduit une classe de fonctions basée sur ce système trigonométrique, principalement caractérisé par le fait que le nombre d'observations (la taille de l'échantillon) est aléatoire.

L'objectif de la premiére partie de ce travail est d'établir les propriétés asymptotiques et les théorèmes de convergences (Convergence en moyenne quadratique, convergence en

moyenne quadratique intégrée, vitesse de convergence en moyenne quadratique , vitesse de convergence en moyenne quadratique intégrée) de cet estimateur dans le cas où la taille de l'échantillon est fixée. A fin d'étudier les qualités statistiques de l'estimateur obtenu , on le compare numériquement par simulation à l'estimateur associé à la base de Saadi et Adjabi .

L'estimateur classique de la fonction de répartition est la fonction de répartition empirique. Les premiers articles consacrés à ce sujet parut dans Kolmongorov (1939), avec depuis cette date, plusieurs articles publiés sur ce sujet : (Yammato (1973), Efron (1993)). La fonction de répartition empirique possède 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, il existe des estimateurs qui sont préférables à la fonction de répartition empirique par exemple , estimateur à noyau introduit par Nadaraya (1964). Ses propriétés sont données par Winter (1973) et Yamato(1973).

Estimation de la fonction de répartition par des séries orthogonales a été introduit par Kronmal et Tarter (1968, 1976). A la suite de Kronmal et Tarter , plusieurs auteurs ont contribué à améliorer cette méthode. Les principales études ont concerné le choix du paramètre de lissage (nombre de termes de la série orthogonale). La règle adoptée pour déterminer le nombre optimum a été développé par Kronmal et Tarter généralisée par Ott et Kronmal (1976). Les inconvénients de cette méthode ont été soulignés par Crain qui a suggéré qu'il pourrait ne pas donner le terme optimal. Hall (1986), a averti au sujet de la mauvaise performance possible et même l'incohérence de la règle dans des situations multimodales. Saadi et Adjabi (2016) ont proposé une nouvelle technique pour sélectionner ce paramètre de lissage qui aide à surmonter ces difficultés. Les difficultés rencontrées avec la méthode de Kronmal -Tarter sont dues au fait que les termes sont analysés un par un, La méthode de Saadi-Adjabi est basée sur la manipulation de nombreux termes ensemble .

Le deuxième objectif de ce travail est d'établir certaines propriétés de l'estimateur de la fonction de répartition basé sur des séries trigonométriques. Nous établissons ces propriétés statistiques (biais, variance , erreur quadratique moyenne et l'erreur quadratique

moyenne intégrée) et ces propriétés asymptotiques (convergence de biais, convergence de la variance, la convergence en moyenne quadratique et la convergence en moyenne quadratique intégrée). A fin d'étudier les qualités statistiques de l'estimateur obtenu , on le compare numériquement par simulation à l'estimateur associé à la base de Saadi et Adjabi.

Afin d'atteindre nos objectif, nous avons structuré notre travail en cinq chapitres : le premier chapitre portera sur l'estimation non paramétrique de la densité de probabilité. Dans cette partie nous allons données les différentes méthodes d'estimation de la densité de probabilité à savoir : la méthode du noyau, estimation par histogramme et la méthode des fonctions orthogonales . Le deuxième chapitre sera motivée par l'application de cette méthode d'estimation de la densité en utilisant un système trigonométrique . Dans le chapitre 3 nous allons étudiées les différentes méthodes d'estimation de la fonction de répartition. Dans le chapitre 4 nous allons présentées un nouvel estimateur de la fonction de répartition basé sur un système trigonométrique.

Enfin, dans le chapitre 5, nous présentons les résultats des simulations conduites à partir de densité cibles connue : loi normale. L'expérimentation numérique nous servira en particulier à étudier les performances des estimateurs proposés .

Ce mémoire se termine par une conclusion générale et quelques propositions d'axes de recherche.

Chapitre 1

Estimation non paramérique de la densité de probabilité

1.1 Introduction

Dans de nombreuses applications, la densité f est inconnue et on dispose d'un n-échantillon $X_1, X_2, ..., X_n$ de variables aléatoires indépendantes et identiquement distribuées admettant f comme densité. Le problème du statisticien consiste alors à utiliser cet échantillon pour construire un estimateur qui soit le plus proche possible de la densité f. Les premiers articles consacrés à ce sujet sont dus au biométricien Karl Pearson [33], Rosenblatt [37] et Cencov [7].

Plusieur estimateurs de la densité de probabilité ont été proposés depuis les travaux de Rosenblatt [37], Cencov [7] et Parzen [32].

La grande majorité d'entre eux se rangent dans une classe trés importantes contribuées par des estimateurs batis à partir d'un noyau (méthode de nouyau et l'estimation par histogramme). une autre classe d'estimateurs basé sur la notion de développement en séries de fonctions orthogonales, fut proposée par Cencov [7]. Bien que moin étudiée que la précédente, elle semble actuellement connaître un nouvel essor (méthode des fonctions orthogonales).

Dans cette partie nous étudions en détail l'estimation de la densité de probabilité par des séries orthogonales.

1.2 Estimation par histogramme

Elle consiste à estimer la densité de la variable aléatoire X en x par n_i le nombre d'occurences de réalisations x_i appartenent à la iéme classe associée à la valeur x.

Etant données des observations $(x_1, x_2, ..., x_n)$ qui sont les réalisations des variables aléatoires réelles indépendantes et identiquement distribuées $X_1, X_2, ..., X_n$ de densité f inconnue sur un intervalle fini $[a, b]$. Pour le faire de manière non paramétrique, il est naturel de se donner k intervalle (classe) $(I_j)_{j=0,1...k-1}$ avec $I_j = [a_{j-1}, a_j[$.

Pour construire l'histogramme nous devons choisir une origine x_0 et une largeur d'intervalle h. la largeur contrôle principalement la qualité de lissa ge. L'histogramme peut être généralisé en autorisant la largeur d'intervalle à varie. Supposons que la droite est coupée en intervalles, la densité estimée est :

$$\hat{f}_h(x) = \frac{1}{n} \frac{card\{x_i \quad dans \quad le \quad même \quad intervalle \quad que \quad x\}}{largeur \quad de \quad l'intervalle \quad contenant \quad x} \tag{1.1}$$

l'estimateur de $f sur [a_j, a_{j+1}]$ est défini par :

$$\begin{aligned} \hat{f}_h(x) &= \frac{card\{x_i \leq a_{j+1}\} - card\{x_{i\leq} a_j\}}{nh} \\ &= \frac{n_j}{nh}, x \in]a_j, a_{j+1}]. \end{aligned} \tag{1.2}$$

la construction de $\hat{f}_h(x)$ peut $\hat{e}tre$ faite même si les données ne proviennent pas d'une loi continue.

1.2.1 Propiétés de l'histogramme

Pour construire un bon histogramme, il faut trouver un juste équilibre entre le nombre d'observations n et le nombre de classes déterminé par k. Pour mesurer la précision de $\hat{f}_h$ en un point x fixé, on peut utiliser l'erreur quadratique moyenne :

$$MSE(\hat{f}_h(x)) = \mathbb{E}[\hat{f}_h(x) - f(x)]^2 \tag{1.3}$$

Il est cependant préférable de mesurer la précision de façons globale en calculant l'erreur quadratique moyenne intégrée :

$$MISE(\hat{f}_h(x)) = \int MSE(\hat{f}_h(x))dx \tag{1.4}$$

D'aprés [18], on a :

$$n_j \underset{loi}{\longrightarrow} \beta(n, p_j)$$

où β est la loi binomiale de paramétre n et $p_j = p(a_j < X \leq a_{j+1})$

pour $x \in [a_j, a_{j+1}]$, on déduit que

$$\mathbb{E}[\hat{f}_h(x)] = \frac{\mathbb{E}(n_j)}{nh} = \frac{p_j}{h}, \tag{1.5}$$

et

$$\mathbb{V}ar[\hat{f}_h(x)] = \frac{p_j(1-p_j)}{nh^2}. \tag{1.6}$$

L'expression de la variance montre que plus le paramétre h est petit, plus $\hat{f}_h(x)$ est variable ; inversement, plus h est grand, moins $\hat{f}_h(x)$ est variable.

En faisant le développement de Taylor à l'ordre 1, on obtient les propriété suivantes [4] :

1.
$$\mathbb{B}iais(\hat{f}_h(x)) = \frac{f'}{2}[h - 2(x - a_j)] + O(h^2) \tag{1.7}$$

2.
$$\mathbb{V}ar(\hat{f}_h(x)) = \frac{f(x)}{nh} - O(n^{-1}) \tag{1.8}$$

D'aprés (1.7) on peut constater que $h \to 0$ entraîne que le biais $(\hat{f}_h(x)) \to 0$. D'aprés (1.8), pour que $\hat{f}_h(x))$ soit peut variable, il faut que $nh \to \infty$.

3.
$$MSE(\hat{f}_h(x)) = \frac{f(x)}{nh} + \frac{f'^2(x)}{4}[h - 2(x - a_j)]^2 + O(h^3) + O(n^{-1}).$$

4.

$$MISE(f_h(x)) = \frac{1}{nh} + \frac{h^2 \int f'^2(x)dt}{12} + O(h^3) + O(n^{-1}). \tag{1.9}$$

Dans (1.9), on voit que le paramétre de lissage h est relié directement au terme $\frac{h^2 \int f'(t)^2 dt}{12}$ provenant du carré de biais intégré et que ce paramètre est inversement portionnel à la variance intégrée. Un petit h donne un histogramme peu biaisé, tandis qu'un grand h et un grand n déterminent un histogramme moins variable.

1.2.2 Choix du paramètre de lissage

En pratique, on choisi h en fonction de n. Les règles les plus utulisées sont :

Règle de sturges [40]

Prendre le nombre de k de classes égal à $1 + \log_2 n$. En pratique, cela revient à prendre

$$h = \frac{x_k - x_1}{k}, o\grave{u}$$

Les $x_{(i)}$ sont les valeurs d'observations d'un échantillon ordonné par ordre croissant.
la régle de Sturges a tendance à produire des histogrammes trop lisses.

Règle de Scott [41]

La valeur qui minimise l'erreur quadratique moyenne intégrée, $MISE$ est donnée par :

$$h_{opt} = [\frac{6}{\int f'(t)^2 dt}]\cdot$$

En prenant pour f la densité de loi normale $N(\mu, \sigma^2)$, d'aprés Scott [41] on peut alors montrer que :

$$h_{opt} = 3.491\sigma n^{-\frac{1}{3}}\cdot$$

En estiment σ par l'ecart type S de l'échantillon, on aura :

$$h_{opt} = 3.491 S n^{-\frac{1}{3}}\cdot$$

Remarque 1.2.1.

L'hitogramme convient bien pour des analyses relativement grossiéres. Néanmoins,ses discontinuités n'apparaissent pas trés naturelles et ce qui est plus grave, les points tombant prés du bords d'une classe et ceux tombant prés du milieu ne sont pas différencies, ceci explique la variabilité des interprétations statistiques que l'on peut faire d'un histogramme suivant le choix de l'origine et des classes. Pour des densités raisonnablement lisses, l'histogramme apparaît donc comme un estimateur sévérement limités.
Dans ce chapitre noun allons présenté une méthode plus robuste que la méthode par histogramme est la méthode de noyau.

1.3 Estimation de la densité de probabilité par la méthode du noyau

Il s'agit de l'estimateur le plus populaire. Il est adapté aux variables aléatoires continues.
Soit $X_1, X_2, ..., X_n$ un échantillon de variables aléatoires indépendantes et identiquement distribuées de fonction de répartition F et d'une densité f.
L'estimateur à noyau de densité, notée $\hat{f}(x)$ est définie par

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} k(\frac{x - X_i}{h}). \tag{1.10}$$

Où k est appelé fonction de poids ou noyau, et h est appelé paramétre de lissage ou fenêtre.

1.3.1 Noyaux usuels :

Noyaux	$K(u)$
Uniforme	$\frac{1}{2}, \lvert u\rvert \leq 1$
Normal	$\frac{1}{\sqrt{2\pi}} \exp(\frac{-u^2}{2}) u \in \mathbb{R}$
Triangulaire	$(1 - \lvert x\rvert), \lvert u\rvert \leq 1$
Epanechncov	$\frac{3}{4\sqrt{5}}(1 - (\frac{4}{5})^2), \lvert u\rvert \leq \sqrt{5}$

1.3.2 Propriétés d'un estimateur à noyau

- $\int_{\mathbb{R}} k(u)du = 1$ et $k(u) \geq 0$.
- L'estimateur à noyau est une fonction de densité.
- $\hat{f}$ a les même propriétés de continuité et de différentiabilité que k :

Si k est continue, $\hat{f}$ sera une fonction continue.

Si k est différentiable, $\hat{f}$ sera une fonction différentiable.

Si k peut prendre des valeurs négatives, alors $\hat{f}$ pourra aussi prendre des valeurs négatives.

1.4 Expressions asymptotique du biais et de la variance

une approximation asymptotique de l'espérence de l'estimateur $\hat{f}(x)$ est donnés sous conditions suivantes sur f, h et k

1. La dérivé seconde $f''(x)$ est continue, de carré intégrable et monotone sur $(-\infty, -M)$ et $(M, +\infty)$ pour $M > 0$;
2. $\lim_{n\to+\infty} h = 0$ et $\lim_{n\to+\infty} nh = 0$;
3. Pour que $\hat{f}(x)$ soit une densité, on suppose que $k(u) \geq 0$ et $\int k(u)du = 1$. La fonction noyau est supposée être symétrique autour de zéro, c.à.d $\int uk(u)du = 0$ et posséde un moment d'ordre 2 fini, c.à.d, $\int u^2 k(u)du < \infty$

Nous avons établi que

$$\mathbb{B}iais(\hat{f}(x)) = \frac{h^2}{2} f''(x)\mu_2 + O(h^2) \tag{1.11}$$

$$\mathbb{V}ar(\hat{f}_h(x)) = \frac{f(x)}{n} R(k) + O(\frac{1}{nh}) \tag{1.12}$$

où $\mu_2 = \int k(u)u^2 du$ et $R(g) = \int g^2(u)du$ pour une fonction g de carré intégrable.

Il faut donc essayer de choisir un h qui fasse un compromis entre le $biais^2$ et la variance. Les expressions asymptotiques du biais et de la variance nous permetent de trouver des expressions asymptotiques pour le MSE et le $MISE$.

Ces expressions ont été obtenues sous la condition (3) sur k et en supposant que la densité de probabilité f avait toutes les dérivéés (continues) nécessaires.

On peut obtenir facilement les approximations asymptotiques suivantes pour le MSE et le $MISE$.

$$MSE(\hat{f}_h(x)) = \frac{h^4}{4}(f''(x))^2\mu_2^2 + \frac{f(x)}{nh} R(k) + O(\frac{1}{nh}) + O(h^4). \tag{1.13}$$

$$MISE((\hat{f}_h(x))) = \frac{h^4}{4}\mu_2^2 \int_{\mathbb{R}} (f''(x))^2 dx + \frac{1}{nh} R(k) + O(h^4 + \frac{1}{nh}) \tag{1.14}$$

Sous des conditions appropriées d'intégrabilité de f et ses dérivées.

On note l'approximation asymptotique de le MSE par

$$AMSE(\hat{f}_h(x)) = \frac{h^4}{4}(f''(x))^2\mu_2^2 + \frac{f(x)}{nh} R(k). \tag{1.15}$$

et l'approximation asymptotique de le $MISE$ par

$$AMISE(\hat{f}_h(x)) = \frac{h^4}{4}\mu_2^2 \int_{\mathbb{R}} (f''(x))^2 dx + \frac{R(k)}{nh} \tag{1.16}$$

1.5 Choix théoriques optimaux du paramétre de lissage

Pour le paramétre de lissage on fait la distinction entre h paramétre de lissage constant (ou global), et $h(x)$ paramétre de lissage variable (local).

Ces choix différents du paramétre de lissage résultent en les estimateurs à noyau suivants :

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} k(\frac{x - X_i}{h}).$$

$$\hat{f}_h(x) = \frac{1}{nh(x)}\sum_{i=1}^{n} k(\frac{x - X_i}{h(x)}).$$

Nous allons décrire des choix théoriques optimaux des paramétres de lissage h et $h(x)$.
Un critére approprié pour sélectionner un paramétre de lissage constant h et le $MISE$. Le paramétres de lissage optimal est la valeur de h qui minimise le $MISE$. Notons cette valeur par h_{MISE}.
Une approximation asymptotique de h_{MISE} est donnée par h_{AMISE}, la valeur de h qui minimise $AMISE\{\hat{f}_n(.)\}$. Il est facile de vérifie que

$$h_{AMISE} = \{\frac{R(k)}{\mu^2 R(f'')}\}^{1/5} n^{-1/5}$$

Et

$$h_{MISE} \approx \{\frac{R(k)}{\mu^2 R(f'')}\}^{1/5} n^{-1/5},$$

c est-à-dire

$$\lim_{n\to\infty} \frac{h_{MISE}}{h_{AMISE}} = 1$$

Un critère approprié pour sélectionner un paramétre de lissage variable (local) $h(x)$ est la mesure de performance locale MSE $f_{n,L(x)}$. Nous introduisons les notations suivantes :

$$h_{MSE} = argmin_h MSE(\hat{f}_{n,L(x)}),$$

et

$$h_{AMSE} = argmin_h AMSE(\hat{f}_{n,L(x)}),$$

sous condition que $f''(x) \neq 0$. Les choix h_{AMISE} et $h_{AMSE}(x)$ sont des choix théoriques, qui ne sont pas utilisables en pratique car il dépendent des quantités inconnues f et f''. En substituant h_{AMISE} dans l'expression de l'AMISE, on montre que pour l'estimateur à noyau

$$n^{4/5} AMISE \hat{f}_{h_{AMISE}} = O(1)$$

1.6 Estimation de la densité de probabilité par des fonctions orthogonales

Dans cette section nous nous intéressons au probléme d'estimation de la densité de probabilité par des séries orthogonales introduite par Cencov [7]

1.6.1 Principe de la méthode

Soit $X_1, X_2, ..., X_n$ une suite des variables aléatoires indépendantes et identiquement distribuées de densité de probabilité f sur $\mathbb{R}$.
Il s'agit d'estimer f à partir des observations $X_1, X_2, ..., X_n$.
Pour cela on suppose que :

- L'espace de Hilbert L^2 est de dimension infinie ;
- $\{e_k, k \in \mathbb{N}\}$ est une base orthogonale dans L^2 ;
- f $\in L^2$ tel que :

$$f(x) = \sum_{k=0}^{\infty} a_k e_k(x), x \in \mathbb{R}. \tag{1.17}$$

- Le développement à l'ordre d_n de $f(x)$ est :

$$f_{d_n}(x) = \sum_{k=0}^{d_n} a_k e_k(x), x \in \mathbb{R}. \tag{1.18}$$

  avec (d_n) est une suite d'entiers qui tend vers ∞ lorsque $n \to \infty$
- Avec $(a_k)_{k\in\mathbb{N}}$ sont les coefficients de Fourier associés à f donnés par :

$$a_k = \int_{\mathbb{R}} e_k(x) f(x) dx = \mathbb{E}[e_k(X)]. \tag{1.19}$$

  Pour estimer $f(x)$ dans L^2 on se propose de construire un estimateur sans biais de $\hat{f}_{d_n}(x)$. Par la methode des moments, on peut estimer les coefficients $\{a_k, k \in \mathbb{N}\}$ par :

$$\hat{a}_k = \frac{1}{n} \sum_{i=1}^{n} e_k(X_i). \tag{1.20}$$

- Ansi $f_{d_n}(x)$ peut être estimée par :

$$\hat{f}_{d_n}(x) = \sum_{k=0}^{d_n} \hat{a}_k e_k(x). \tag{1.21}$$

1.6.2 Propriètés statistiques de l'estimateur

a. Les coefficients $(\hat{a}_k)_{k=0,\dots,d_n}$ sont des estimateur sans biais de $(a_k)_{k=0,\dots,d_n}$.

En effet,

$$\begin{aligned}\mathbb{E}(\hat{a}_k) &= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} e_k(X_i)] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[e_k(X_i)] \\ &= \mathbb{E}[e_k(X)] \\ &= a_k. \end{aligned} \tag{1.22}$$

b. Le biais de $\hat{f}_{d_n}(x)$ est par définition :

$$\begin{aligned}\mathbb{B}iais(\hat{f}_{d_n}(x)) &= \mathbb{E}(\hat{f}_{d_n}(x)) - f(x) \\ &= \sum_{k=0}^{d_n} a_k e_k(x) - \sum_{k=0}^{\infty} a_k e_k(x) \\ &= -\sum_{k=d_n+1}^{\infty} a_k e_k(x). \end{aligned} \tag{1.23}$$

Ce qui implique que $\hat{f}_{d_n}(x)$ est un estimateur biaisé de $f(x)$.

c. L'erreur quadratique moyenne intégrée de l'estimateur est donnée par le théorème suivant :

Théorème 1.6.1. *(Kronmal-Tarter) [44]*

Si $\int f^2(x)dx < \infty$, *alors :*

$$MISE(\hat{f}_{d_n}(x)) = \int f^2(x)dx - \sum_{k=0}^{d_n} a_k^2 + \sum_{k=0}^{d_n}\mathbb{V}ar(\hat{a}_k). \tag{1.24}$$

1.6.3 Exemples

Dans cette partie, nous allons présenté quelques cas particuliers d'estimateurs basés sur les systémes orthogonaux.

a.Estimateur de Saadi et Adjabi [38]

La base proposé par Saadi et Adjabi est donnée par

$$e_k(x) = \frac{1}{\sqrt{2\pi}}(\cos(kx) + \sin(kx))1_{[-\pi,\pi]}(x), k = 0, 1, ... \tag{1.25}$$

L'estimateur de la densité associé à la base trigonométrique est alors de la forme suivant :

$$\hat{f}_{d_n}(x) = \frac{1}{4\pi n}\sum_{i=1}^{n}\left[\frac{\sin[\frac{(2d_n+1)(X_i-x)}{2}]}{\sin[\frac{X_i-x}{2}]} + \frac{\sin[\frac{(2d_n+1)(\frac{\pi}{2}-(X_i+x))}{2}]}{\sin[\frac{\frac{\pi}{2}-(X_i+x)}{2}]}\right]. \tag{1.26}$$

Propriétés statistiques de l'estimateur

Biais de l'estimateur

$$\mathbb{B}iais(\hat{f}_{d_n}(x)) = -\sum_{k=d_n+1}^{\infty} a_k e_k(x),$$

ce qui implique, finalement, que $\hat{f}_{d_n}(x)$ est un estimateur biaisé de $f(x)$.

Variance de l'estimateur

$$\begin{aligned}\mathbb{V}ar(\hat{f}_{d_n}(x)) &= \frac{d_n+1}{4\pi^2 n} + \frac{1}{2\pi\sqrt{2\pi}n}\sum_{k=0}^{d_n}\beta_{2k} - \frac{1}{2n\pi}\sum_{k=0}^{d_n}a_k^2 \\ &+ \frac{1}{4\pi n^2}\sum_{k=0}^{d_n}\sin(2kx) + \frac{1}{2n\pi\sqrt{2\pi}}\sum_{k=0}^{d_n}\sin(2kx)\beta_{2k} - \frac{1}{2\pi n}\sum_{k=0}^{d_n}a_k^2\sin(2kx) \\ &+ \sum_{k=0}^{d_n}\sum_{j=0}^{d_n}\frac{1}{n}[\frac{1}{\sqrt{2\pi}}\gamma_{k-j} + \frac{1}{\sqrt{2\pi}}\beta_{k+j} - a_k a_j]e_k(x)e_j(x).\end{aligned}$$

Erreur quadratique moyenne

$$\begin{aligned} MSE(\hat{f}_{d_n}(x)) &= \frac{d_n+1}{4\pi^2 n} + \frac{1}{2\pi\sqrt{2\pi}n}\sum_{k=0}^{d_n}\beta_{2k} - \frac{1}{2n\pi}\sum_{k=0}^{d_n}a_k^2 \\ &+ \frac{1}{4\pi n^2}\sum_{k=0}^{d_n}\sin(2kx) + \frac{1}{2n\pi\sqrt{2\pi}}\sum_{k=0}^{d_n}\sin(2kx)\beta_{2k} \\ &- \frac{1}{2\pi n}\sum_{k=0}^{d_n}a_k^2\sin(2kx) + \sum_{k=0}^{d_n}\sum_{j=0}^{d_n}\frac{1}{n}[\frac{1}{\sqrt{2\pi}}\gamma_{k-j} \\ &+ \frac{1}{\sqrt{2\pi}}\beta_{k+j} - a_k a_j]e_k(x)e_j(x) + [\sum_{k=d_{n+1}^{\infty}} a_k e_k(x)]^2. \end{aligned}$$

Erreur quadratique moyenne intégrée

$$MISE(\hat{f}_{d_n}(x)) = \int_{-\pi}^{\pi} f^2(x)dx + \sum_{k=0}^{d_n}[\frac{1}{2\pi n} + \frac{1}{\sqrt{2\pi}n}\beta_{2k} - \frac{n-1}{n}a_k^2].$$

b. L'estimateur associé aux fonctions d'Hermite [21]

Les fonctions d'Hermite sont données par les formules suivantes :

$$e_j(x) = (2^j j!\pi^{\frac{1}{2}})^{-\frac{1}{2}} Q_j(x)\exp(-\frac{x^2}{2}), x \in \mathbb{R}; j = 0, \ldots$$

où $Q_j(x)$ est le $j^{\text{ième}}$ polynôme d'Hermite défini par :

$$Q_j(x) = (-1)^j \exp(-x^2)\frac{d^j}{dx^j}\exp(x^2); x \in \mathbb{R}, \quad j = 0, \ldots$$

L'estimateur associé est donné par :

$$\hat{f}_{d_n}(x) = \frac{d_n+1}{2n}\sum_{i=1}^{n}[\frac{Q_{j+1}(X_i)Q_j(x) - Q_j(X_i)Q_{j+1}(x)}{X_i - x}].$$

Convergence en moyenne quadratique intégrée

Théorème 1.6.2. *(Devroy et Györfl) [10]*

Pour que $\mathbb{E}|\hat{f}_{d_n}(x) - f(x)|^2$ tend vers 0 pour tout f de L^2, il est nécéssaire et suffisant que $\frac{\sqrt{d_n}}{n} \to 0$; tend vers 0 lorsque n $\to \infty$

Convergence simple

Théorème 1.6.3. *(Bleuez et Bosq) [5]*

les conditions suivantes sont équivalentes :

- $\frac{\sqrt{d_n}}{n} \to 0$;
- $\hat{f}_{d_n}(x) \to f(x), x \in \mathbb{R}$;
- $\mathbb{E}|\hat{f}_{d_n}(x) - f(x)| \to f(x), x \in \mathbb{R}$;
- $\mathbb{E}|\hat{f}_{d_n}(x) - f(x)|^2 \to f(x), x \in \mathbb{R}$.

c. L'estimateur associé aux fonctions de Laguerre

Les fonctions de Laguerre sont données par les formules :

$$L_i(x) = [\frac{\Gamma(d_n+1)}{\Gamma(d_n+1+\alpha)} x^{-\alpha}\exp(x)]^{\frac{1}{2}} \frac{1}{i!}\frac{d^i}{dx^i}(x^{i+\alpha}\exp(-x)), i \geq 0, \alpha > 0, x > 0.$$

Où $\Gamma(.)$ est la fonction gamma d'Euler, définie pour tout réel positif a par $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$

L'estimateur de Laguerre associé est alors défini par :

$$\hat{f}_{d_n}(x) = \frac{\Gamma(d_n+1)}{n\Gamma(d_n+1+\alpha)} \sum_{i=1}^{n} [\frac{l_{j+1}(X_i)l_j(x) - l_j(X_i)l_{j+1}(x)}{X_i - x}].$$

d. L'estimateur associé à la base de Dirichlet [4]

Nous supposons que $I = [-\pi, \pi]$ un intervalle de $\mathbb{R}$ muni de la mesure de lebesgue et la base orthogonale est définie par :

$$e_0(x) = \frac{1}{\sqrt{2\pi}}, e_{2k}(x) = \frac{cos(kx)}{\sqrt{\pi}}, e_{2k+1}(x) = \frac{\sin(kx)}{\sqrt{\pi}}; k = 1, \ldots$$

pour d_n impair, l'estimateur de Dirichlet est donné par :

$$\hat{f}_{d_n}(x) \quad = \begin{cases} \frac{1}{2\pi n}\sum_{i=1}^{n} \frac{\sin[d_n\frac{(X_i-x)}{2}]}{\sin[\frac{X_i-x}{2}]} & \text{si } X_i \neq x \\ \frac{d_n}{2\pi} \quad \text{sinon.} \end{cases}$$

e. L'estimateur associé aux fonctions de Legendre

Les fonctions de Legendre sont définies par :

$$p_i(x) = \sqrt{\frac{2i+1}{2}}\frac{1}{2^i i!}\frac{d^i}{dx^i}((x^2-1)^i), x \in [-1.1], i \geq 0.$$

L'estimateur associé est alors défini par :

$$\hat{f}_{d_n}(x) = \frac{d_n+1}{n\sqrt{2d_n+1}\sqrt{2d_n+3}} \sum_{i=1}^{n} \frac{p_{d_n}(X_i)p_{d_n+1}(x) - p_{d_n+1}(X_i)p_{d_n}(x)}{x - X_i}.$$

1.7 Choix pratique de la base

Le choix de la base dépend d'abord du support de la densité à estimer. Si le support de f est un intervalle compact, on pourra choisir les fonctions trigonométriques ou les fonctions de Legendre. Sur $\mathbb{R}_+$, on pourra utiliser les fonctions de Laguerre ou les fonctions d'Hermite. Quand on ne possède aucune information sur le support de f on peut utiliser les fonctions d'Hermite. Les fonctions d'Hermite donnent de bons résultats au voisinage de la loi normale réduite puisque le premier élément de la base $e_0(x) = \pi^{-\frac{1}{2}} \exp(-\frac{x^2}{2})$ est la densité d'une variable aléatoire de loi normale centrée réduite. Au voisinage d'une loi normale quelconque on peut considerer des fonctions d' Hermite modifiées données par

$$e_j^1(x) = e_j(\frac{x-\overline{X}}{S_n}), j \in \mathbb{N}, \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S_n = [\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})]^{\frac{1}{2}}. \tag{1.27}$$

On voit qu'il l n'y a pas de solution évidente qui se dégage. En effet, les systèmes orthonormaux sont très variés et il n'existe pas de théorèmes qui permettrait de conseiller un système particulier.

1.7.1 Choix du paramétre de lissage

La base étant sopposé fixée, il reste à choisir le paramétre de lissage d_n. Pour cela,on cherche à minimiser l'esrreur quadratique moyenne intégrée $MISE(\hat{f}_{d_n}(x))$. Il existe plusieurs méthodes pour le choix du paramétre de lissage.Dans ce chapitre on va étudier les méthodes suivantes :

✓ La méthode de Kronmal-Tarter.

✓ La méthode de Bosq.

Méthode de Kronmal-Tarter

L'emploi de (1.21) pour estimer $f(x)$ n'est pas possible qu'après avoir déterminer le nombre optimum de terme d_n de la somme. Il est naturel de choisir d_n de sorte que l'er-

reur quadratique moyenne intégrée $MISE(\hat{f}_{d_n}(x))$ soit minimum. La règle adoptée pour déterminer la valeur optimum d_n repose sur l'algorithme suivant :
Apartir de $d_n = 1$ on augmente la valeur de d_n d'une unité jusqu'à ce que $MISE(\hat{f}_{d_n}(x))$ augmente on donne alors à d_n la valeur qui précède juste l'augmentation $MISE(\hat{f}_{d_n}(x))$. On ajoutera donc à la somme (1.21) le d_n terme si et seulement si

$$\Delta_{d_n} = MISE(\hat{f}_{d_n}(x)) - (\hat{f}_{d_n-1}(x)) \leq 0. \tag{1.28}$$

En tenant compte de de théorème (1.6.1), Δ_{d_n} se met sous la forme :

$$\begin{aligned}
\Delta_{d_n} &= MISE(\hat{f}_{d_n}(x)) - (\hat{f}_{d_n-1}(x)) \qquad (1.29)\\
&= \int f^2(x)dx + \sum_{k=0}^{d_n}[\mathbb{V}ar(\hat{a}_k) - a_k^2] - \int f^2(x)dx + \sum_{k=0}^{d_n-1}[\mathbb{V}ar(\hat{a}_k) - a_k^2]\\
&= \mathbb{V}ar(\hat{a}_{d_n}) - a_{d_n}^2\\
&= \mathbb{V}ar(\frac{1}{n}\sum_{i=1}^{n})e_{d_n}(X_i)) - a_{d_n}^2\\
&= \frac{1}{n}\mathbb{V}ar(e_{d_n}(X_i)) - a_{d_n}^2
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{n}\int e_{d_n}(x)f(x)dx - \frac{1}{n}a_{d_n}^2 - a_{d_n}^2 \qquad (1.30)\\
&= \frac{1}{n}[\int e_{d_n}(x)f(x)dx - (n+1)a_{d_n}^2]\\
&= \frac{n+1}{n}\mathbb{V}ar(e_{d_n}(X)) - \mathbb{E}(e_{d_n}(X))^2
\end{aligned}$$

posons alors

$$\theta_i = e_{d_n}(X_i), i = 1, ..., n, \bar{\theta} = \frac{1}{n}\sum_{i=1}^{n}\theta_i \tag{1.31}$$

On peut alors définir un estimateur symétrique sans biais de Δ_{d_n} donné par :

$$\hat{\Delta_{d_n}} = \frac{1}{n}[\frac{n+1}{n-1}\sum_{i=1}^{n}(\theta_i - \bar{\theta})^2 - \sum_{i=1}^{n}\theta_i^2]. \tag{1.32}$$

On se fixe maintenant un entier positif D, l'optimum d_n^* est alors de la forme :

$$d_n^* = \begin{cases} inf\{d_n, 1 \leq d_n \leq D\} & \hat{\Delta}_d > 0; \\ \text{D} & sinon. \end{cases}$$

Méthode de Bosq [3]

Bosq a proposé un nouveau estimateur de paramétre de lissage donné par :

$$\hat{d}_n = max\{j : 0 \leq j \leq d_n, |\hat{a}_j| \geq \gamma_n\}, \tag{1.33}$$

avec

$$\gamma_n = c\sqrt{\frac{logn}{n}}, c > 0.$$

Théorème 1.7.1. *(Bosq) [3]*

1. *Si* $\frac{d_n}{n} \longrightarrow 0$ *on a*

$$MISE(\hat{f}_{\hat{d}_n}(x)) \longrightarrow 0. \tag{1.34}$$

2. *Si* $\sum_n \mid a_j \mid < \infty$ *et* $\sum_n^{d_n} \exp[-\frac{n}{d_n^2}a] < \infty, a > 0,$

alors

$$\sup_{x \in E} \mid \hat{f}_{\hat{d}_n}(x) - f(x) \mid \underset{p.s}{\longrightarrow} 0. \tag{1.35}$$

1.8 Conclusion

Nous avons présenté dans ce chapitre les différentes méthodes d'estimation de la densité de probabilité : estimation par histogramme, estimation par la méthode de noyau et la méthode des fonctions orthogonales. Nous avons présenté quelque cas particuliéres d'estimateurs basés sur des systémes trigonométriques Saadi et Adjabi, Direchlet et ceux associés aux fonctions d'Hermit, de Laguerre et de legendre.

Chapitre 2

Estimation de la densité de propabilité par un systéme trigonométrique

2.1 Introduction

Sam Efromovich (2010) a introduit un estimateur de la desité de probabilité basé sur ce systéme trigonométrique sur $[0, 1]$. L'auteur a donné ses propriétés statistiques (biais, la variance, l'erreur quadratique moyenne et l'erreur quadratique moyenne intégrées).
D'autres résultats furent également obtenus par les auteurs Lagha et Adjabi (2016) qui ont introduit une classe de fonctions basée sur un systéme trigonométrique, pricipalement caractérisée par le fait que la taille de l'échantillon est aléatoire :
L'objectif de cette partie est d'étudier les propriétés statistiques asymptotiques de cet estimateur, la convergence en moyenne quadratique, la convergence en moyenne quadratique intégrée, la vitesse de convergence en moyenne quadratique et la vitesse de convergence en moyenne quadratique intégrée.

2.2 Construction de l'estimateur

Soit $(X_1, ..., X_n)$ un suite de variable aléatoires indépendantes et identiquement distribuées de densité de probabilité inconnue f sur $[0, 1]$. Il s'agit d'estimer $f(x)$ à partir des observations $(x_1, ..., x_n)$. Pour cela on considére que :
La base est donnée par :

$$e_k(x) = \begin{cases} 1, & \text{si k=0 ;} \\ \sqrt{2}\cos(\pi kx), & \text{si k=,1,2,...} \end{cases} \tag{2.1}$$

Montrons que (2.1) est une base orthogonale dans $[0, 1]$, c'est-à-dire :

$$\int_0^1 e_k(x)e_j(x)dx = \begin{cases} 1, & \text{si j = k ;} \\ 0, & \text{sinon.} \end{cases}$$

En effet, pour $k = j$, nous avaons :

$$\begin{aligned} \int_0^1 e_k(x)e_j(x)dx &= \int_0^1 e_k^2(x)dx \\ &= 2\int_0^1 \cos^2(\pi kx)dx \end{aligned} \tag{2.2}$$

On a les propriétés suivantes

$$\cos^2(a) = \frac{1}{2}(1 + \cos(2a)) \tag{2.3}$$

$$\sin^2(a) = \frac{1}{2}(1 - \cos(2a)) \tag{2.4}$$

$$\cos(a)\cos(b) = \frac{1}{2}[\cos(a + b) + \cos(a - b)] \tag{2.5}$$

En tenant compte, (2.3) on obtient :

$$\begin{aligned} \int_0^1 e_k(x)e_j(x)dx &= \int_0^1 (1 + \cos(2\pi kx))dx \\ &= \int_0^1 dx + \int_0^1 \cos(2\pi kx)dx \\ &= 1 \end{aligned} \tag{2.6}$$

Pour $k \neq j$,

$$\int_0^1 e_k(x)e_j(x)dx = 2\int_0^1 \cos(\pi kx)\cos(\pi jx)dx \tag{2.7}$$

D'aprés (2.5) l'égalité devient :

$$\begin{aligned}\int_0^1 e_k(x)e_j(x)dx &= \int_0^1 \cos(\pi kx + \pi jx)) + \cos(\pi kx - \pi jx))dx \\ &= \int_0^1 \cos(\pi kx + \pi jx))dx + \int_0^1 \cos(\pi kx - \pi jx))dx \\ &= 0 \end{aligned} \tag{2.8}$$

Ce qui prouve que la base est orthogonale dans $[0, 1]$, par conséquent, la densité de probabilité $f(x)$ peut se mettre sous la forme :

$$f(x) = \sum_{k=0}^{\infty} a_k e_k(x),\ x \in [0, 1]. \tag{2.9}$$

où

$$a_k = \int_0^1 e_k(x)f(x)dx = \mathbb{E}[e_k(X)], \tag{2.10}$$

Le développement à l'ordre d_n de $f(x)$ s'écrit comme suit :

$$f_{d_n}(x) = \sum_{k=0}^{d_n} a_k e_k(x), \tag{2.11}$$

avec (d_n) est une suite d'entiers tendent vers 0 lorsque n tend vers l'infini.

Avec la nouvelle base orthogornale, les coefficients de Fourier associés à la densité de probabilité $f(x)$ s'écrivent comme suit :

$$a_k = \begin{cases} 1, & \text{si k = 0 ;} \\ \sqrt{2}\int_0^1 \cos(\pi kx)f(x)dx, & \text{si k = 1,...} \end{cases} \tag{2.12}$$

Les coefficients a_k peuvent être estimés par :

$$\hat{a}_k = \frac{1}{n}\sum_{i=1}^{n} e_k(X_i), \text{si k=1,...} \tag{2.13}$$

L'estimateur $\hat{f}_{d_n}(x)$ de $f(x)$ est donnée par :

$$\hat{f}_{d_n}(x) = \sum_{k=0}^{d_n} \hat{a}_k e_k(x) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{d_n} e_k(X_i)e_k(x). \tag{2.14}$$

Qui peut être dévloppé comme suit :

$$\begin{aligned}
\hat{f}_{d_n}(x) &= \sum_{k=0}^{d_n} \hat{a}_k e_k(x) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{d_n} e_k(X_i)e_k(x) \\
&= \frac{1}{n}\sum_{i=1}^{n}(1+2\sum_{k=1}^{d_n}\cos(\pi kx)\cos(\pi kX_i)) \\
&= 1+\frac{2}{n}\sum_{i=1}^{n}\sum_{k=1}^{d_n}\cos(\pi kx)\cos(\pi kX_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}[\frac{1}{2}+\sum_{i=1}^{d_n}\cos(\pi k(X_i+x))+\frac{1}{2}+\sum_{i=1}^{d_n}\cos(\pi k(X_i-x))] \\
&= \frac{1}{n}\sum_{i=1}^{n}[\frac{\sin(\frac{\pi(2d_n+1)(X_i+x)}{2})}{2\sin(\frac{\pi(X_i+x)}{2})}+[\frac{\sin(\frac{\pi(2d_n+1)(X_i-x)}{2})}{2\sin(\frac{\pi(X_i-x)}{2})}] \\
&= \frac{1}{2n}\sum_{i=1}^{n}\sum_{i=1}^{n}[\frac{\sin(\frac{\pi(2d_n+1)(X_i+x)}{2})}{\sin(\frac{\pi(X_i+x)}{2})}+[\frac{\sin(\frac{\pi(2d_n+1)(X_i-x)}{2})}{\sin(\frac{\pi(X_i-x)}{2})}], x\in[0,1], X_i \neq \quad (2.15)
\end{aligned}$$

2.3 propriétés des coefficients de Fouries

Les propriétés statistiques et asymptotiques des $(\hat{a}_{k_{k\in\mathbb{N}}})$ seront très utiles pour établir les propriétés statistiques et asymptotiques de l'estimateur $\hat{f}_{d_n}(x)$ qui seront données dans la section suivante.

Théorème 2.3.1.

Soient $(\hat{a}_k))_{k\in\mathbb{N}}$ les estimateurs de coefficients de Fourier associés à f, alors

$$\mathbb{E}(\hat{a}_k) = \begin{cases} 1, & si\ k{=}0\,; \\ a_k, & sinon. \end{cases} \tag{2.16}$$

$$\mathbb{V}ar(\hat{a}_k) = \begin{cases} 0, & si\ k{=}0\,; \\ \frac{1}{n}+\frac{1}{n\sqrt{2}}a_{2k}-\frac{a_k^2}{n}, & sinon. \end{cases} \tag{2.17}$$

$$\mathbb{C}ov(a_k,a_j) = \frac{1}{n\sqrt{2}}a_{k+j}+\frac{1}{n\sqrt{2}}a_{k-j}-\frac{1}{n}a_k a_j \tag{2.18}$$

Démonstration.

D'après la définition de l'espérance mathématique, l'espérance de $(\hat{a}_k)_{k\in\mathbb{N}}$ est de la forme :

pour $k = 0$

$$\begin{aligned}
\mathbb{E}(\hat{a}_0) &= \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} e_0(X_i)) \\
&= \int_0^1 e_0(x)f(x)dx \\
&= \int_0^1 f(x)dx \\
&= 1. \qquad (2.19)
\end{aligned}$$

Pour $k \neq 0$

$$\begin{aligned}
\mathbb{E}(\hat{a}_k) &= \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} e_k(X_i)) \\
&= \mathbb{E}(\frac{\sqrt{2}}{n}\sum_{i=1}^{n} \cos(\pi k X_i)) \\
&= \sqrt{2}\mathbb{E}(\cos(\pi k X)) \\
&= \sqrt{2}\int_0^1 \cos(\pi k x)f(x)dx \\
&= a_k. \qquad (2.20)
\end{aligned}$$

La variance de $(\hat{a}_k)_{k\in\mathbb{N}}$ *peut être développée comme suit :*

$$\begin{aligned}
\mathbb{V}ar(\hat{a}_k) &= \mathbb{V}ar(\frac{1}{n}\sum_{i=1}^{n} e_k(X_i)) \\
&= \frac{1}{n}\mathbb{V}ar(e_k(X)) \\
&= \frac{1}{n}[\mathbb{E}(2\cos^2(\pi k X)) - (\mathbb{E}(\sqrt{2}\cos(\pi k X)))^2] \qquad (2.21)
\end{aligned}$$

En tenant compte, de la propriété (2.3) on obtient :

$$\begin{aligned}
\mathbb{V}ar(\hat{a}_k) &= \frac{1}{n} + \frac{1}{n}\mathbb{E}(\cos(2\pi k X)) - \frac{a_k^2}{n} \\
&= \frac{1}{n} + \frac{1}{n}\int_0^1 (\cos(2\pi k x)f(x))dx - \frac{a_k^2}{n} \\
&= \frac{1}{n} + \frac{1}{n\sqrt{2}}a_{2k} - \frac{a_k^2}{n}. \qquad (2.22)
\end{aligned}$$

La covariance est par définition :

$$\mathbb{C}ov(\hat{a}_k, \hat{a}_j) = \mathbb{E}(\hat{a}_k\hat{a}_j) - \mathbb{E}(\hat{a}_k)\mathbb{E}(\hat{a}_j), \tag{2.23}$$

qui peut s'exprimer aussi comme suit :

$$\begin{aligned}
\mathbb{C}ov(\hat{a}_k, \hat{a}_j) &= \mathbb{E}[(\frac{1}{n}\sum_{i=1}^{n} e_k(X_i))(\frac{1}{n}\sum_{i=1}^{n} e_j(X_i))] - a_k a_j \\
&= \frac{1}{n^2}[\mathbb{E}\sum_{i=1}^{n} e_k(X_i)e_j(X_i)] + \frac{1}{n^2}\mathbb{E}[\sum_{i=1}^{n}\sum_{\substack{l=1,\\ l\neq i}}^{n} e_k(X_i)e_j(X_l)] - a_k a_j \\
&= \frac{1}{n}\mathbb{E}(e_k(X)e_j(X)) + [\frac{n-1}{n}]\mathbb{E}(e_k(X))\mathbb{E}(e_j(X)) - a_k a_j \\
&= \frac{2}{n}\mathbb{E}(\cos(\pi kX)\cos(\pi jX)) + [\frac{n-1}{n}]a_k a_j - a_k a_j
\end{aligned} \tag{2.24}$$

D'aprés la propriéré (2.5) l'égalité devient :

$$\begin{aligned}
\mathbb{C}ov(\hat{a}_k, \hat{a}_j) &= \frac{1}{n}\mathbb{E}(\cos(k+j)\pi X)) + \frac{1}{n}\mathbb{E}(\cos(k-j)\pi X)) - \frac{1}{n}a_k a_j \\
&= \frac{1}{n\sqrt{2}}a_{k+j} + \frac{1}{n\sqrt{2}}a_{k-j} - \frac{1}{n}a_k a_j
\end{aligned} \tag{2.25}$$

Comportement asymptotique de la variance et de l'erreur quadratique des $(\hat{a}_k)_{k\in\mathbb{N}}$

Théorème 2.3.2.

Soient $(\hat{a}_k)_{k\in\mathbb{N}}$ les estimateurs de coefficients de Fourier associés à f , alors

$$\lim_{n\longrightarrow\infty} \mathbb{V}(\hat{a}_k) = 0. \tag{2.26}$$

Démonstration.

$$\lim_{n\longrightarrow\infty} \mathbb{V}(\hat{a}_k) = \lim_{n\longrightarrow\infty}(\frac{1}{n} + \frac{1}{n\sqrt{2}}a_{2k} - \frac{a_k^2}{n}) = 0.$$

Théorème 2.3.3.

Soient $(\hat{a}_k)_{k\in\mathbb{N}}$ les estimateurs de coefficients de Fourier associés à f , alors

$$\lim_{n\longrightarrow\infty} \mathbb{E} \mid \hat{a}_k - a_k \mid^2 = 0 \tag{2.27}$$

Démonstration.

$$\lim_{n\longrightarrow\infty} \mathbb{E} \mid \hat{a}_k - a_k \mid^2 = \lim_{n\longrightarrow\infty} \mathbb{V}(\hat{a}_k) = 0$$

Donc, finalement, on peut conclure que les estimateurs $(\hat{a}_k)$ sont asymptotiquemet convergent et vérifiant la convergence en moyenne quadratique.

2.4 Propriétés statistiques de l'estimateur

2.4.1 Biais de l'estimateur

Théorème 2.4.1.

Soit $\hat{f}_{d_n}(x)$ l'estimateur de $f(x)$ sur $[0,1]$, alors $\hat{f}_{d_n}(x)$ est un estimateur biaisé de $f(x)$.

Démonstration.

En tenant compte de (2.16) le biais de $\hat{f}_{d_n}(x)$ s'écrit comme suit :

$$\begin{aligned}
\mathbb{B}iais(\hat{f}_{d_n}(x)) &= \mathbb{E}(\hat{f}_{d_n}(x)) - f(x) \\
&= \mathbb{E}(\sum_{k=0}^{d_n} \hat{a}_k e_k(x)) - f(x) \\
&= \sum_{k=0}^{d_n} \mathbb{E}(\hat{a}_k) e_k(x) - f(x). \\
&= \sum_{k=0}^{d_n} a_k e_k(x) - \sum_{k=0}^{\infty} a_k e_k(x) \\
&= -\sum_{k=d_{n+1}}^{\infty} a_k e_k(x). \qquad (2.28)
\end{aligned}$$

ce qui implique, finalement, que $\hat{f}_{d_n}(x)$ est un estimateur biaisé de $f(x)$.

2.4.2 Variance de l'estimateur

Théorème 2.4.2.

Soit $\hat{f}_{d_n}(x)$ l'estimateur de $f(x)$ sur $[0,1]$, alors

$$\begin{aligned}\mathbb{V}ar(\hat{f}_{d_n}(x)) &= \frac{d_n}{2}+\sum_{k=1}^{d_n}\cos(2\pi kx)(\frac{1}{n}+\frac{1}{\sqrt{2}n}a_{2k}-\frac{1}{n}a_k^2)\\ &+ \sum_{k=1}^{d_n}\sum_{j=1}^{d_n}\frac{1}{n}[\frac{1}{\sqrt{2}}a_{k+j}\frac{1}{\sqrt{2}}a_{k-j}-a_ka_j]e_k(x)e_j(x)\end{aligned} \tag{2.29}$$

Démonstration.

La variance de $\hat{f}_{d_n}(x)$ est par définition :

$$\begin{aligned}\mathbb{V}ar(\hat{f}_{d_n}(x)) &= \mathbb{V}ar[\sum_{k=1}^{d_n}\hat{a}_ke_k(x)]\\ &= \sum_{k=1}^{d_n}e_k^2(x)\mathbb{V}ar(\hat{a}_k)+\sum_{k=1}^{d_n}\sum_{j=1}^{d_n}\mathbb{C}ov(\hat{a}_k,\hat{a}_j)e_k(x)e_j(x).\\ &= \sum_{k=1}^{d_n}(1+\cos(2\pi kx))(\frac{1}{n}+\frac{1}{n\sqrt{2}}a_{2k}-\frac{1}{n}a_k^2)\\ &+ \sum_{k=1}^{d_n}\sum_{j=1}^{d_n}\frac{1}{n}[\frac{1}{\sqrt{2}}a_{k+j}\frac{1}{\sqrt{2}}a_{k-j}-a_ka_j]e_k(x)e_j(x)\\ &= \frac{d_n}{2}+\sum_{k=1}^{d_n}\cos(2\pi kx)(\frac{1}{n}+\frac{1}{\sqrt{2}n}a_{2k}-\frac{1}{n}a_k^2)\\ &+ \sum_{k=1}^{d_n}\sum_{j=1}^{d_n}\frac{1}{n}[\frac{1}{\sqrt{2}}a_{k+j}\frac{1}{\sqrt{2}}a_{k-j}-a_ka_j]e_k(x)e_j(x)\end{aligned} \tag{2.30}$$

2.4.3 Erreur quadratique moyenne

Théorème 2.4.3.

Soit $\hat{f}_{d_n}(x)$ l'estimateur de $f(x)$ sur $[0,1]$, alors

$$\begin{aligned}MSE(\hat{f}_{d_n}(x)) &= \frac{d_n}{2}+\sum_{k=1}^{d_n}\cos(2\pi kx)(\frac{1}{n}+\frac{1}{\sqrt{2}n}a_{2k}-\frac{1}{n}a_k^2)\\ &+ \sum_{k=1}^{d_n}\sum_{j=1}^{d_n}\frac{1}{n}[\frac{1}{\sqrt{2}}a_{k+j}\frac{1}{\sqrt{2}}a_{k-j}-a_ka_j]e_k(x)e_j(x)+[\sum_{k=d_{n+1}}^{\infty}a_ke_k(x)]^2\end{aligned}$$

Démonstration.

En tenant compte des théorèmes (2.4.2), (2.4.1) et de la relation suivante :

$$MSE(\hat{f}_{d_n}(x)) = \mathbb{V}ar(\hat{f}_{d_n}(x)) + \mathbb{B}ias^2(\hat{f}_{d_n}(x)), \tag{2.31}$$

ce qui permet de déduire l'expression du $MSE(\hat{f}_{d_n}(x))$.

2.4.4 Erreur quadratique moyenne intégrée

Théorème 2.4.4.

Soit $\hat{f}_{d_n}(x)$ l'estimateur de $f(x)$ sur $[0,1]$, alors

$$MISE(\hat{f}_{d_n}(x)) = \int_0^1 f^2(x)dx + \sum_{k=0}^{d_n}[\frac{1}{n} + \frac{1}{n\sqrt{2}}a_{2k} - \frac{n+1}{n}a_k^2]. \tag{2.32}$$

Démonstration.

$$\begin{aligned} MISE(\hat{f}_{d_n}(x)) &= \int_0^1 MSE(\hat{f}_{d_n}(x))dx \\ &= \int_0^1 \mathbb{E}(\hat{f}_{d_n}(x) - f(x))^2 \\ &= \int_0^1 \mathbb{E}(\hat{f}_{d_n}(x))^2dx - 2\int_0^1 \mathbb{E}(\hat{f}_{d_n}(x))f(x)dx + \int_0^1 f^2(x)dx. \end{aligned} \tag{2.33}$$

Le premier terme de (2.33) peut s'écrire comme suit :

$$\int_0^1 \mathbb{E}(\hat{f}_{d_n}(x))^2dx = \int_0^1 \mathbb{V}ar(\hat{f}_{d_n}(x))dx + \int_0^1 (\mathbb{E}(\hat{f}_{d_n}(x)))^2dx. \tag{2.34}$$

En tenant compte de fait que la base est orthogonale dans $[0,1]$, alors le premier et le deuxième terme de (2.34) peuvent êtres développés comme suit :

$$\begin{aligned} \int_0^1 \mathbb{V}ar(\hat{f}_{d_n}(x)) &= \int_0^1 \mathbb{V}ar(\sum_{k=0}^{d_n} \hat{a}_k e_{k(x)})dx \\ &= \sum_{k=0}^{d_n} \mathbb{V}ar(\hat{a}_k) \int_0^1 e_k^2(x)dx \\ &= \sum_{k=0}^{d_n} \mathbb{V}ar(\hat{a}_k) \\ &= \sum_{k=0}^{d_n}[\frac{1}{n} + \frac{1}{n\sqrt{2}}a_{2k} - \frac{a_k^2}{n}], \end{aligned} \tag{2.35}$$

et

$$\begin{aligned}\int_0^1 (\mathbb{E}(\hat{f}_{d_n}(x))^2 dx &= \int_0^1 [\mathbb{E}(\sum_{k=0}^{d_n} \hat{a}_k e_k(x))]^2 dx \\ &= \int_0^1 (\sum_{k=0}^{d_n} a_k e_k(x))^2 dx \\ &= \sum_{k=0}^{d_n} a_k^2 \int_0^1 e_k^2(x) dx \\ &= \sum_{k=0}^{d_n} a_k^2. \end{aligned} \tag{2.36}$$

Par conséquent,

$$\int_0^1 \mathbb{E}(\hat{f}_{d_n}(x))^2 dx = \sum_{k=0}^{d_n} \mathbb{V}ar(\hat{a}_k) + \sum_{k=0}^{d_n} a_k^2. \tag{2.37}$$

Calculons maintenant le deuxième terme de (2.33),

$$\begin{aligned}-2\int_0^1 \mathbb{E}(\hat{f}_{d_n}(x)) f(x) dx &= -2\int_0^1 \sum_{k=0}^{d_n} a_k e_k(x) f(x) dx \\ &= -2\sum_{k=0}^{d_n} a_k \int_0^1 e_k(x) f(x) dx \\ &= -2\sum_{k=0}^{d_n} a_k^2. \end{aligned} \tag{2.38}$$

Finalement, la combinaison de (2.35), (2.36) et (5.2) permet d'écrire :

$$\begin{aligned}MISE(\hat{f}_{d_n}(x)) &= \sum_{k=0}^{d_n} \mathbb{V}ar(\hat{a}_k) + \sum_{k=0}^{d_n} a_k^2 - 2\sum_{k=0}^{d_n} a_k^2 + \int_0^1 f^2(x) dx \\ &= \int_0^1 f^2(x) dx + \sum_{k=0}^{d_n} [\frac{1}{n} + \frac{1}{n\sqrt{2}} a_{2k} - \frac{a_k^2}{n}] - \sum_{k=0}^{d_n} a_k^2 \\ &= \int_0^1 f^2(x) dx + \sum_{k=0}^{d_n} [\frac{1}{n} + \frac{1}{n\sqrt{2}} a_{2k} - \frac{n+1}{n} a_k^2]. \end{aligned} \tag{2.39}$$

2.5 Propriétés asymptotiques de l'estimateur

Nous donnons deux hypothèses trés importantes sur le paramètre de lissage d_n afin d'établir les differents modes de convergence.

$$d_n \longrightarrow \infty \quad \text{quand} \quad n \longrightarrow \infty, \tag{2.40}$$

et

$$d_n = o(\sqrt{n}). \tag{2.41}$$

Biais asymptotique

Nous montrons que $\hat{f}_{d_n}(x)$ est asymptotiquement sans biais.

Théorème 2.5.1.

Si $d_n \longrightarrow \infty$ quand $n \longrightarrow \infty$, l'estimateur $\hat{f}_{d_n}(x)$ est asymptotiquement sans biais.

Démonstration.

$$\begin{aligned} \lim_{n\longrightarrow\infty} \mathbb{B}iais(\hat{f}_{d_n}(x)) &= \lim_{n\longrightarrow\infty} (\mathbb{E}(\hat{f}_{d_n}(x)) - f(x) \\ &= -\lim_{n\longrightarrow\infty} \sum_{k=d_n+1}^{\infty} a_k e_k(x) \\ &= 0, \end{aligned} \tag{2.42}$$

car les coefficients $(a_k)_{k\in\mathbb{N}}$ tendent vers 0 quand n tend vers ∞. Ce qui prouve que $\hat{f}_{d_n}(x)$ est asymptiquement sans biais.

Variance Asymptotique

Théorème 2.5.2.

Si d_n est d'ordre inférieur à $\sqrt{n}$ ie $(d_n = o(\sqrt{n}))$, alors

$$\lim_{n\longrightarrow\infty} \mathbb{V}ar(\hat{f}_{d_n}(x)) = 0. \tag{2.43}$$

Démonstration.

La variance de $\hat{f}_{d_n}(x)$ peut se mettre sous la forme suivante :

$$\begin{aligned} \mathbb{V}ar(\hat{f}_{d_n}(x)) &= \mathbb{V}ar[\frac{1}{2n}\sum_{i=1}^{n}[\frac{\sin[\frac{\pi(2d_n+1)(X_i+x)}{2}]}{\sin[\frac{\pi(X_i+x)}{2}]} + \frac{\sin[\frac{\pi(2d_n+1)(X_i-x))}{2}]}{\sin[\frac{(\pi(X_i+x))}{2}]}] \\ &= \frac{1}{4n}\mathbb{V}ar[\frac{\sin[\frac{\pi(2d_n+1)(X+x)}{2}]}{\sin[\frac{\pi(X+x)}{2}]} + \frac{\sin[\frac{\pi}{(2d_n+1)(X-x))}2]}{\sin[\frac{(\pi(X-x))}{2}]}]. \end{aligned} \tag{2.44}$$

Par définition de la variance, on sait que

$$\mathbb{V}ar(X) \leq \mathbb{E}(X^2). \tag{2.45}$$

On obtient

$$\mathbb{V}ar[\frac{\sin[\frac{\pi(2d_n+1)(X+x)}{2}]}{\sin[\frac{\pi(X+x)}{2}]} + \frac{\sin[\frac{\pi(2d_n+1)(X-x))}{2}]}{\sin[\frac{(\pi(X+x))}{2}]}] \leq \mathbb{E}[\frac{\sin[\frac{\pi(2d_n+1)(X+x)}{2}]}{\sin[\frac{\pi(X+x)}{2}]} + \frac{\sin[\frac{\pi(2d_n+1)(X-x))}{2}]}{\sin[\frac{\pi(X-x)}{2}]}]^2.$$

Par ailleurs,

$$\begin{aligned}
\mathbb{E}[\frac{\sin[\frac{\pi(2d_n+1)(X+x)}{2}]}{\sin[\frac{\pi(X+x)}{2}]} + \frac{\sin[\frac{\pi(2d_n+1)(X-x))}{2}]}{\sin[\frac{\pi(X-x))}{2}]}]^2 &= \mathbb{E}[\frac{\sin[\frac{\pi(2d_n+1)(X+x)}{2}]}{\sin[\frac{\pi(X+x)}{2}]}]^2 + \mathbb{E}[\frac{\sin[\frac{\pi(2d_n+1)(X-x)}{2}]}{\sin[\frac{\pi(X-x)}{2}]}]^2 \\
&+ 2\mathbb{E}[\frac{\sin[\frac{\pi(2d_n+1)(X+x)}{2}]}{\sin[\frac{\pi(X+x)}{2}]}\frac{\sin[\frac{\pi(2d_n+1)(X-x)}{2}]}{\sin[\frac{\pi(X-x)}{2}]}] \\
&= \int_0^1 f(y)[\frac{\sin[\frac{\pi(2d_n+1)(y+x)}{2}]}{\sin[\frac{\pi(y+x)}{2}]}]^2 dy \\
&+ \int_0^1 f(y)[\frac{\sin[\frac{\pi(2d_n+1)(y-x)}{2}]}{\sin[\frac{\pi(y-x)}{2}]}]^2 dy \\
&+ 2\int_0^1 f(y)[\frac{\sin[\frac{\pi(2d_n+1)(y+x)}{2}]}{\sin[\frac{\pi(y+x)}{2}]}\frac{\sin[\frac{\pi(2d_n+1)(y-x)}{2}]}{\sin[\frac{\pi(y-x)}{2}]}dy.
\end{aligned}$$

D'autre part, on a

$$\frac{\sin kx}{\sin x} \leq k.$$

Il vient donc :

$$\begin{aligned}
\mathbb{V}ar(\hat{f}_{d_n}(x)) &\leq \frac{1}{4n}[(2d_n+1)^2\int_0^1 f(y)dy + (2d_n+1)^2\int_0^1 f(y)dy + 2(2d_n+1)^2\int_0^1 f(y)dy] \\
&\leq \frac{(2d_n+1)^2}{n}
\end{aligned} \tag{2.46}$$

Àprésent, il est possible d'établire la convergence en moyenne quadratique. Qui sera l'objectif de théorème suivant.

Convergence en moyenne quadratique

Théorème 2.5.3.

Si d_n est d'ordre inférieur à $\sqrt{n}$ ie $(d_n = o(\sqrt{n}))$, *et* $d_n \longrightarrow \infty$ *quand* $n \longrightarrow \infty$. *L'estimateur* $\hat{f}_{d_n}(x)$ *est alors un estimateur convergent en moyenne quadratique. C'est-à-dire :*

$$\lim_{n\longrightarrow\infty} MSE(\hat{f}_{d_n}(x)) = \lim_{n\longrightarrow\infty} \mathbb{E}[\hat{f}_{d_n}(x) - f(x)]^2 dx = 0. \tag{2.47}$$

Démonstration.

En effet,

$$\mathbb{E}[\hat{f}_{d_n}(x) - f(x)]^2 \quad = \quad \mathbb{V}ar(\hat{f}(x)) + \mathbb{B}iais^2(\hat{f}(x)) \tag{2.48}$$

tenant compte de (2.42) et (2.43) on obtient

$$\lim_{n\longrightarrow\infty} MSE(\hat{f}_{d_n}(x)) = 0. \tag{2.49}$$

Convergence en moyenne quadratique intégrée

Théorème 2.5.4.

Si d_n est d'ordre inférieur à $\sqrt{n}$ ie $(d_n = o(\sqrt{n}))$ *et* $d_n \longrightarrow \infty$ *quand* $n \longrightarrow \infty$. *L'estimateur* $\hat{f}_{d_n}(x)$ *est alors un estimateur convergent en moyenne quadratique intégrée. C'est-à-dire :*

$$\lim_{n\longrightarrow\infty} MISE(\hat{f}_{d_n}(x)) = \lim_{n\longrightarrow\infty} \int_0^1 \mathbb{E}[\hat{f}_{d_n}(x) - f(x)]^2 dx = 0.$$

Démonstration.

D'après le théorème (2.5.3)

$$\lim_{n\longrightarrow\infty} \mathbb{E}[\hat{f}_{d_n}(x) - f(x)]^2 = 0,$$

la convergence étant uniforme :

$$\lim_{n\longrightarrow\infty} MISE(\hat{f}_{d_n}(x)) = \lim_{n\longrightarrow\infty} \int_0^1 \mathbb{E}[\hat{f}_{d_n}(x) - f(x)]^2 dx = 0.$$

L'intérêt majeur de ce théorème est la condition $d_n = o(\sqrt{n})$ *qui implique que le nombre de terme* d_n *qui intervient dans* $\hat{f}_{d_n}(x)$ *doit toujours être relativement petit par rapport à la taille de l'échantillon.*

Vitesse de convergence en moyenne quadratique

Théorème 2.5.5.

Si

$$(\sum_{k=d_n+1}^{\infty} \mid a_k k^{\alpha} \mid^q)^{\frac{2}{q}} < \infty, \alpha > 0 \quad q > 0, \tag{2.50}$$

et

$$d_n = (\gamma \log n)^{\frac{1}{2}}, \quad \gamma \in]0, \frac{1}{2}[. \tag{2.51}$$

Alors

$$\mathbb{E} \mid \hat{f}_{d_n}(x) - f(x) \mid^2 = O((\log n)^{-\delta}), \tag{2.52}$$

avec $\delta > 0$.

Démonstration.

L'erreur quadratique moyenne associée à $\hat{f}_{d_n}(x)$ est par définition :

$$\mathbb{E} \mid \hat{f}_{d_n}(x) - f(x) \mid^2 = \mathbb{V}ar(\hat{f}_{d_n}(x)) + \mathbb{B}iais^2(\hat{f}_{d_n}(x)), \tag{2.53}$$

le biasi de $\hat{f}_{d_n}(x)$ peut également se mettre sous la forme :

$$\begin{aligned} \mid \mathbb{B}iais(\hat{f}_{d_n}(x)) \mid &= \mid \mathbb{E}(\hat{f}_{d_n}(x)) - f(x) \mid \\ &= \mid \mathbb{E}(\sum_{k=0}^{d_n} \hat{a}_k e_k(x)) - \sum_{k=0}^{\infty} a_k e_k(x) \mid \\ &= \mid \sum_{k=0}^{d_n} \mathbb{E}(\hat{a}_k) e_k(x) - \sum_{k=0}^{\infty} a_k e_k(x) \mid \\ &= \mid \sum_{k=0}^{d_n} a_k e_k(x) - \sum_{k=0}^{\infty} a_k e_k(x) \mid \\ &= \mid \sum_{k=d_n+1}^{\infty} a_k e_k(x) \mid . \end{aligned} \tag{2.54}$$

Or,

$$\begin{aligned} \mid \sum_{k=d_n+1}^{\infty} a_k e_k(x) \mid &\leq \sum_{k=d_n+1}^{\infty} \mid a_k \mid\mid e_k(x) \mid \\ &\leq \sup_{x\in[0,1]} \mid e_k(x) \mid \sum_{k=d_n+1}^{\infty} \mid a_k \mid . \end{aligned} \tag{2.55}$$

De plus :

$$\begin{aligned} \sup_{x\in[0,1]} \mid e_k(x) \mid \sum_{k=d_n+1}^{\infty} \mid a_k \mid &= \frac{2}{\sqrt{2\pi}} \sum_{k=d_n+1}^{\infty} \mid a_k \mid \\ &= 2\sqrt{2} \sum_{k=d_n+1}^{\infty} \mid a_k k^{\alpha} \mid \frac{1}{k^{\alpha}}. \end{aligned} \tag{2.56}$$

D'autre part,

$$\frac{2}{\sqrt{2\pi}}\sum_{k=d_n+1}^{\infty} \mid a_k k^{\alpha} \mid \frac{1}{k^{\alpha}} \leq \frac{2}{\sqrt{2\pi}}(\sum_{k=d_n+1}^{\infty} \mid a_k k^{\alpha} \mid^q)^{\frac{1}{q}}(\sum_{k=d_n+1}^{\infty}\frac{1}{k^{p\alpha}})^{\frac{1}{p}} = 2\sqrt{2}M(\sum_{k=d_n+1}^{\infty}\frac{1}{k^{p\alpha}})^{\frac{1}{p}},$$

avec

$$M = (\sum_{k=d_n+1}^{\infty} \mid a_k k^{\alpha} \mid^q)^{\frac{1}{q}}, p > 0, \frac{1}{p}+\frac{1}{q} = 1. \tag{2.57}$$

Nous avons :

$$\sum_{k=d_n+1}^{\infty}\frac{1}{k^{p\alpha}} \leq \int_{d_n}^{\infty}\frac{1}{x^{p\alpha}}dx = \frac{1}{\alpha p-1}(\frac{1}{d_n})^{\alpha p-1}. \tag{2.58}$$

Il vient donc :

$$\mid \sum_{k=d_n+1}^{\infty} a_k e_k(x) \mid^2 \leq \frac{2}{\pi}M^2\frac{1}{(\alpha p-1)^{\frac{2}{p}}}(\frac{1}{d_n})^{2\alpha-\frac{2}{p}}. \tag{2.59}$$

Posons

$$c = \frac{2}{\pi}M^2[\frac{1}{\alpha p-1}]^{\frac{2}{p}}, \alpha' = 2\alpha-\frac{2}{p}. \tag{2.60}$$

Donc

$$\mid \sum_{k=d_n+1}^{\infty} a_k e_k(x) \mid^2 \leq c(\frac{1}{d_n})^{\alpha'}. \tag{2.61}$$

Par conséquent,

$$\mid \mathbb{B}iais(\hat{f}_{d_n}(x)) \mid^2 = O((\frac{1}{d_n})^{\alpha'}). \tag{2.62}$$

De plus d'après, on a :

$$\mathbb{V}ar(\hat{f}_{d_n}(x)) = O(\frac{d_n^2}{n}). \tag{2.63}$$

(2.73) permet d'écrire :

$$\mathbb{E} \mid \hat{f}_{d_n}(x) - f(x) \mid^2 = O(\frac{d_n^2}{n}) + O((\frac{1}{d_n})^{\alpha'}). \tag{2.64}$$

Posons

$$d_n = (\gamma \log n)^{\frac{1}{2}}, \gamma \in]0, \frac{1}{2}[, \tag{2.65}$$

il vient donc :

$$\begin{aligned} \frac{d_n^2}{n} &\leq \frac{1}{n}\exp(\log d_n^2) \\ &\leq \frac{1}{n}\exp(\gamma \log n). \end{aligned} \tag{2.66}$$

D'autre part :

$$\begin{aligned}\frac{1}{n}\exp(\gamma\log n) &= \frac{1}{n}\exp(\log n^{\gamma}) \\ &= \frac{1}{n}n^{\gamma} = \frac{1}{n^{1-\gamma}}.\end{aligned} \tag{2.67}$$

Finalement, on peut conclure que :

$$\begin{aligned}\mathbb{E}(\hat{f}_{d_n}(x) - f(x))^2 &= O(\frac{1}{n^{1-\gamma}}) + O((\frac{1}{\log n})^{\alpha'}) \\ &= O((\log n)^{-\delta}),\end{aligned} \tag{2.68}$$

avec

$$\delta = \min(1-\gamma, \alpha'). \tag{2.69}$$

Le théorème suivant nous donne la vitesse de convergence en moyenne quadratique intégrée.

2.5.1 Vitesse de convergence en moyenne quadratique intégrée

Théorème 2.5.6. .

Si

$$\sum_{k=d_n+1}^{\infty} a_k^2 = O(d_n^{-r}), \quad r > 0, \tag{2.70}$$

et

$$d_n = (\alpha\log n), \quad 0 < \alpha < 1. \tag{2.71}$$

Alors

$$MISE(\hat{f}_{d_n}(x)) = \mathbb{E}\int_{-\pi}^{\pi}(\hat{f}_{d_n}(x) - f(x))^2 dx = O((\log n)^{-\delta}), \tag{2.72}$$

avec

$$\delta = \min(1-\alpha, r). \tag{2.73}$$

Démonstration.

L'erreur quadratique moyenne intégrée associée à $\hat{f}_{d_n}(x)$ est donnée par :

$$MISE(\hat{f}_{d_n}(x)) = \int_0^1 f^2(x)dx + \sum_{k=0}^{d_n}[\mathbb{V}ar(\hat{a}_k) - a_k^2].$$

La somme des variances des $(\hat{a}_k)_{\{k\in\mathbb{N}\}}$ peut être s'exprimée comme suit :

$$\begin{aligned}\sum_{k=0}^{d_n}\mathbb{V}ar(\hat{a}_k) &= \sum_{k=0}^{d_n}\mathbb{V}ar(\frac{1}{n}\sum_{i=1}^{n}e_k(X_i))\\ &= \frac{1}{n}\sum_{k=0}^{d_n}\mathbb{V}ar(e_k(X))\\ &= \frac{1}{n}\sum_{k=0}^{d_n}[\mathbb{E}(e_k(X))^2-(\mathbb{E}(e_k(X)))^2]\\ &= \frac{1}{n}\sum_{k=0}^{d_n}\int_0^1 f(x)(e_k(x))^2dx-\frac{1}{n}\sum_{k=0}^{d_n}a_k^2, \qquad (2.74)\end{aligned}$$

et on a

$$\int_0^1 f^2(x)dx=\sum_{k=0}^{\infty}a_k^2\int_0^1 e_k^2(x)dx=\sum_{k=0}^{\infty}a_k^2=\sum_{k=0}^{d_n}a_k^2+\sum_{k=d_n+1}^{\infty}a_k^2. \qquad (2.75)$$

On aura :

$$\begin{aligned}MISE(\hat{f}_{d_n}(x)) &= \frac{1}{n}\sum_{k=0}^{d_n}\int_0^1 f(x)(e_k(x))^2dx+\sum_{k=0}^{d_n}a_k^2+\sum_{k=d_n+1}^{\infty}a_k^2-\frac{1}{n}\sum_{k=0}^{d_n}a_k^2-\sum_{k=0}^{d_n}a_k^2\\ &= \frac{1}{n}\sum_{k=0}^{d_n}\int_0^1 f(x)(e_k(x))^2dx+\sum_{k=d_n+1}^{\infty}a_k^2-\frac{1}{n}\int_0^1(\sum_{k=0}^{d_n}a_ke_k(x))^2dx\\ &= \frac{1}{n}\sum_{k=0}^{d_n}\int_0^1 f(x)(e_k(x))^2dx-\frac{1}{n}\sum_{k=0}^{d_n}a_k^2+\sum_{k=d_n+1}^{\infty}a_k^2.\end{aligned}$$

Par ailleurs,

$$\begin{aligned}\frac{1}{n}\sum_{k=0}^{d_n}\int_0^1 f(x)e_k^2(x)dx-\frac{1}{n}\sum_{k=0}^{d_n}a_k^2+\sum_{k=d_n+1}^{\infty}a_k^2 &\le \frac{1}{n}\sum_{k=0}^{d_n}\int_0^1 f(t)e_k^2(x)dx+\sum_{k=d_n+1}^{\infty}a_k^2\\ &\le \frac{d_n+1}{n}\sup_{x\in[0,1]}e_k^2(x)\int_0^1 f(x)dx+\sum_{k=d_n+1}^{\infty}a_k^2.\end{aligned}$$

Compte tenant de (2.70), nous avons :

$$\frac{d_n+1}{n}\sup_{x\in[0,1]}e_k^2(x)\int_0^1 f(x)dx+\sum_{k=d_n+1}^{\infty}a_k^2=O(\frac{d_n}{n})+O(d_n^{-r}). \qquad (2.76)$$

D'autre part :

$$\begin{aligned}\frac{d_n}{n} &\le \frac{1}{n}\exp\log(d_n)\\ &\le \frac{1}{n}\exp(d_n)\\ &\le \frac{1}{n}\exp(\alpha\log n). \qquad (2.77)\end{aligned}$$

Or,

$$\begin{aligned}\frac{1}{n}\exp(\alpha\log n) &= \frac{1}{n}\exp(\log(n^{\alpha})) \\ &= \frac{1}{n^{1-\alpha}}. \end{aligned} \tag{2.78}$$

Par conséquent,

$$\begin{aligned} MISE(\hat{f}_{d_n}(x)) &= O(\frac{1}{n^{1-\alpha}}) + O((\alpha\log n)^{-r}) \\ &= O((\log n)^{-\delta}), \end{aligned} \tag{2.79}$$

avec $\delta = \min(1-\alpha, r)$.

2.6 Détermination du nombre optimum de termes par la méthode de Kronmal et Tarter

Pour déterminer la valeur optimum d_n^* selon la méthode de Kronmal et Tarter on procède comme suit :

À partir de $d_n = 1$ on augmente la valeur de d_n d'une unité jusqu'à ce que le $MISE$ augmente. On donne alors à d_n la valeur qui précède juste l'augmentation du $MISE$. On ajoutera à la somme (1.21) le $d_n^{\text{ième}}$ si et seulement si

$$\Delta_{d_n} = MISE(\hat{f}_{d_n}(x)) - MISE(\hat{f}_{d_n-1}(x)) \leq 0. \tag{2.80}$$

On a

$$\begin{aligned}\Delta_{d_n} &= \int_0^1 f^2(x)dx + \sum_{k=0}^{d_n}[\mathbb{V}ar(\hat{a}_k) - a_k^2] - \int_0^1 f^2(x)dx - \sum_{k=0}^{d_n-1}[\mathbb{V}ar(\hat{a}_k) - a_k^2] \\ &= \mathbb{V}ar(\hat{a}_{d_n}) - a_{d_n}^2. \end{aligned}$$

Il suffit donc de tester la quantité

$$\Delta_{d_n} = \mathbb{V}ar(\hat{a}_{d_n}) - a_{d_n}^2. \tag{2.81}$$

compte tenant de (1.29) Δ_{d_n} *s'écrit comme suit :*

$$\Delta_{d_n} = \mathbb{V}ar(\hat{a}_{d_n}) - a_{d_n}^2 \tag{2.82}$$

$$= \frac{n+1}{n}\mathbb{V}ar(e_{d_n}(X)) - \mathbb{E}(e_{d_n}(X))^2. \tag{2.83}$$

L'estimateur sans biais de Δ_{d_n} est de la forme :

$$\hat{\Delta}_{d_n} = \frac{1}{n}[\frac{n+1}{n-1}\sum_{i=1}^{n}(\theta_i - \overline{\theta})^2 - \sum_{i=1}^{n}\theta_i^2]. \tag{2.84}$$

Avec

$$\theta_i = e_d(X_i) = \sqrt{2}(\cos(k\pi X_i)), i = 1, ..., n, \tag{2.85}$$

et

$$\overline{\theta} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{2}(\cos(k\pi X_i)). \tag{2.86}$$

On se fixe maintenant un entier positif D, l'optimal d_n^* s'écrit :

$$d_n^* = \begin{cases} \inf\{d_n : 1 \leq d_n \leq D\}; \hat{\Delta}_{d_n} > 0 \\ D \qquad\qquad\qquad\qquad sinon. \end{cases}$$

2.7 Conclusion

Nous avons préenté dans ce chapitre un estimateur de densité de probabilité basé sur un systéme trigonométrique, nous avons donné ces propriétes statistiques (biais, variance, l'erreur quadratique moyenne, l'erreur quadratique moyenne intégrée), ces propriétés asymptotique et aussi , la vitesse de convergence en moyenne quadratique et la vitesse de convergence en moyenne quadratique intégrée.

Chapitre 3

Estimation non paramétrique de la fonction de répartition

La fonction de répartition F caractérise la loi de probabilité d'une variable aléatoire. Elle permet d'avoir un aperçu des principales caractéristiques de la distribution et c'est en terme de comportement locale 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 les probabilités d'ensembles. l'objectif de ce chapitre est de présenté des différentes méthodes d'estimation non paramétrique de fonction de répartition.

3.1 Fonction de répartition empirique

Soit $X_1, \ldots, X_n$ un échantillon de variables aléatoires indépendantes et identiquement distribuées à valeurs dans $\mathbb{R}$ avec de fonction de répartition $F(x)$.
Un bon estimateur de F est la fonction de répartition empirique, notée F_n, définie par

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x), x \in \mathbb{R} \tag{3.1}$$

où I(A) est la fonction indicatrice de l'événement A donnée par :

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

3.1.1 Propriétés statistiques de l'estimateur

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

1. $D_n^+ = \sup_{x\in\mathbb{R}}(Z_n(x))$,
2. $D_n^- = \sup_{x\in\mathbb{R}}(-Z_n(x))$
3. $D_n = \sup_{x\in\mathbb{R}} \mid Z_n(x) \mid$.

Kolmogorov-Smirnov [23] introduisent et étudient ces trois statistiques et démontrent que leurs distribution ne dépend pas de F. en autre, à l'aide des théorèmes de Donsker [12] Doob (1949), 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 (1956) déterminent une borne de la forme

$$p(D_n^- > \lambda) = C\exp(-2\lambda^2),$$

où C est une constante indéterminée. L'ensemble de cette démarche et de ces résultats sont analysés dans Hennequin et Tortrat (1965). Pour des propriétés liées aux statistiques d'ordre et de rang on pourra consulter le livre de Caperaa et Van Cutsem (1988).
De nombreux auteurs essayent ensuite de trouver la meilleure constante C dans l'inégalité précédente. Devroye et Wise (1979), font peu à peu diminuer cette constante. Finalement, Massart (1990) démontre qu'il est possible de prendre $C = 1$ pourvu que

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

Il montre aussi que, quel que soit λ,

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

F_n est également l'estimateur non paramétrique du maximum de vraisemblance Kiefer-Wolfowitz (1956), Efron-ibshirani (1993) et, en général, une transformé $t(F)$ a pour estimateurs du maximum de vraisemblance $t(F)$. D'autre part, parmi les estimateurs sans biais de $F(x)$, $F_n(x)$ est également l'unique estimateur de variance minimale Lehmann (1983).

Cette dernière vaut par ailleurs $F(x)(1 - F(x))/n$. Lehmann (1983) élargit les recherches à une famille de fonction englobant F_n. Il étudie les fonctions de type

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

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

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

si et seulement si $\omega_n \to e_0$ avec

$$I_A(x) = \begin{cases} 1, & \text{si } x > 0\,; \\ 0, & \text{sinon.} \end{cases}$$

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

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

La fonction de répartition empirique ne tient pas compte d'une éventuelle information que nous pouvons avoir sur la fonction à estimer. Modarres (2002) utilise une possible symétrie pour bâtir un nouvel estimateur à partir de F_n :

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

et démontre que cet estimareur est l'estimateur du maximum de vraisemblance.

Remarque 3.1.1.

La fonction de répartition empirique a de bonne propriétés de convergence mais posséde certains inconvénients comme celui de ne pas prendre en compte ene éventuelle information supplémentaires ou bien le fait d'être une fonction en escalier.
Il existe des estimateurs qui sont préférables à la fonction de répartition empirique par exemple l'estimateur de noyau.

3.2 Estiamation non paramétrique de la fonction de répartition par la méthode du noyau

L'estimateur à noyau de la densité

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} k(\frac{x - X_i}{h})$$

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 (1954), Rosenblatt [37] et Parzen [32]. La littérateure 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 $\tilde{F}_h$ à noyau de F par

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

où

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

Ces propriétés sont connues depuis longtemps, par exemple sa convergence uniforme vers F avec f continue Nadaraya(1965), Yamato puis sans conditions sur f Singh-Gasser(1984) ou sa normalité asymptotique Nadaraya (1965) et Leadbetter (1972) . Winter démontre aussi qu'il vérifie la propriété de Chung-Smirnov, c'est-à-dire

$$\limsup_{n\to\infty} \{(\frac{2n}{loglogn})^{\frac{1}{2}} \sup_{-\infty} < t < \infty \mid \tilde{F}_n(t) - F(t) \mid\} \leq 1$$

avec probabilité 1. Azzalini (1981) trouve un expression asymptotique pour l'erreur quadratique moyenne ou MSE ($\mathbb{E}(\tilde{F}_n(x) - F(x))^2$) et détermine la fenêtre asymptotiquement optimale permettant d'avoir un MSE plus faible que pour F_n. Reiss (1981) 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

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

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

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

Falk(1983) 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(1983) 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 $\phi(k) = \int k(x)^2$ défini par Epanechnikov (1969) comme une mesure de la performance de noyau en estimation de la densité. Au sens de ϕ, le noyau d'Epanechnikov suivant

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

est le meilleur mais les noyaux gaussiens ou uniformes ont des performances trés proches. En utilisant le critère ψ le noyau d'Epanechnikov (1969) est alors de loin le meilleur des trois.
Falk (1983) montre ensuite que cette inefficacité relative s'applique aussi aux estimateurs des quantiles q_n De par rapport aux quantiles $\tilde{q}_n$ de $\tilde{F}_n$. Enfin, Golubev-Levit(1996) 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 $MISE$, le meilleur noyau est le noyau uniforme bien que les performance d'autre noyaux (Epanechnikov, normal, trian-

gulaire) ne soient, en pratique, que légèrement moins bonnes . Il est intéressant de noter que ce ne sera pas le meilleur noyau dans le cadre d'estimation de la densité.

3.3 Estimation par lissage local

Afin d'obtenir une estimation plus régulière de la fonction de répartition, Berlinet (1981) et Lejeune(1992) lisent la fonction de répartition dans différents espaces. Lejeune utilise 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. Ce résultat est ensuite élargi par Abdous, Berlinet et Huang(2004) qui proposent d'estimer differentes fonctionnelles $\phi(x, F)$ de la fonction F au point x. pour cela il substituent $\phi(x, F_n)$ à $\phi(x, F)$ et, dans le cas où $\phi(x, F)$ a r dérivées continues, choisissent de minimiser de critère suivant

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

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

$$K^{m,r}(u) = (\sum_{k=0}^{r} Q_k(u) \frac{d^m}{dw^m} Q_k(w) \mid_{w=0}) K(u)$$

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

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

En considérant l'estimateur

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

de $\theta^{(m)}(x) = \phi^{(m)}(x, F)$. Berlinet(1981) et Thomson(1990) 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 équation

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

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 noyau et K le noyau reproduisant de V.

3.4 Estimateur splines

Les méthodes splines connaissent un large spectre d'application et par la simplicité de leur mise en œuvre, 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'observation doit être suffisamment grand pour éviter les phénomènes classique de sur ou sous-lissage. Pour plus le précision sur le sujet on pourra se référer à Besse (1989) et Thomson (1990).

Berlinet (1981) 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 certaine conditions aux limites.

3.5 Estimation non paramétrique par des séries orthogonales

3.5.1 Principe de la méthode

Soit $X_1, ..., X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées de densité de probabilité $f(.)$ et de fonction de répartition $F(x)$ sur $[a, b] \in \mathbb{R}$. Il s'agit d'estimer $F(x)$ à partir des observations $X_1, ..., X_n$. Pour cela on suppose que :

- $\{e_k, k \in \mathbb{N}\}$ un système orthogonal dans $\mathbb{L}_2([a, b])$
- $F \in \mathbb{L}_2([a, b])$ tel que :

$$F(x) = \sum_{k=0}^{\infty} A_k e_k(x), x \in [a, b]. \tag{3.2}$$

- Le développement à l'ordre d_n de $F(x)$ est donné par :

$$F_{d_n}(x) = \sum_{k=0}^{d_n} A_k e_k(x), x \in [a, b]. \tag{3.3}$$

- Les $\{A_k\}_{\{k\in\mathbb{N}\}}$ sont les coefficients de Fourier associés à $F(x)$ satisfont

$$A_k = \int_a^b e_k(x) F(x) dx. \tag{3.4}$$

Supposons que $\sum_{k=0}^{\infty} A_k e_k(x)$ uniformément convergente sur $[a, b]$ et chaque fonction $e_k(x)$ est continue. Pour tout $x \in [a, b]$,l'estimateur de $F(x)$ est donné par :

$$\hat{F}_{d_n}(x) = \sum_{k=0}^{d_n} \hat{A}_k e_k(x), \tag{3.5}$$

avec (d_n) est une suite d'entiers telle que $d_n \longrightarrow \infty$ lorsque $n \longrightarrow \infty$, et

$$\hat{A}_k = \frac{1}{n} \sum_{i=1}^{n} \int e_k(x) \epsilon(x - X_i) dx, \quad avec \ \ \epsilon(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 & \text{si } x \geq 0. \end{cases} \tag{3.6}$$

Chapitre 4

Estimation non paramétrique de la fonction de répartition par des séries trigonométriques

4.1 Introduction

L'objectif de se chapitre est de présenter un estimateur de la fonction de répartition basé sur un systéme trigonométrique. Nous mettons en place, d'une part, les premiers résultats concernant les propriétés statistiques de l'estimateur, d'autre part un certain nombre d'outils permettant d'étudier le comportement asymptotique de notre estimateur. Cette étude ne sera pas complète si l'on ne s'interesse pas à l'étude du probléme de selection de paramétre de lissage. A cette dérniére problématique est dévouée la dérniére partie de ce chapitre

4.2 Principe de la méthode

Soit $X_1, ..., X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées de densité de probabilité $f(x)$ et de fonction de répartition $F(x)$ sur $[0, 1]$, . Il s'agit d'estimer $F(x)$ à partir des observations $X_1, ..., X_n$. Pour cela on considére que :

– La base orthogonale proposée est donnée par :

$$e_k(x) = \begin{cases} 1, & \text{k = 0 ;} \\ \sqrt{2}\cos(\pi k x), & \text{si k = 1,2,...} \end{cases} \tag{4.1}$$

– $F \in \mathbb{L}_2([0,1])$ où :

$$F(x) = \sqrt{2}\sum_{k=0}^{\infty} A_k \cos(\pi k x), x \in [0,1]. \tag{4.2}$$

– Le développement à l'ordre d_n de $F(x)$ est donné par :

$$F_{d_n}(x) = \sqrt{2}\sum_{k=0}^{d_n} A_k \cos(\pi k x), x \in [0,1]. \tag{4.3}$$

– Les $\{A_k\}_{\{k\in\mathbb{N}\}}$ sont les coefficients de Fourier associés à $F(x)$ satisfont

$$A_k = \sqrt{2}\int_0^1 (\cos(\pi k x))F(x)dx, \tag{4.4}$$

avec (d_n) est une suite d'entiers telle que $d_n \longrightarrow \infty$ lorsque $n \longrightarrow \infty$.

Avec la nouvelle base orthogonale, les estimateurs des coefficientsde de Fouier associés à la fonction de répartition F(x) s'écrivent comme suit :

$$\hat{A}_k = \frac{1}{n}\sum_{i=1}^{n}\int e_k(x)\epsilon(x - X_i)dx, \;\; avec \;\; \epsilon(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 & \text{si } x \geq 0. \end{cases} \tag{4.5}$$

Pour $k \neq 0$, nous avons :

$$\begin{aligned} \hat{A}_k &= \frac{\sqrt{2}}{n}\sum_{i=1}^{n}\int_{X_i}^{1}\cos(\pi k x)dx \\ &= \frac{\sqrt{2}}{\pi k n}\sum_{i=1}^{n}\sin(-\pi k X_i) \end{aligned} \tag{4.6}$$

Pour $k = 0$, $\hat{A}_0$ s'écrit conmme suit :

$$\begin{aligned} \hat{A}_0 &= \frac{1}{n}\sum_{i=1}^{n}\int_{X_i}^{1} dx \\ &= \frac{1}{n}\sum_{i=1}^{n}(1 - X_i) \\ &= 1 - \overline{X} \end{aligned} \tag{4.7}$$

Par conséquent,

$$\hat{A}_k = \begin{cases} \frac{\sqrt{2}}{k\Pi n}\sum_{i=1}^{n}\sin(-k\pi X_i) & \text{si } k \neq 0 \\ 1 - \overline{X} & \text{si } k = 0. \end{cases} \tag{4.8}$$

En utilisant (4.8), l'estimateur $\hat{F}_{d_n}(x)$ de $F(x)$ est donné par :

$$\begin{aligned}\hat{F}_{d_n}(x) &= \sum_{k=0}^{d_n} \hat{A}_k e_k(x) \\ &= \sqrt{2}\sum_{k=0}^{d_n} \hat{A}_k \cos(\pi k x) \\ &= 1 - \overline{X} + \sqrt{2}\sum_{k=0}^{d_n} \hat{A}_k \cos(\pi k x) \\ &= 1 - \overline{X} + \frac{2}{\pi n}\sum_{k=0}^{d_n}\sum_{i=1}^{n} \frac{1}{k}\sin(-\pi k X_i)\cos(\pi k x). \end{aligned} \tag{4.9}$$

Par conséquent, l'estimateur de la fonction de répartition associé à la nouvelle base trigonométrique est de la forme :

$$\hat{F}_{d_n}(x) = 1 - \overline{X} + \frac{2}{\pi n}\sum_{k=0}^{d_n}\sum_{i=1}^{n} \frac{1}{k}\sin(-\pi k X_i)\cos(\pi k x). \tag{4.10}$$

4.2.1 Propriétés des coéfficients de Fourier

Propriétés Statistiques

Théorème 4.2.1.

Soient $\{\hat{A}_k\}_{k\in\mathbb{N}}$ (4.8) les estimateurs des coefficients de Fourier $(A_k)_{k\in\mathbb{N}}$ associés à $F(x)$ sur $[0,1]$, alors les coefficients $(\hat{A}_k)_{k\in\mathbb{N}}$ sont des estimateurs sans biais de $(A_k)_{k\in\mathbb{N}}$, c'est-à-dire :

$$\mathbb{E}(\hat{A}_k) = A_k, \forall k = 0, 1, ... \tag{4.11}$$

Démonstration.

Pour $k = 0$, on a

$$\mathbb{E}(\hat{A}_0) = \mathbb{E}(1 - \overline{X}) = 1 - \mu.$$

Par ailleurs,

$$A_0 = \int_0^1 F(x)dx. \tag{4.12}$$

En appliquant une intégration par parties, nous obtenons

$$A_0 = [xF(x)]_0^1 - \int_0^1 xf(x)dx] = 1 - \mu = \mathbb{E}(\hat{A}_0)$$

Pour $k \neq 0$, l'espérance mathématiques de $(\hat{A}_k)$ s'écrit comme suit :

$$\begin{aligned}\mathbb{E}(\hat{A}_k) &= \mathbb{E}(\frac{\sqrt{2}}{\pi kn}\sum_{i=1}^{n}\sin(-\pi kX_i)) \\ &= \frac{\sqrt{2}}{\pi k}\mathbb{E}(\sin(-\pi kX)) \\ &= \frac{\sqrt{2}}{\pi k}\int_0^1 \sin(-\pi kx)f(x)dx \end{aligned} \tag{4.13}$$

De la même façon, en appliquant une intégration par parties, nous obtenons

$$\begin{aligned}\mathbb{E}(\hat{A}_k) &= \frac{\sqrt{2}}{\pi k}([\sin(\pi kx)F(x)]_0^1 + \pi k\int_0^1 \cos(\pi kx)F(x)dx) \\ &= \sqrt{2}\int_0^1 \cos(\pi kx)F(x)dx \\ &= A_k. \end{aligned} \tag{4.14}$$

Finalement,

$$\mathbb{E}(\hat{A}_k) = A_k, \forall k = 0, 1, ...$$

Ce qui implique que les $(\hat{A}_k)_{k\in\mathbb{N}}$ sont des estimateurs sans biais de $(A_k)_{k\in\mathbb{N}}$.

Théorème 4.2.2.

Soient $\{\hat{A}_k\}_{k\in\mathbb{N}}$ (4.8) les estimateurs des coefficients de Fourier $\{A_k\}_{k\in\mathbb{N}}$ associés à $F(x)$ sur $[0,1]$, et $X_1, ..., X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées, avec $\mathbb{E}(X) = \mu$ et $\mathbb{V}ar(X) = \sigma^2$. Alors

$$\mathbb{V}ar(\hat{A}_k) = \begin{cases} \frac{1}{(\pi k)^2 n} - \frac{1}{\sqrt{2}(\pi k)^2 n}a_{2k} - \frac{1}{n}A_k^2 & si\ k \neq 0 \\ \frac{\sigma^2}{n} & sinon. \end{cases}$$

Démonstration.

Pour $k = 0$ et d'après la définition des $\{\hat{A}_k\}_{k\in\mathbb{N}}$, nous avons :

$$\mathbb{V}ar(\hat{A}_0) = \mathbb{V}ar(1 - \overline{X}) = \frac{\sigma^2}{n}. \tag{4.15}$$

Pour $k \neq 0$, la variance de $(\hat{A}_k)$ se met sous la forme suivante :

$$\begin{aligned}
\mathbb{V}ar(\hat{A}_k) &= \mathbb{V}ar[\frac{\sqrt{2}}{\pi kn}\sum_{i=1}^{n}\sin(-\pi kX_i)] \\
&= \frac{2}{(\pi k)^2 n}\mathbb{V}ar(\sin(-\pi kX)) \\
&= \frac{2}{(\pi k)^2 n}[\mathbb{E}(\sin(-\pi kX)^2 - [\mathbb{E}(\sin(-\pi kX)]^2] \\
&= \frac{2}{(\pi k)^2 n}\mathbb{E}(\frac{1-\cos(2\pi kX)}{2}) - \frac{1}{n}A_k^2 \\
&= \frac{1}{(\pi k)^2 n} - \frac{1}{\sqrt{2}(\pi k)^2 n}a_{2k} - \frac{1}{n}A_k^2 \qquad (4.16)
\end{aligned}$$

Théorème 4.2.3.

Soient $\{\hat{A}_k\}_{k\in\mathbb{N}}$ (4.8) les estimateurs des coefficients de Fourier $\{A_k\}_{k\in\mathbb{N}}$ associés à $F(x)$ sur $[0,1]$, alors

$$\mathbb{C}ov(\hat{A}_k, \hat{A}_j) = \frac{1}{\sqrt{2}\pi^2 jkn}a_{k-j} - \frac{1}{\sqrt{2}\pi^2 jkn}a_{k+j} - \frac{1}{n}A_kA_j, k, j = 1, 2, \ldots. \qquad (4.17)$$

Démonstration.

La covariance des $(\hat{A}_k)$ est par définition :

$$\mathbb{C}ov(\hat{A}_k, \hat{A}_j) = \mathbb{E}(\hat{A}_k\hat{A}_j) - \mathbb{E}(\hat{A}_k)\mathbb{E}(\hat{A}_j). \qquad (4.18)$$

D'autre part, nous avons :

$$\begin{aligned}
\mathbb{E}(\hat{A}_k\hat{A}_j) &= \mathbb{E}[(\frac{\sqrt{2}}{\pi kn}\sum_{i=1}^{n}\sin(-\pi kX_i))(\frac{\sqrt{2}}{\pi jn}\sum_{i=1}^{n}\sin(-\pi jX_i))] \\
&= \frac{2}{(\pi n)^2 jk}[\sum_{i=1}^{n}\mathbb{E}(\sin(-\pi kX_i)\sin(-\pi jX_i)) + \sum_{i=1}^{n}\sum_{\substack{l=1,\\ l\neq i}}^{n}\mathbb{E}(\sin(-\pi kX_i)\sin(-\pi jX_l))] \\
&= \frac{2}{\pi^2 jkn}\mathbb{E}(\sin(-\pi kX)\sin(-\pi jX)) + \frac{2(n-1)}{\pi^2 jkn}\mathbb{E}(\sin(-\pi kX)\mathbb{E}(\sin(-\pi jX)) \\
&= \frac{1}{\pi^2 jkn}[\mathbb{E}(\cos(-\pi X(k-j))) - \mathbb{E}(\cos(-\pi X(k+j)))] + \frac{(n-1)}{n}A_kA_j \\
&= \frac{1}{\sqrt{2}\pi^2 jkn}a_{k-j} - \frac{1}{\sqrt{2}\pi^2 jkn}a_{k+j} + \frac{n-1}{n}A_kA_j, \qquad (4.19)
\end{aligned}$$

et

$$\mathbb{E}(\hat{A}_k)\mathbb{E}(\hat{A}_j) = A_kA_j, \qquad (4.20)$$

d'où d'aprés (4.19) et (4.20) on a

$$\mathbb{C}ov(\hat{A}_k, \hat{A}_j) = \frac{1}{\sqrt{2}\pi^2 jkn} a_{k-j} - \frac{1}{\sqrt{2}\pi^2 jkn} a_{k+j} - \frac{1}{n} A_k A_j. \tag{4.21}$$

Théorème 4.2.4.

Soient $\{\hat{A}_k\}_{k\in\mathbb{N}}$ *(4.8) les estimateurs des coefficients de Fourier* $\{A_k\}_{k\in\mathbb{N}}$ *associés à* $F(x)$ *sur* $[0,1]$*, alors*

$$\lim_{n\longrightarrow\infty} \mathbb{V}ar(\hat{A}_k) = 0, \forall k = 0, 1, ... \tag{4.22}$$

Démonstration.

Pour $k = 0$*, on a*

$$\lim_{n\longrightarrow\infty} \mathbb{V}ar(\hat{A}_0) = \lim_{n\longrightarrow\infty} \frac{\sigma^2}{n} = 0.$$

Pour $k \neq 0$*, nous avons*

$$\lim_{n\longrightarrow\infty} \mathbb{V}ar(\hat{A}_k) = \lim_{n\longrightarrow\infty} [\frac{1}{(\pi k)^2 n} - \frac{1}{\sqrt{2}\pi kn} a_{2k} - \frac{1}{n} A_k] = 0.$$

Théorème 4.2.5.

Soient $\{\hat{A}_k\}_{k\in\mathbb{N}}$ *les estimateurs de coefficients de Fourier* $\{A_k\}_{k\in\mathbb{N}}$ *associés à* $F(x)$ *sur* $[0,1]$*, alors*

$$\lim_{n\longrightarrow\infty} MSE(\hat{A}_k) = \lim_{n\longrightarrow\infty} \mathbb{E} \mid \hat{A}_k - A_k \mid^2 = 0. \tag{4.23}$$

Démonstration.

En introduisant l'erreur quadratique moyenne (MSE) associée à $(\hat{A}_k)$ *, alors le MSE de* $(\hat{A}_k)$ *se met sous la forme suivante :*

$$\begin{aligned} MSE(\hat{A}_k) &= \mathbb{E} \mid \hat{A}_k - A_k \mid^2 \\ &= \mathbb{E} \mid \hat{A}_k - E[\hat{A}_k] \mid^2 \\ &= \mathbb{V}ar(\hat{A}_k). \end{aligned}$$

Donc,

$$\lim_{n\longrightarrow\infty} MSE(\hat{A}_k) = \lim_{n\longrightarrow\infty} \mathbb{V}ar(\hat{A}_k) = 0, \forall k = 0, 1, ...$$

Par conséquent,

$$\lim_{n\longrightarrow\infty} MSE(\hat{A}_k) = 0, \forall k = 0, 1, ...$$

Ce qui nous amène à déduire que les $(\hat{A}_k)$ sont asymptotiquement convergent et vérifiant la convergence en moyenne qudratique.

4.2.2 Propriétés statistiques de l'estimateur

Biais de l'estimateur

Théorème 4.2.6.

Soient $\{\hat{A}_k\}_{k\in\mathbb{N}}$ *(4.8) les estimateurs des coefficients de Fourier associés à* $F(x)$ *sur* $[0,1]$. *Alors l'estimateur* $\hat{F}_{d_n}(x)$ *de* $F(x)$ *est biaisé.*

Démonstration.

Le biais de l'estimateur s'écrit :

$$\mathbb{B}iais(\hat{F}_{d_n}(x)) = \mathbb{E}(\sum_{k=0}^{d_n}\hat{A}_k e_k(x)) - \sum_{k=0}^{\infty} A_k e_k(x) = -\sum_{k=d_n+1}^{\infty} A_k e_k(x). \tag{4.24}$$

$\hat{F}_{d_n}(x)$ *est donc un estimateur biaisé de* $F(x)$.

Variance de l'estimateur

Théorème 4.2.7.

Soit $X_1, ..., X_n$ *une suite de variables aléatoires indépendantes et identiquement distribuées avec* $\mathbb{E}(X) = \mu$ *et* $\mathbb{V}ar(X) = \sigma^2$. *Soient* $\{\hat{A}_k\}_{k\in\mathbb{N}}$ (4.8) *les estimateurs de coefficients de Fourier associés à* $F(x)$ *sur* $[0,1]$. *Alors*

$$\begin{aligned}\mathbb{V}ar(\hat{F}_{d_n}(x)) &= \frac{\sigma^2}{n} + \sum_{k=1}^{d_n} e_k^2(x)\left(\frac{1}{(\pi k)^2 n} - \frac{1}{\sqrt{2}(\pi k)^2 n}a_{2k} - \frac{1}{n}A_k^2\right)\\ &+ \sum_{i=1}^{d_n}\sum_{j=1}^{d_n}\left[\frac{1}{\sqrt{2}\pi^2 jkn}a_{k-j} - \frac{1}{\sqrt{2}\pi^2 jkn}a_{k+j} - \frac{1}{n}A_k A_j\right]e_k(x)e_j(x)\\ &- \frac{2\sqrt{2}}{\pi n}\sum_{k=1}^{d_n}\frac{\gamma_k}{k}e_k(x) + \frac{2\mu}{n}\sum_{k=1}^{d_n}A_k e_k(x).\end{aligned} \tag{4.25}$$

Avec $\gamma_k = \mathbb{E}[X\sin(-\pi k X)]$.

Démonstration.

La variance de $\hat{F}_{d_n}(x)$ *est par définition :*

$$\mathbb{V}ar(\hat{F}_{d_n}) = \mathbb{V}ar(1-\overline{X}) + \mathbb{V}ar(\sum_{k=1}^{d_n}\hat{A}_k e_k(x)) + 2\mathbb{C}ov(1-\overline{X}, \sum_{k=1}^{d_n}\hat{A}_k e_k(x)) \tag{4.26}$$

Le premier terme de (4.26) s'exprime comme suit :

$$\mathbb{V}ar(1-\overline{X}) = \mathbb{V}ar(\overline{X}) = \frac{\sigma^2}{n}, \tag{4.27}$$

et le second terme de (4.26) *se met sous la forme :*

$$\begin{aligned}\mathbb{V}ar(\sum_{k=1}^{d_n}\hat{A}_k e_k(x)) &= \sum_{k=1}^{d_n}\mathbb{V}ar(\hat{A}_k)e_k^2(x)+\sum_{k=1}^{d_n}\sum_{\substack{j=1,\\ j\neq k}}^{d_n}\mathbb{C}ov(\hat{A}_k,\hat{A}_j)e_k(x)e_j(x)\\ &= \sum_{k=1}^{d_n}(e_k^2(x))(\frac{1}{(\pi k)^2 n}-\frac{1}{\sqrt{2}(\pi k)^2 n}a_{2k}-\frac{1}{n}A_k^2)\\ &+ \sum_{k=1}^{d_n}\sum_{\substack{j=1,\\ j\neq k}}^{d_n}(\frac{1}{\sqrt{2}\pi^2 jkn}a_{k-j}-\frac{1}{\sqrt{2}\pi^2 jkn}a_{k+j}-\frac{1}{n}A_kA_j)e_k(x)e_j(x),\end{aligned}$$

D'autre part, on a

$$\mathbb{C}ov(1-\overline{X},\sum_{k=1}^{d_n}\hat{A}_k e_k(x))=\mathbb{E}[(1-\overline{X})(\sum_{k=1}^{d_n}\hat{A}_k e_k(x))]-\mathbb{E}(1-\overline{X})\mathbb{E}(\sum_{k=1}^{d_n}\hat{A}_k e_k(x)) \quad (4.28)$$

Le premier terme de (4.28) peut être développé comme suit :

$$\begin{aligned}\mathbb{E}[(1-\overline{X})(\sum_{k=1}^{d_n}\hat{A}_k e_k(x))] &= \mathbb{E}(\sum_{k=1}^{d_n}\hat{A}_k e_k(x))-\mathbb{E}(\overline{X}\sum_{k=1}^{d_n}\hat{A}_k e_k(x))\\ &= \sum_{k=1}^{d_n}A_k e_k(x)-\frac{\sqrt{2}}{\pi n^2}\mathbb{E}(\sum_{k=1}^{d_n}\frac{1}{k}e_k(x)\sum_{i=1}^{n}\sum_{i=j}^{n}X_i\sin(-\pi kX_j))\\ &= \sum_{k=1}^{d_n}A_k e_k(x)-\frac{\sqrt{2}}{\pi n^2}\sum_{k=1}^{d_n}\frac{1}{k}e_k(x)[\sum_{i=1}^{n}\mathbb{E}(X_i\sin(-\pi kX_i))\\ &+ \frac{\sqrt{2}}{\pi n}-\mathbb{E}(\sum_{i=1}^{n}\sum_{\substack{j=1,\\ j\neq i}}^{n}X_i\sin(-\pi kX_j))\\ &= \sum_{k=1}^{d_n}A_k e_k(x)-\frac{\sqrt{2}}{\pi n}\sum_{k=1}^{d_n}\frac{1}{k}e_k(x)\mathbb{E}(X\sin(-\pi kX))\\ &- \frac{\sqrt{2}(n-1)}{\pi n}\sum_{k=1}^{d_n}\frac{1}{k}e_k(x)\mathbb{E}(X)\mathbb{E}(\sin(-\pi kX))\\ &= \sum_{k=1}^{d_n}A_k e_k(x)-\frac{\sqrt{2}}{\pi n}\sum_{k=1}^{d_n}\frac{1}{k}e_k(x)\mathbb{E}(X\sin(-\pi kX))\\ &- \mu\frac{(n-1)}{n}\sum_{k=1}^{d_n}A_k e_k(x), \quad (4.29)\end{aligned}$$

et,le second terme de (4.28) *s'exprime comme suit :*

$$\mathbb{E}(1-\overline{X})\mathbb{E}(\hat{A}_k e_k(x))=(1-\mu)(\sum_{k=1}^{d_n}A_k e_k(x)) \quad (4.30)$$

Par conséquent,

$$2\mathbb{C}ov(1-\overline{X},\sum_{k=1}^{d_n}\hat{A}_k e_k(x)) = \frac{-2\sqrt{2}}{\pi n}\sum_{k=1}^{d_n}\frac{1}{k}e_k(x)\mathbb{E}(X\sin(-\pi kX))+\frac{2\mu}{n}\sum_{k=1}^{d_n}A_k e_k(x). \quad (4.31)$$

Donc la combinaison de (4.27), (4.28) et (4.31) permet d'écrire :

$$\begin{aligned}\mathbb{V}ar(\sum_{k=1}^{d_n}\hat{A}_k e_k(x)) &= \frac{\sigma^2}{n}+\sum_{k=1}^{d_n}e_k^2(x)(\frac{1}{(\pi k)^2 n}-\frac{1}{\sqrt{2}(\pi k)^2 n}a_{2k}-\frac{1}{n}A_k^2)\\ &+ \sum_{i=1}^{d_n}\sum_{j=1}^{d_n}[\frac{1}{\sqrt{2}\pi^2 jkn}a_{k-j}-\frac{1}{\sqrt{2}\pi^2 jkn}a_{k+j}-\frac{1}{n}A_kA_j]e_k(x)e_j(x)\\ &- \frac{2\sqrt{2}}{\pi n}\sum_{k=1}^{d_n}\frac{\gamma_k}{k}e_k(x)+\frac{2\mu}{n}\sum_{k=1}^{d_n}A_k e_k(x). \quad (4.32)\end{aligned}$$

d'où le résultat

Erreur quadratique moyenne

Théorème 4.2.8.

Soient $X_1,...,X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées avec $\mathbb{E}(X)=\mu$ et $\mathbb{V}ar(X)=\sigma^2$. Soient $\{\hat{A}_k\}_{k\in\mathbb{N}}$ (4.8) les estimateurs de coefficients de Fourier $\{A_k\}_{k\in\mathbb{N}}$ associés à $F(x)$ sur $[0,1]$. Si $d_n=o(\sqrt{n})$ et $d_n\longrightarrow\infty$ lorsque $n\longrightarrow\infty$, alors

$$\begin{aligned}\mathbb{E}\mid\hat{F}_{d_n}(x)-F(x)\mid^2 &= \frac{\sigma^2}{n}+\sum_{k=1}^{d_n}e_k^2(x)(\frac{1}{(\pi k)^2 n}-\frac{1}{\sqrt{2}(\pi k)^2 n}a_{2k}-\frac{1}{n}A_k^2)\\ &+ (\sum_{i=1}^{d_n}\sum_{\substack{j=1,\\ j\neq i}}^{d_n}\frac{1}{\sqrt{2}\pi^2 jkn}a_{k-j}-\frac{1}{\sqrt{2}\pi^2 jkn}a_{k+j}-\frac{1}{n}A_kA_j)e_k(x)e_j(x)\\ &+ \frac{2\sqrt{2}}{\pi n}\sum_{k=1}^{d_n}\frac{1}{k}e_k(x)\gamma_k+\frac{2\mu}{n}\sum_{k=1}^{d_n}A_k e_k(x)+[\sum_{k=d_n+1}^{\infty}A_k e_k]^2.\end{aligned}$$

Démonstration.

On a :

$$\mathbb{E}\mid\hat{F}_{d_n}(x)-F(x)\mid^2=\mathbb{V}ar(\hat{F}_{d_n}(x))+\mathbb{B}ias^2(\hat{F}_{d_n}(x)).$$

En utilisant le théorème (4.2.6) et (4.2.7) pour obtenir le résultat :

$$\begin{aligned}\mathbb{E} \mid \hat{F}_{d_n}(x) - F(x) \mid^2 &= \frac{\sigma^2}{n} + \sum_{k=1}^{d_n} e_k^2(x)\left(\frac{1}{(\pi k)^2 n} - \frac{1}{\sqrt{2}(\pi k)^2 n} a_{2k} - \frac{1}{n} A_k^2\right) \\ &+ \left(\sum_{i=1}^{d_n} \sum_{\substack{j=1, \\ j\neq i}}^{d_n} \frac{1}{\sqrt{2}\pi^2 jkn} a_{k-j} - \frac{1}{\sqrt{2}\pi^2 jkn} a_{k+j} - \frac{1}{n} A_k A_j\right) e_k(x) e_j(x) \\ &- \frac{2\sqrt{2}}{\pi n} \sum_{k=1}^{d_n} \frac{1}{k} e_k(x)\gamma_k + \frac{2\mu}{n} \sum_{k=1}^{d_n} A_k e_k(x) + [\sum_{k=d_n+1}^{\infty} A_k e_k]^2.\end{aligned}$$

Théorème 4.2.9.

Soit $X_1, ..., X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées avec $\mathbb{E}(X) = \mu$ et $\mathbb{V}ar(X) = \sigma^2$, . Soient $\{\hat{A}_k\}_{k\in\mathbb{N}}$ (4.8) les estimateurs de coefficients de Fourier associés à $F(x)$ sur $[0, 1]$, alors

$$MISE(\hat{F}_{d_n}(x)) = \sum_{k=1}^{d_n}\left(\frac{1}{(\pi k)^2 n} - \frac{1}{\sqrt{2}(\pi k)^2 n} a_{2k} - \frac{A_k^2}{n}\right) \sum_{k=d_n+1}^{\infty} A_k^2.$$

Démonstration.

L'erreur quadratique moyenne intégrée peut se mettre sous la forme :

$$\begin{aligned}
MISE(\hat{F}_{d_n}(x)) &= \mathbb{E}\int_0^1 \mid \hat{F}_{d_n}(x) - F(x) \mid^2 dx \\
&= \int_0^1 F^2(x)dx - 2\int_0^1 \mathbb{E}(\hat{F}_{d_n}(x))F(x)dx + \int_0^1 \mathbb{E}(\hat{F}_{d_n}(x))^2 dx \\
&= \int_0^1 F^2(x)dx - 2\int_0^1 \mathbb{E}(\sum_{k=0}^{d_n} \hat{A}_k e_k(x))F(x)dx + \int_0^1 \mathbb{E}(\sum_{k=0}^{d_n} \hat{A}_k e_k(x))^2)dx \\
&= \int_0^1 F^2(x)dx - 2\sum_{k=0}^{d_n} A_k \int_0^1 e_k(x)F(x)dx + \sum_{k=0}^{d_n}[\mathbb{V}ar(\hat{A}_k) + (\mathbb{E}(\hat{A}_k))^2] \int_0^1 e_k^2(x)dx \\
&= \int_0^1 F^2(x)dx - 2\sum_{k=0}^{d_n} A_k^2 + \sum_{k=0}^{d_n}[\mathbb{V}ar(\hat{A}_k) + A_k^2] \\
&= \sum_{k=0}^{\infty} A_k^2 - 2\sum_{k=0}^{d_n} A_k^2 + \sum_{k=0}^{d_n} A_k^2 + \sum_{k=0}^{d_n} \mathbb{V}ar(\hat{A}_k).
\end{aligned}$$

Ce qui implique que :

$$\begin{aligned}
MISE(\hat{F}_{d_n}(x)) &= \sum_{k=0}^{d_n} \mathbb{V}ar(\hat{A}_k) + \sum_{k=d_n+1}^{\infty} A_k^2 \\
&= \sum_{k=1}^{d_n}(\frac{1}{(\pi k)^2 n} - \frac{1}{\sqrt{2}(\pi k)^2 n} a_{2k} - \frac{A_k^2}{n} + \sum_{k=d_n+1}^{\infty} A_k^2
\end{aligned}$$

Biais Asymptotique

Théorème 4.2.10.

Si $d_n \longrightarrow \infty$ lorsque $n \longrightarrow \infty$, l'estimateur $\hat{F}_{d_n}(x)$ est asymptotiquement sans biais.

Démonstration.

$$\begin{aligned}
\lim_{n\longrightarrow\infty} \mathbb{E}(\hat{F}_{d_n}(x)) &= \lim_{n\longrightarrow\infty} \mathbb{E}[\sum_{k=0}^{d_n} \hat{A}_k e_k(x)] \\
&= \lim_{n\longrightarrow\infty} \sum_{k=0}^{d_n} A_k e_k(x) \\
&= \sum_{k=0}^{\infty} A_k e_k(x) \\
&= F(x).
\end{aligned}$$

Variance Asymptotique

Théorème 4.2.11.

Soit $X_1, ..., X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées avec $\mathbb{E}(X) = \mu$ et $\mathbb{V}ar(X) = \sigma^2$. Soient $\{\hat{A}_k\}_{k\in\mathbb{N}}$ (4.8) les estimateurs des coefficients de Fourier associés à $F(.)$ sur $[0, 1]$.

Si $d_n = o(\sqrt{n})$ et $d_n \longrightarrow \infty$ lorsque $n \longrightarrow \infty$, alors

$$\lim_{n\longrightarrow\infty} \mathbb{V}ar[\hat{F}_{d_n}(x)] = 0.$$

Démonstration.

La variance de $\hat{F}_{d_n}(x)$ est par définition :

$$\mathbb{V}ar[\hat{F}_{d_n}(x)] = \mathbb{V}ar[\sum_{k=0}^{d_n} \hat{A}_k e_k(x)] = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}ar[\sum_{k=0}^{d_n}(\int_{X_i}^{1} e_k(y)dy)e_k(x)].$$

On sait que $\mathbb{V}ar(X) \leq \mathbb{E}(X^2)$, alors

$$\mathbb{V}ar[\hat{F}_{d_n}(x)] \leq \frac{1}{n}\int_0^1 [\int_z^1 \sum_{k=0}^{d_n} e_k(y)e_k(x)dy]^2 f(z)dz. \tag{4.33}$$

Posons

$$\omega = \frac{1}{n}\int_0^1 [\int_z^1 \sum_{k=0}^{d_n} e_k(y)e_k(x)dy]^2 f(z)dz, \tag{4.34}$$

qui peut être développé comme suit :

$$\begin{aligned}
\omega &= \int_0^1 [\int_z^1 [1 + \sum_{i=1}^{d_n} 2\cos(\pi ky)\cos(\pi kx)]dy]^2 f(z)dz \\
&= \frac{1}{n}\int_0^1 [\int_z^1 dy + \int_0^1 (\frac{1}{2} + \sum_{k=1}^{d_n}\cos(\pi k(y+x)) + \frac{1}{2} + \sum_{k=1}^{d_n}\cos(\pi k(y-x))) \\
&= \frac{1}{4n}\int_0^1 [\int_z^1 [\frac{\sin[\frac{\pi(2d_n+1)(X+x)}{2}]}{\sin[\frac{\pi(X+x)}{2}]} + \frac{\sin[\frac{\pi(2d_n+1)(X-x))}{2}]}{\sin[\frac{\pi(X-x)}{2}]}]^2 dy]^2 f(z)dz
\end{aligned} \tag{4.35}$$

Nous avons

$$\frac{\sin(kx)}{\sin(x)} \leq k$$

$$\begin{aligned}
\mathbb{V}ar(\hat{F}_{d_n}(x)) &\leq \frac{1}{4n}[\int_0^1 [\int_z^1 (2d_{n+1})dy + \int_z^1 (2d_{n+1})dy]^2 f(z)dz] \\
&\leq \frac{1}{n}[\int_0^1 [(2d_n+1)(1-z)]^2 f(z)dz] \\
&\leq \frac{(2d_n+1)^2}{n}[\int_0^1 z^2 f(z)dz + \int_0^1 f(z)dz - 2\int_0^1 zf(z)dz] \\
&\leq \frac{(2d_n+1)^2}{n}(\sigma^2 + (1-\mu)^2) \qquad (4.36) \\
& \qquad (4.37)
\end{aligned}$$

Ce qui implique,finalement, que

$$\lim_{n\longrightarrow\infty} \mathbb{V}ar(\hat{F}_{d_n}(x)) = 0$$

Convergence en moyenne quadratique

Théorème 4.2.12.

Soit $X_1, ..., X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées avec $\mathbb{E}(X) = \mu$ et $\mathbb{V}ar(X) = \sigma^2$. Soient $\{\hat{A}_k\}_{k\in\mathbb{N}}$ (4.8) les estimateurs de coefficients de Fourier $\{A_k\}_{k\in\mathbb{N}}$ associés à $F(x)$ sur $[0,1]$. Si $d_n = o(\sqrt{n})$ et $d_n \longrightarrow \infty$ lorsque $n \longrightarrow \infty$, alors

$$\lim_{n\longrightarrow\infty} \mathbb{E} \mid \hat{F}_{d_n}(x) - F(x) \mid^2 = 0.$$

Démonstration.

L'erreur quadratique moyenne peut être s'exprimée comme suit :

$$\begin{aligned}
\mathbb{E} \mid \hat{F}_{d_n}(x) - F(x) \mid^2 &= \mathbb{E} \mid \hat{F}_{d_n}(x) - \sum_{k=0}^{d_n} A_k e_k(x) - \sum_{k=d_n+1}^{\infty} A_k e_k(x) \mid^2 \\
&= \mathbb{E} \mid \hat{F}_{d_n}(x) - \mathbb{E}(\hat{F}_{d_n}(x)) - \sum_{k=d_n+1}^{\infty} A_k e_k(x)) \mid^2 \\
&= \mathbb{V}ar(\hat{F}_{d_n}(x)) - 2\mathbb{E}[\hat{F}_{d_n}(x) - \mathbb{E}(\hat{F}_{d_n}(x))] \sum_{k=d_n+1}^{\infty} A_k e_k(x) + [\sum_{k=d_n+1}^{\infty} A_k e_k(x)]^2.
\end{aligned}$$

D'après le théorème (4.2.11), $\mathbb{V}ar(\hat{F}_{d_n}(x))$ tende vers 0 lorsque n tend vers 0.

Le deuxième terme égal à 0 car $\mathbb{E}[\hat{F}_{d_n}(x) - \mathbb{E}(\hat{F}_{d_n}(x))] = \mathbb{E}(\hat{F}_{d_n}(x)) - \mathbb{E}(\hat{F}_{d_n}(x))$. Le

troisième terme tend vers 0 lorsque n tend vers l'infini, car les coefficients de Fourier $\{A_k\}_{k\in\mathbb{N}}$ *tendent vers* 0 *lorsque n tend vers l'infini. On conclut, finalement, que*

$$\lim_{n\longrightarrow\infty} \mathbb{E} \mid \hat{F}_{d_n}(x) - F(x) \mid^2 dx = 0. \tag{4.38}$$

Convergence en moyenne quadratique intégrée

Théorème 4.2.13.

Soit $X_1, ..., X_n$ *une suite de variables aléatoires indépendantes et identiquement distribuées avec* $\mathbb{E}(X) = \mu$ *et* $\mathbb{V}ar(X) = \sigma^2$ *. Soient* $\{\hat{A}_k\}_{k\in\mathbb{N}}$ *(4.8) les estimateurs de coefficients de Fourier associés à* $F(.)$ *sur* $[0, 1]$.
Si $d_n = o(\sqrt{n})$ *et* $d_n \longrightarrow \infty$ *lorsque* $n \longrightarrow \infty$, *alors*

$$\lim_{n\longrightarrow\infty} \mathbb{E} \int_0^1 \mid \hat{F}_{d_n}(x) - F(x) \mid^2 dx = 0.$$

Démonstration.

D'après (4.2.12)

$$\lim_{n\longrightarrow\infty} \mathbb{E} \mid \hat{F}_{d_n}(x) - F(x) \mid^2 = 0,$$

la convergence est uniforme Par conséquent,

$$\lim_{n\longrightarrow\infty} \mathbb{E} \int_0^1 \mid \hat{F}_{d_n}(x) - F(x) \mid^2 dx = 0.$$

4.3 Détermination du nombre optimum de termes d_n

4.3.1 Méthode de Kronmal-Tarter

Il est naturel de choisir d_n de sorte que l'erreur quadratique moyenne intégrée $MISE(\hat{F}_{d_n}(x))$ soit minimum. La règle adoptée pour déterminer la valeur optimum $\hat{d}_n$ repose sur l'algorithme suivant : A partir de $d_n = 1$ on augmente la valeur de d_n d'une unité jusqu'à ce que $MISE$ augmente on donne alors à d_n la valeur qui précède juste l'augmentation de $MISE$. On ajoutera donc à la somme (4.9) le $d^{\text{ième}}$ terme si et seulement si

$$\Delta_{d_n} = MISE(\hat{F}_{d_n}(x)) - MISE(\hat{F}_{d_n-1}(x)) \leq 0.$$

Notons que l'expression (4.33) est en fonction des coefficients de Fourier associés à la densité de probabilité $f(.)$. Donc, le signe de $[MISE(\hat{F}_{d_n}(x)) - MISE(\hat{F}_{d_n-1}(x))]$ est égal au signe de $[MISE(\hat{f}_{d_n}(x)) - MISE(\hat{f}_{d_n-1}(x))]$.

Par conséquent, le paramètre de lissage qui minimise l'erreur quadratique moyenne intégrée associée à la fonction de répartion est donné par :

$$\hat{d}_n = \begin{cases} \inf\{d_n, 1 \leq d_n \leq D\} & \hat{\Delta}_d > 0 \\ D & sinon. \end{cases}$$

Avec

$$\hat{\Delta}_d = \frac{1}{n}[\frac{n+1}{n-1}\sum_{i=1}^{n}(\theta_i - \overline{\theta})^2 - \sum_{i=1}^{n}\theta_i^2]. \tag{4.39}$$

4.3.2 Méthode de Saadi et Adjabi

Les principales études ont concerné le choix du paramètre de lissage (nombre de termes de la série orthogonale). La règle adoptée pour déterminer le nombre optimum a été développé par Kronmal et Tarter généralisée par Ott et Kronmal (1976). Les inconvénients de cette méthode ont été soulignés par Crain qui a suggéré qu'il pourrait ne pas donner le terme optimal. Hall (986), a averti au sujet de la mauvaise performances possible et même l'incohérence de la règle dans des situations multimodales. Saadi et Adjabi (2016) ont proposé une nouvelle technique pour sélectionner ce paramètre de lissage qui aide à surmonter ces difficultés. Les difficultés rencontrées avec la méthode de Kronmal -Tarter sont dues au fait que les termes sont analysés un par un, La méthode de Saadi-A djabi est basée sur la manipulation nombreux termes ensemble .

Théorème 4.3.1.

Soit $X_1, ..., X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées avec $\mathbb{E}(X) = \mu$ et $\mathbb{V}ar(X) = \sigma^2$. Soient $\{\hat{A}_k\}_{k\in\mathbb{N}}$ (4.5) les estimateurs de coefficients de Fourier associés à $F(x)$ sur $[a, b]$. Si $F(.) \in L_2([a, b])$, alors

$$MISE(\hat{F}_{d_n}(x)) = \int_a^b F^2(x)dx + \sum_{k=0}^{d_n}[\mathbb{V}ar(\hat{A}_k) - A_k^2] = b - 2\mathbb{E}(XF(X)) + \sum_{k=0}^{d_n}[\mathbb{V}ar(\hat{A}_k) - A_k^2].$$

Le paramétre de lissage optimal d_n^* minimise

$$b - \frac{2}{n}\sum_{i=1}^{n}\psi_{-i}(X_i) + \sum_{k=0}^{d_n}\frac{1}{n}[\frac{n+1}{n-1}\sum_{i=1}^{n}(\vartheta_i - \overline{\vartheta})^2 - \sum_{i=1}^{n}\vartheta_i^2], \tag{4.40}$$

où,

$$\vartheta_i = \int_{X_i}^{b} e_k(x)dx,$$

$$\overline{\vartheta} = \frac{1}{n}\sum_{i=1}^{n} \vartheta_i,$$

et

$$\psi_{-i}(x) = \frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\sum_{k=0}^{d_n}(\int_{X_j}^{b} e_k(y)dy)xe_k(x).$$

Démonstration.

Nous avons :

$$\begin{aligned} MISE(\hat{F}_{d_n}(x)) &= \mathbb{E}\int_a^b \mid \hat{F}_{d_n}(x) - F(x) \mid^2 dx \\ &= \int_a^b F^2(x)dx + \sum_{k=0}^{d_n}[\mathbb{V}ar(\hat{A}_k) - A_k^2]. \end{aligned}$$

En appliquant une intégration par partie, on aura :

$$\begin{aligned} \int_a^b F^2(x)dx &= [bF^2(b) - aF^2(a)] - 2\int_a^b xf(x)F(x)dx \\ &= [bF^2(b) - aF^2(a)] - 2\mathbb{E}(XF(X)) \\ &= b - 2\mathbb{E}(XF(X)). \end{aligned}$$

Par conséquent,

$$MISE(\hat{f}_{d_n}(x)) = b - 2\mathbb{E}(Xf(X)) + \sum_{k=0}^{d_n}[\mathbb{V}ar(\hat{A}_k) - A_k^2].$$

Par ailleurs :

$$\begin{aligned} \mathbb{V}ar(\hat{A}_k) - A_k^2 &= \mathbb{V}ar(\frac{1}{n}\sum_{i=1}^{n}\int_{X_i}^{b} e_k(x)dx) - A_k^2 \\ &= \frac{1}{n}\mathbb{V}ar(\int_X^b e_k(x)dx) - [\mathbb{E}\int_X^b e_k(x)dx)]^2 \\ &= \frac{n+1}{n}\mathbb{V}ar(\int_X^b e_k(x)dx) - \mathbb{E}[\int_X^b e_k(x)dx)]^2. \end{aligned}$$

Posons

$$\vartheta_i = \int_{X_i}^{b} e_k(x)dx,$$

et

$$\overline{\vartheta} = \frac{1}{n}\sum_{i=1}^{n}\vartheta_i.$$

En définissant un estimateur symétrique de $\sum_{k=0}^{d_n}[\mathbb{V}ar(\hat{A}_k) - A_k^2]$ *donné*

$$\frac{1}{n}\sum_{k=0}^{d_n}[\frac{n+1}{n-1}\sum_{i=1}^{n}(\vartheta_i - \overline{\vartheta})^2 - \sum_{i=1}^{n}\vartheta_i^2].$$

Posons

$$R = \mathbb{E}(XF(X)) = \int_a^b xF(x)f(x)dx.$$

Un estimateur de R *est donné par :*

$$\hat{R} = \frac{1}{n}\sum_{i=1}^{n}\psi_{-i}(X_i),$$

avec

$$\psi_{-i}(X_i) = X_i\hat{F}_{d_n,-i}(X_i),$$

et

$$\hat{F}_{d_n,-i}(x) = \frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\sum_{k=0}^{d_n}(\int_{X_j}^{b} e_k(y)dy)e_k(x).$$

Le paramètre de lissage optimal est alors :

$$d_n^* = arg\min_{d_n}[b - \frac{2}{n}\sum_{i=1}^{n}\psi_{-i}(X_i) + \frac{1}{n}\sum_{k=0}^{d_n}[\frac{n+1}{n-1}\sum_{i=1}^{n}(\vartheta_i - \overline{\vartheta})^2 - \sum_{i=1}^{n}\vartheta_i^2]. \tag{4.41}$$

Où,

$$\vartheta_i = \sqrt{2}\int_{X_i}^{b}(\cos(\pi kx)dx \;\; and \;\; \overline{\vartheta} = \frac{\sqrt{2}}{n}\sum_{i=1}^{n}[\int_{X_i}^{b}(\cos(\pi kx)dx]. \tag{4.42}$$

4.4 Conclusion

Dans se chapitre nous avons proposé un estimateur de fonction de répartition basé sur un systéme trigonométrique. Cet estimateur possède de bonnes propriétés de convergence et asymptotiques.

Chapitre 5

Simulation

Nous présentons dans ce chapitre le travail de simulation effectué porr justifier de la qualité des estimateurs proposés. Pour cela, on va les comparer numériquement par simulation aux estimateurs associés a la base de Saadi et Adjabi.
L'expérimentation numérique nous servira en particulier à :

- Comparer les performances de l'estimateur construit à partir de la base d'Efromovich à ce lui de Saadi et Adjabi, et ceci en comparant les erreurs quadratiques intégrées des deux estimateurs ;
- Étudier l'influence de la taille de l'échantillon sur ces différents algorithmes.

5.1 Estimateur de la densité

5.1.1 Plan de simulation

Nous allons faire des simulations et observer deux estimateurs en question, calculés à partir d'échantillons simulés, censés représenter une loi connue,dont la densté de probabilité est f.
Nous utilisons pour les simulations des échantillons de loi connue de taille de plus en plus grande (50, 100, 1000, 1500, 3000, 3500, 5000, 7500, 10000). Pour la densité cible choisie et pour chaque taille d'échantillon 10 à 20 répétitions d'experiences ont été conduites.
La densité cible choisie pour tirer des échantillons est la loi normale $\mathbb{N}(0, 1)$.

5.1.2 Algorithme

L'algorithme de simulation que nous avons utilisé comporte les étapes suivantes :

◇ Simuler un échantillon de taille n ;

◇ Calculer le paramétre de lissage optimal de Kronmal-Tarter associé à la base d'Efromovich ainsi que l'erreur quadratique intégrée optimale associée pour la densité ;

◇ Calculer le paramètre de lissage optimal de Kronmal-Tarter associé à l'estimateur de Saadi et Adjabi ainsi que l'erreur quadratique intégrée optimale associée pour la densité ;

Méthode de Kronmal-Tarter

La méthode consiste à :

1. Simuler un échantillon $(X_1, \ldots, X_n)$, $0 \leq X_i \leq 1$, de loi $\mathcal{N}(0,1)$.
2. La taille des échantillons est fixée successivement à 50, 100, 1000, 1500, 2000, 2500, 3000, 3500, 5000, 7500 et 10000.
3. La base trigonométrique est donnée par :

$$e_k(x) = \begin{cases} 1, & \text{si k=0 ;} \\ \sqrt{2}\cos(\pi k x), & \text{si k=1,2,...} \end{cases}$$

4. D est égal à la partie entière de $(log(n))^{(1/2)}$, tel que :
   $D \longrightarrow \infty$, et $\frac{D}{n} \longrightarrow 0$ lorsque $n \longrightarrow \infty$.
5. Calculer le paramètre de lissage optimum associé à la base d'Efromovich d^*_{E-kt} :

$$d^*_{E-kt} = \begin{cases} \inf\{d_n : 1 \leq d_n \leq D\}; \hat{\Delta}_E > 0 \\ D \qquad\qquad\qquad\qquad\qquad sinon. \end{cases}$$

   où

$$\hat{\Delta}_E = \frac{1}{n}[\frac{n+1}{n-1}\sum_{i=1}^{n}[\sqrt{2}\cos(\pi k X_i) - \frac{1}{n}\sum_{i=1}^{n}\sqrt{2}\cos(\pi k X_i)]^2 - \sum_{i=1}^{n}(\sqrt{2}\cos(\pi k X_i))^2].$$

6. Calculer l'erreur quadratique moyenne intégrée optimale associée à cette base :

$$MISE^*_{E-kt} = \int_0^1 f^2(x)dx + \sum_{k=0}^{d^*_{E-kt}}(\mathbb{V}ar(\hat{a}_k) - a_k^2).$$

Pour $f(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$, on a :

$$MISE^*_{E-kt} = \frac{1}{\sqrt{\pi}} \int_0^1 [\exp(-\frac{x^2}{2})]^2 dx + \sum_{k=0}^{d^*_{E-kt}} (\mathbb{V}ar(\hat{a}_k) - a_k^2),$$

avec

$$a_k = \frac{1}{\sqrt{\pi}} \int_0^1 \exp(-\frac{x^2}{2}) \cos(\pi k x) dx,$$

et

$$\hat{a}_k = \frac{\sqrt{2}}{n} \sum_{i=1}^{n} \cos(\pi k X_i).$$

7. Calculer le paramètre de lissage optimal associé à l'estimateur de Saadi et Adjabi :

$$d^*_{SA-kt} = \begin{cases} \inf\{d_n : 1 \leq d_n \leq D\}; \hat{\Delta}_{SA} > 0 \\ D \qquad\qquad\qquad\qquad sinon. \end{cases}$$

où

$$\hat{\Delta}_{SA} = \frac{1}{n}[\frac{n+1}{n-1} \sum_{i=1}^{n} (\frac{1}{\sqrt{2\pi}}(\cos(X_i) + \sin(X_i)) - \frac{1}{n} \sum_{i=1}^{n} (\frac{1}{\sqrt{2\pi}}(\cos(X_i) + \sin(X_i)))^2 -$$

$$\sum_{i=1}^{n} (\frac{1}{\sqrt{2\pi}}(\cos(X_i) + \sin(X_i)))^2].$$

8. Calculer l'erreur quadratique moyenne intégrée optimale associée à l'estimateur de Saadi et Adjabi :

$$MISE^*_{SA-kt} = \int_{-\pi}^{\pi} f^2(x) dx + \sum_{k=0}^{d^*_{SA-kt}} (\mathbb{V}ar(\hat{a}_k) - a_k^2).$$

Pour $f(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$, on a :

$$MISE^*_{SA-kt} = \frac{1}{2\pi} \int_{-\pi}^{\pi} [\exp(-\frac{x^2}{2})]^2 dx + \sum_{k=0}^{d^*_{SA-kt}} (\mathbb{V}ar(\hat{a}_k) - a_k^2),$$

avec

$$a_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(-\frac{x^2}{2})(\cos(kx) + \sin(kx)) dx,$$

et

$$\hat{a}_k = \frac{1}{\sqrt{2\pi} n} \sum_{i=1}^{n} [\cos(kX_i) + \sin(kX_i)].$$

Résultats de la simulation

Les résultats de la simulation sont donnés sous forme de tableaux. Les tableaux contiennent les résultats suivants :

◇ d^*_{E-kt} : paramètre de lissage optimal associé à la base d'Efromovich en utilisant la méthode de Kronmal-Tarter ;

◇ d^*_{SA-kt} : paramètre de lissage optimal associé à l'estimateur de Saadi et Adjabi en utilisant la méthode de Kronmal-Tarter

◇ $MISE^*_{E-kt}$: erreur quadratique moyenne intégrée optimale associée à la base d'Efromovich en utilisant la méthode de Kronmal-Tarter ;

◇ $MISE^*_{SA-kt}$: erreur quadratique moyenne intégrée optimale associée à l'estimateur de Saadi et Adjabi en utilisant la méthode de Kronmal-Tarter ;

n	d^*_{E-kt}	d^*_{SA-kt}	$MISE^*_{E-kt}$	$MISE^*_{SA-kt}$
50	1	2	$4.6873 * 10^{-3}$	$2.9294 * 10^{-2}$
100	2	3	$3.3062 * 10^{-3}$	$8.3682 * 10^{-3}$
1000	3	3	$5.3565 * 10^{-4}$	$7.4108 * 10^{-3}$
1500	3	3	$4.5293 * 10^{-4}$	$4.1662 * 10^{-3}$
2500	3	3	$1.7157 * 10^{-4}$	$3.2491 * 10^{-4}$
3000	4	3	$4.0028 * 10^{-5}$	$2.9105 * 10^{-4}$
5000	4	4	$2.0775 * 10^{-5}$	$1.1493 * 10^{-4}$
7500	4	4	$1.0758 * 10^{-5}$	$1.1112 * 10^{-4}$
10000	4	4	$5.5617 * 10^{-6}$	$6.5867 * 10^{-5}$

TABLE 5.1 – Paramètre de lissage optimal et variation du $MISE$ optimal en fonction de n pour la Loi normale $\mathcal{N}(0,1)$, en utilisant la méthode de Kronmal-Tarter pour la base d'Efromovich et la base de Saadi et Adjabi

Représentation graphique

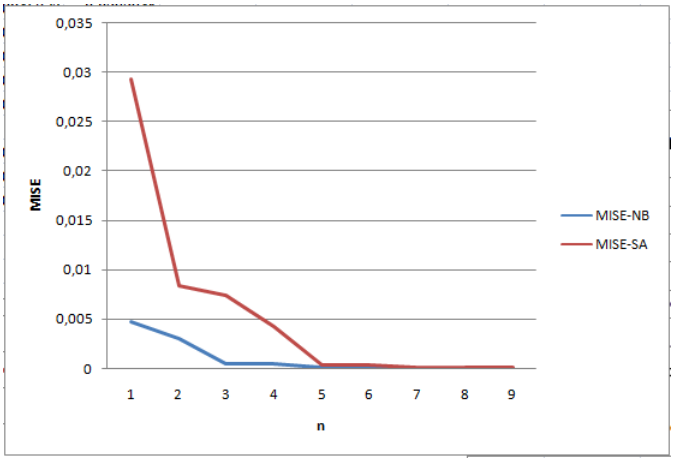

FIGURE 5.1 – Variation de MISE optimal associé à la nouvelle base et a la base de Saadi et Adjabi en ulilisant la méthode de Kronmal-Tarter

5.1.3 Interprétation

La méthode de Kronmal-Tarter donnes les mêmes valeurs du paramètre de lissage [Tab. 5.1] quand on estime la densité par la base d'Efromovich ou par la base de Saadi et Adjabi.

Estimateur de la densité

- Le tableau [Tab. 5.1] rend compte de la supériorité de l'estimateur de la densité construit à partir de la base d'Efromovich sur l'estimateur de Saadi et Adjabi, car les valeurs de l'erreur quadratique moyenne intégrée $MISE^*_{E-kt}$ sont toujours inférieures aux valeurs de l'erreur quadratique moyenne intégrée $MISE^*_{SA-kt}$ associées à l'estimateur de Saadi et Adjabi. Ceci est confirmé sur le graphe ou l'on constate que la courbe de $MISE^*_{E-kt}$ est toujours au dessous de celle de $MISE^*_{SA-kt}$.

- L'erreur quadratique moyenne intégrée diminue, quand on augmente la taille de l'échantillon.

5.1.4 Estimateurs de la fonction de répartition

Méthode de Kronmal-Tarter

La méthode consiste à :

1. Simuler un échantillon $(X_1, \ldots, X_n)$, $0 \leq X_i \leq 1$, de loi $\mathcal{N}(0,1)$.
2. La taille des échantillons est fixée successivement à 50, 100, 1000, 1500, 2000, 2500, 3000, 3500, 5000, 7500 et 10000.
3. La base trigonométrique est :

$$e_k(x) = \begin{cases} 1, & \text{si k=0 ;} \\ \sqrt{2}\cos(\pi k x), & \text{si k=0,1,2,...} \end{cases}$$

4. D est égal à la partie entière de $(log(n))^{(1/2)}$, tel que :
   $D \longrightarrow \infty$, et $\frac{D}{n} \longrightarrow 0$ lorsque $n \longrightarrow \infty$.
5. Calculer le paramètre de lissage optimum associé à la nouvelle base d^*_{NB-kt} :

$$d^*_{E-kt} = \begin{cases} \inf\{d_n : 1 \leq d_n \leq D\}; \hat{\Delta}_E > 0 \\ D \qquad\qquad\qquad\qquad sinon. \end{cases}$$

   où

$$\hat{\Delta}_E = \frac{1}{n}[\frac{n+1}{n-1}\sum_{i=1}^{n}[\sqrt{2}\cos(\pi k X_i) - \frac{1}{n}\sum_{i=1}^{n}\sqrt{2}\cos(\pi k X_i)]^2 - \sum_{i=1}^{n}(\sqrt{2}\cos(\pi k X_i))^2].$$

6. Calculer l'erreur quadratique moyenne intégrée optimale associée à la nouvelle base :

$$MISE^*_{E-kt} = \int_0^1 F^2(x)dx + \sum_{k=0}^{d^*_{E-kt}}(\mathbb{V}ar(\hat{A}_k) - A_k^2).$$

$$A_k = \sqrt{2}\int_0^1 F(x)^2 \sin(-\pi k x)dx,$$

   et

$$\hat{A}_k = \frac{\sqrt{2}}{\pi k n}\sum_{i=1}^{n}\sin(-\pi k X_i).$$

7. Calculer le paramètre de lissage optimal associé à l'estimateur de Saadi et Adjabi :

$$d^*_{SA-kt} = \begin{cases} \inf\{d_n : 1 \leq d_n \leq D\}; \hat{\Delta}_{SA} > 0 \\ D \qquad\qquad\qquad\qquad sinon. \end{cases}$$

où

$$\hat{\Delta}_{SA} = \frac{1}{n}[\frac{n+1}{n-1}\sum_{i=1}^{n}(\frac{1}{\sqrt{2\pi}}(\cos(X_i)+\sin(X_i)) - \frac{1}{n}\sum_{i=1}^{n}(\frac{1}{\sqrt{2\pi}}(\cos(X_i)+\sin(X_i)))^2 -$$

$$\sum_{i=1}^{n}(\frac{1}{\sqrt{2\pi}}(\cos(X_i)+\sin(X_i)))^2].$$

8. Calculer l'erreur quadratique moyenne intégrée optimale associée à l'estimateur de Saadi et Adjabi :

$$MISE^*_{SA-kt} = \int_{-\pi}^{\pi} F^2(x)dx + \sum_{k=0}^{d^*_{SA-kt}} (\mathbb{V}ar(\hat{A}_k) - A_k^2).$$

avec

$$A_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} F(x)^2(\cos(kx)+\sin(kx))dx,$$

et

$$\hat{A}_k = \frac{1}{\sqrt{2\pi}n}\sum_{i=1}^{n}\int_{X_i}^{\pi}[\cos(kX_i)+\sin(kX_i)]dx.$$

Résultats de la simulation

Les résultats de la simulation sont donnés sous forme de tableaux et de graphiques. Les tableaux contiennent les résultats suivants :

◇ d^*_{E-kt} : paramètre de lissage optimal associé à la base d'Efromovich en utilisant la méthode de Kronmal-Tarter ;

◇ d^*_{SA-kt} : paramètre de lissage optimal associé à l'estimateur de Saadi et Adjabi en utilisant la méthode de Kronmal-Tarter

◇ $MISE^*_{E-kt}$: erreur quadratique moyenne intégrée optimale associée à la nouvelle base en utilisant la méthode de Kronmal-Tarter ;

◇ $MISE^*_{SA-kt}$: erreur quadratique moyenne intégrée optimale associée à l'estimateur de Saadi et Adjabi en utilisant la méthode de Kronmal-Tarter ;

n	d^*_{E-kt}	d^*_{SA-kt}	$MISE^*_{E-kt}$	$MISE^*_{SA-kt}$
50	1	2	$8.4172 * 10^{-4}$	0.00398
100	2	3	$2.7323 * 10^{-4}$	0.00359
1000	3	3	$1.1656 * 10^{-4}$	0.00353
1500	3	3	$5.6466 * 10^{-5}$	0.00347
2500	3	3	$2.9269 * 10^{-5}$	0.00346
3000	4	3	$2.2804 * 10^{-5}$	0.00337
5000	4	4	$2.1300 * 10^{-5}$	0.00335
7500	4	4	$8.4695 * 10^{-6}$	0.00332
10000	4	4	$4.9435 * 10^{-6}$	0.00311

TABLE 5.2 – Paramètre de lissage optimal et variation du $MISE$ optimal en fonction de n pour la Loi normale $\mathcal{N}(0,1)$, en utilisant la méthode de Kronmal-Tarter pour la base d'Efromovich et la base de Saadi et Adjabi

Représentation graphique

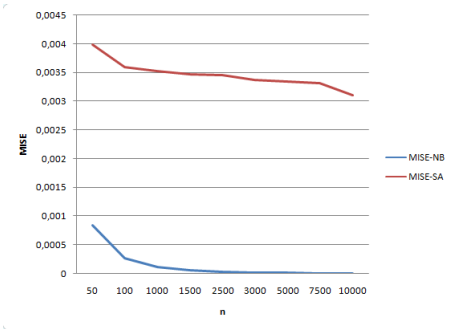

FIGURE 5.2 – Variation de MISE optimal associé à la nouvelle base et a la base de Saadi et Adjabi en ulilisant la méthode de Kronmal-Tarter

5.1.5 Interprétation

La méthode de Kronmal-Tarter donne les mêmes valeurs du paramètre de lissage [Tab. 5.2] quand on estime la fonction de répartition par la base d'Efromovich ou par la base de Saadi et Adjabi.

Estimateur de la fonction de répartition

- Le tableau [Tab. 5.2] rend compte de la supériorité de l'estimateur de la fonction de répartition construit à partir de la nouvelle base sur l'estimateur de Saadi et Adjabi, car les valeurs de l'erreur quadratique moyenne intégrée $MISE^*_{E-kt}$ associées à l'estimateur construit à partir de base d'Efromovih sont toujours inférieures aux valeurs de l'erreur quadratique moyenne intégrée $MISE^*_{SA-kt}$ associées à l'estimateur de Saadi et Adjabi. Ceci est confirmé sur le graphe ou l'on constate que la courbe de $MISE^*_{E-kt}$ est toujours au dessous de celle de $MISE^*_{SA-kt}$.

- L'erreur quadratique moyenne intégrée associée à l'estimateur construit à partir de la nouvelle base diminue, quand on augmente la taille de l'échantillon.

5.2 conclusion

La simulation a permis de constater que les performances des estimateurs construit associés a la base d'Efromovich sont meilleures que celles de l'estimateur de Saadi et Adjabi.

Conclusion générales

Ce travail est une contribution au problème d'estimation de la densité de probabilité et la fonction de répartition.

Le comportement asymptotique (consistance, vitesse de convergence faible et forte...) des estimateurs classiques : estimateur à noyau, estimation par Histogramme de la densité de probabilité et la fonction de répartition a été étudié par de nombreux auteurs. Les estimateurs basés sur des séries orthogonales ont été introduit, mais leur comportement asymptotique a été relativement moin étudié que celui des estimateurs classiques et la plupart sont appliquées à l'estimation de la densité de probabilité. Notre ojectif est de développer une théorie plus approfondie en intégrant les estimateurs proposés dans une même théorie, et de trouver de nouvels estimateurs.

Dans le premier chapitre , nous avons exposé les différentes méthodes d'estimation de la densité de probabilité, à savoir l'estimation par les séries orthogonales, l'estimation par histogramme et l'estimation par la méthode du noyau. Nous nous sommes intéressés à la méthode des séries orthogonales vu sa souplesse d'utilisation et elle présente de bonnes propriétés asymptotiques.

La deuxième partie est motivée par l'application de cette méthode d'estimation de la densité en utilisant une base trigonométrique. L'estimateur construit à partir de cette base a de bonnes propriétés de convergence : il est asymptotiquement sans biais,il est convergent en moyenne quadratique, convergent en moyenne quadratique inégrée, convergent en probabilité.

Dans le chapitre 3 nous avons donné les différentes méthodes d'estimation de la fonction de répatition à savoir : l'estimateur empirique, l'estimateur à noyau et l'estimateur

à la base de spline.

Dans le quatrième chapitre, nous avons introduit un nouvel estimateur de la fonction de répartition basé sur un système trigonométrique et nous avons établi ses résultats asymptotiques.

Afin d'étudier les qualités statistiques des estimateurs obtenus , on les compare numériquement par simulation aux estimateurs associés à la base de Saadi et Adjabi (2016). Les résultats numériques montrent que :

- les performances des estimateurs s'améliorent lorsque la taille de l'échantillon augmente.
- Les performances de l'estimateur associées à la base d'Efomovich sont meilleurs que celles de l'estimateur de saadi et Adjabi .

5.3 Perspectives de recherche

- Estimer les quantiles en utilisant l'estimateur non paramétrique de la fonction de répartition.
- Estimmer la fonctions de répartion dans le cas multidimensionnel.
- Estimer le mode de la densité de probabilité en utilisant l'estimateur non paramétrique de la densité de probabilité par des fonctions orthogonales.

Bibliographie

Bibliographie

[1] A.Berlinet. *Convergence des estimateurs splies de la densité.* publications de l'institut de statistique de l'université de paris 26, p.1-16, (1981).

[2] A. Berlinet, C.THOMAS-AGNAN. *Rreproducing Kernel in Hilbert spaces in probability and statistics* Kluwer, Boston, 2004.

[3] D.Bosq. *Estimation adaptatif de la fonction de densité par projection tronquée.* C.R.Acad.Sci.Paris,Ser.I,334 :591-595,2002.

[4] D.Bosq, J.Lecoutre. *Théorie de l'estimation fonctionnelle.* Economicaedition, 1967.

[5] J.Bleues, D.Bosq.*condition nécessaire et suffisante de convergence de l'estimateur densité par la méthode des séries orthogonales.*Rev.Roum.Math.Pures et App,(24) :869-886,1979.

[6] P. Caperaa, B. Van CUSTEM. *Méthodes et modèles en statistique non paramétrique.* Bordas, Paeis, 1988.

[7] N. Cencov. *Evaluation of unknow distribution density from observation.* SovMaths, (3) :1559-1562, 1962

[8] L. Devroy, G. Wise. *On the recovery of discrete probability densities from imperfect measurements.* Journal of the Franklin Istitue 307, p. 1-20, (1979).

[9] L.Devroye. *A course in density estimation.* Birkhauser, Boston, 1987.

[10] L.Devroy, L.Györfl. *Non paramtric density estimation.* Wiley, New York, 1985.

[11] P.Diggle, P.Hall. *A The selection of terms in an orthogonal serie density estimator.* J.Amer. Star. Assoc, 81, 230-233,1986.

[12] M. Donsker. *Justification and estension of Doob's heuristic approach to the Kolmogrov-Smirnov theorems.*The Annals of Mathematical Statistics 23, p. 277-281,(1952).

[13] J. Doob. *Heuristic approach to the Kolmogrov-smirnov theorem.* The annals of Mathematical statistics 20, p. 393-403, (1949).

[14] S.Efromovich.*orthogonal series density estimation.* Interdisciplinary Review, Computational Statistics, 2,467-46, 2010.

[15] M.Falk. *Relative efficiency and deficiency of kernel type estimators of smooth distribution function.* Statistica Neerlandica, 37, p.73-83, 1983.

[16] T.Gasser, P.Hall, and B.Presnell *Nonparametric estimation of the mode a distribution of random curves.* Journal of the Royal Statistical Society. 60, 681-691, 1984.

[17] U. Grenander. *On the theory of mortality measurement part II.* SkandinaviskAktuarietidskrift 39 , p. 125-153, (1956).

[18] P.hall,R.Wolff and Q.Yao. *Methods for estimating a conditional distrib function.*Journal of Americain statistical Association 94,p.154-163.(1999).

[19] H.Hart. *On the choice of truncation point in Fourier series density estimators.* J Stat Comput. Simul, 21,95-116,1985.

[20] P. Hennequin et A. *Théorie des probabilities et quelques applications.* Masson, Paris, 1965.

[21] S.Julian.*An Assessment of Hermite Function Based Approximation of Mutuel Information Applied to Independant Component Analysis.*10.745-756.(2008).

[22] J. Kiefer et J. Wolfowitz. *Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters* The Annals of Mathematical Statistics 27 , p.887-906, (1956).

[23] A. Kolmogrov. *sulla determinazione empirica di una legge de distribuzione.* Giornaledell'Instituto Italiano degliAttuari 4 p. 83-91, (1933).

[24] R. Kronmal and M. Tarter. *The estimation of probability densities and cumulatives by fourier series methods.* J. Amer. Statist. Assoc, (63) : 925-952, 1968.

[25] K.Lagha. S.Adjabi *Nonparametric sequential estimation of the probability density function by orthogonal series*. Communication in Statistics-Theory and Methods, DOL : 10.1080/03610926.2015.1115075.

[26] M.Lejeune, P.SARDA. *Smooth estimators of distribution and density functions* Computational Statistics and Data Analysis 14, p. 457-471, (1992).

[27] E. Lehmann. *Theoriy of point estimation* Wiley, New-York, 1983.

[28] M.Lock, *Optimising density estimates based on an weighted and weighted mean integrated squared error*. Unpublished PH. D. Dissertation,University of California, Berkely, Group in Biostatistics,1990.

[29] P. Massart. *The tight constant in the Dvoretzky-Kiefer-Wolfoxitz inequality*. The Annals of Probability 18 ; p.1269-1283, (1990).

[30] R. Modarres. *Efficient nonparametric estimation of a distribution fuction* Computational Statistics and Data Analysis 39, p. 75-95, (2002).

[31] E.Nadarya.*Some new estimates for distribution function*. Theory of Probability and its Application, 9, 497-500.

[32] E.Parzen. *On the estimation of a probability density function and mode*. Annals of Mathematical Statistics, 33 :1065-1027, 1962.

[33] K. Pearson. *On the systematic fitting of curves to observations and measurements*.Biometrica, 1, 265-303. 2, 1-23.

[34] H. Peter. *Comparaison of Two Othogonal Series Methods of Estimating a density and its derivatives on an Interval*.Journal of Multivariate Analysis 12, 432-449 1982.

[35] M. Rdemo. *Empirical choice of histograms and kernel density estimators*. Scandinavian Journal of Statistics, 9 :65-78, 1982.

[36] M.Reiss. *Noparametric estimation of the smoth distribution function*. Scandinavian Journal of Statistics,P.116-119, 1981.

[37] M.Rosenblatt. *Remarks on some nonparametric estimates of a density function*. Annals of Mathematical Statistics, 27 :8332-837, 1956.

[38] N. Saadi, S. Adjabi. *On the estimation of the probability density by trigonometric series*.Communications in Statistics-Theory and Methods, 38(3583-3595), 2009.

[39] A.Sanson. *Orthogonal function.* Pure and Applied Math. Interscience Publ, New York, 1959.

[40] G. Saporta. *Probabilités, Analyse des données et Statistiques.*Technip, Paris, 1990.

[41] D. Scott R. A. Tapia and J.R. Thompson. *Kernel density estimation revisited.* Nonlinear Analysis, Theory, Method and Applications, 1 :339-372, 1977.

[42] A.Svoretzky, J.Kiefer et J.Wolfowitz *Asymptotic minimax character of the sample distribution function and the classical multinomial estimator.* The Annals of Mathematical Statistics. 33, P.642-669, 1956.

[43] G.Szego. *Orthogonal polynomials.* American Mathematical Society, New York, 1959.

[44] M.Tarter, M.Lock. *Model-free curve estimation.* New York : chapman and Hall, 1993.

[45] H.Yamato. *Uniform convergence of an estimator of a distribution function.* Bulletin on Mathematical statistics, 15, 69-78, 1973.

Résumé

Le travail développé dans ce mémoire se situ à l'intersection entre deux thématiques importantes de la statistique non paramétrique, à savoir l'estimation non paramétrique de la densité de probabilité, d'une fonction de répartition. L'approche utilisée est la méthode des fonctions orthogonales. Nous avons estimé la densité de probabilité et nous donnons les propriétés statistiques et asymptotiques de l'estimateur obtenu. Une application réalisée pour une base trigonométrique. Nous obtenons la forme de l'estimateur de la fonction de répartition ainsi que ses propriétés statistiques.

Mots-clés : Estimation non paramétrique, Densité de probabilité, Fonction de répartition.

Abstract

The work developed in this paper is at the intersection of two important nonparametric statistics themes, namely the nonparametric estimation of the probability density, of a distribution function. The approach used is the orthogonal function method. We have estimated the probability density and we give the statistical and asymptotic properties of the obtained estimator. An application made for a trigonometric basis. We obtain the form of the estimator of the distribution function as well as its statistical properties.

Keywords: nonparametric estimation, probability density, distribution function.